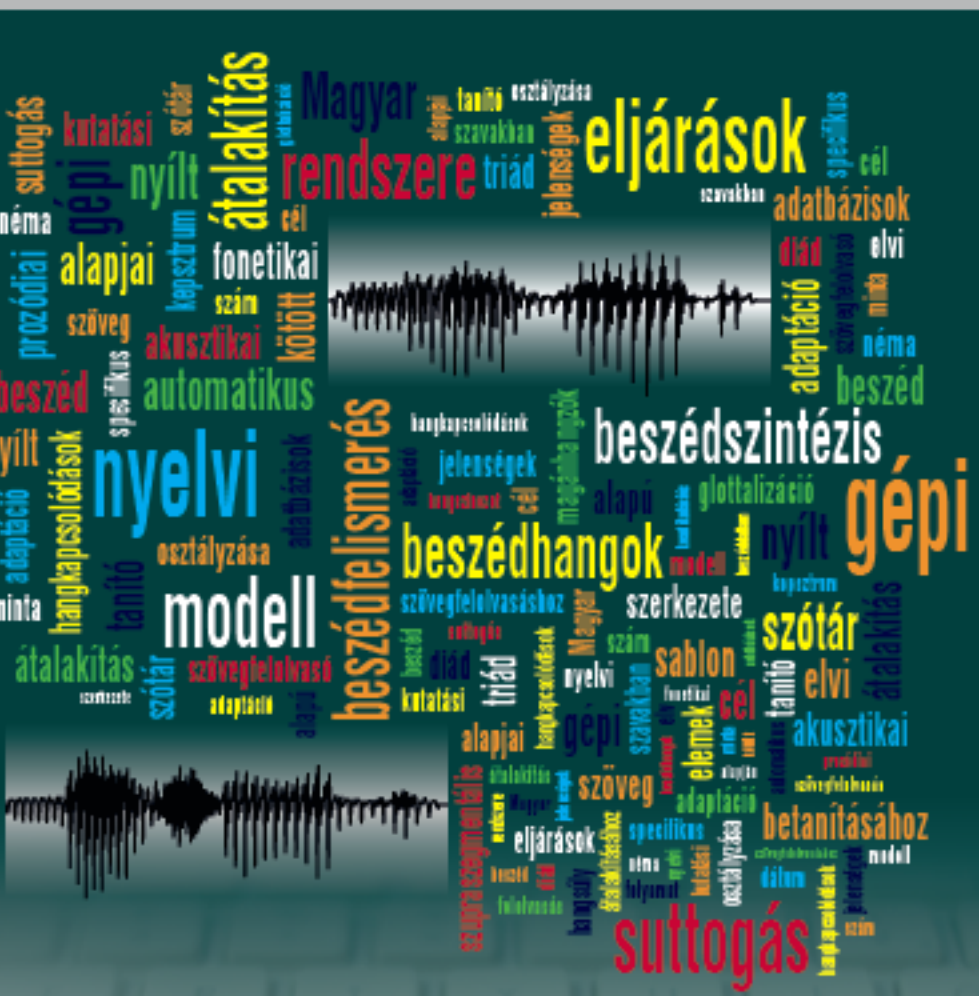
A MAGYAR BESZÉD

BESZÉDKUTATÁS, BESZÉDTECHNOLÓGIA, BESZÉDINFORMÁCIÓS RENDSZEREK

Szerkesztette: Németh Géza – Olaszy Gábor



AKADÉMIAI KIADÓ

A MAGYAR BESZÉD

A MAGYAR BESZÉD

Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek

Szerkesztette

NÉMETH GÉZA
OLASZY GÁBOR

AKADÉMIAI KIADÓ, BUDAPEST, 2010

A könyv megjelenését támogatták
BME Távközlési és Médiainformatikai Tanszék
A Magyar Tudományos Akadémia Műszaki Tudományok Osztálya
Magyar Telekom
Nyelv és Beszédtechnológiai Platform
Interton Elektroakusztikai Kft.
LogMeIn Kft.

Áttekintő szerkesztő: Gordos Géza

Szöveggondozás: Laczkó Klára
Ábrák: Balogh Ágnes, Bartalis Mátyás

Szerzők
Abari Kálmán, Bartalis Mátyás, Bőhm Tamás, Csapó Tamás Gábor, Campbell Nick, Czap László, Fegyó Tibor, Kiss Géza, Mihajlik Péter, Németh Géza, Olaszy Gábor, Szaszák György, Takács György, Tatai Péter, Tóth Bálint, Vicsi Klára, Viktóriusz Ákos, Zainkó Csaba

Lektorok
Dr. Csirik János, Dr. Czap László, Dr. Markó Alexandra, Dr. Takács György

ISBN 978 963 05 XXXXX

Kiadja az Akadémai Kiadó,
az 1759-ben alapított
Magyar Könyvkiadók és Könyvterjesztők Egyesülésének tagja
1117 Budapest, Prielle Kornélia u. 19.
www.akademiaikiado.hu

Első magyar nyelvű kiadás: 2010

Printed in Hungary

Tartalomjegyzék

BESZÉDTECHNOLÓGIA

FÜGGELÉK

Előszó

„*A nyelv mint eszköz a legbonyolultabb gépezet. Ezt a gépezetet működésében tanulmányozni, alkatrészeire bontani, e részek szerepét, egymásba illeszkedésük módjait vizsgálni, titkait megfejteni, maga is, bonyolultságában olykor fárasztónak tetsző, de a belemélyülő számára végtelenül érdekes, és soha véget nem érő tanulmány.*" (Bárczi 1963)

Bárczi Géza 1963-ban írt soraival adjuk át e könyvet a Tisztelt Olvasónak. A fenti gondolat a beszédre mint a nyelv hangzó megjelenési formájára is igaz. A beszéd kifejező ereje végtelen variáltságú, hiszen a földön mintegy 7000 nyelven (Lewis 2009) beszélnek emberek, és gondolataikat, érzéseiket ezzel a kommunikációs eszközzel adják át társaiknak. Mi, magyarok abban a szerencsés helyzetben vagyunk, hogy a 3 millió fölötti anyanyelvű (elsődlegesen tanult) beszélővel rendelkezők száma szerint sorrendbe helyezett 172 nyelv (az összes beszélt nyelv kevesebb, mint 2,5%-a) közé tartozó – annak is az első felében, a rangsor 73. helyén található – nyelven beszélünk.

Ebben a könyvben a magyar beszédet mutatjuk be a 21. század tudományos és technikai eredményeinek tükrében. Célunk megmutatni és rögzíteni a magyar beszéd akusztikai szerkezeti képét a 21. század elején, ismertetni a beszédtechnológia mint új, interdiszciplináris tudományág eddig elért eredményeit, problémaköreit és alkalmazásait, főleg hazai vonatkozásban. Vannak benne viszonylag időtálló fejezetek (például a beszédakusztikai és a jelfeldolgozási témakörök), és vannak olyan alkalmazási és technológiai fejezetek (például a beszédtechnológiai alkalmazásokra vonatkozók), amelyek a mai kor szintjét ismertetik. A könyvhöz rendelt honlap (`http://magyarbeszed.tmit.bme.hu`) pedig sok olyan adatot tartalmaz, amit a könyvben a terjedelem korlátozottsága miatt nem lehet elhelyezni. A 21. század egyik új iparága a beszédtechnológia. Ez a könyv az első szisztematikus összefoglalás magyar nyelven a magyar beszédre vonatkozóan ebben az aspektusban. A könyv célja és remélhetőleg érdeme, hogy mindezeket egységes keretbe összefogva tárgyalja, érzékeltetve az olvasóval a fenti idézet valószerűségét.

A beszédtechnológia az elmúlt évtizedben került az információs és kommunikációs technológiák (rövidítve IKT) szélesebb témakörén belül is a kutatások előterébe. Ahogy a munkahelyről a mindennapi életbe és a gyerekek kezébe is átkerülnek az egyre több funkcióval bíró eszközök és szolgáltatások, úgy válik egyre fontosabbá az, hogy használatuk érdekes, vonzó, könnyen megtanulható és kezelhető legyen. Az emberi beszédkommunikáció évezredek alatt kialakult formáinak alkalmazása a gépek és az ember közötti információcserére óriási lehetőségeket rejt. Ennek az útnak egyenlőre nagyon az elején járunk, hiszen a gépek az emberi kommunikációs képességeknek (különösen is a jelentések értelmezésének, a szemantikus szintnek) a töredékével rendelkeznek. Ezért felmerülhet, hogy a gépi megoldásoknak nem is feltétlenül az ember, hanem az emberrel szoros kapcsolatban álló más élőlény (például a kutya) kommunikációs formáit kell követni. Ezen a területen, az ún. etoinformatika és etokommunikáció témakörében is ígéretes kutatások indultak meg a közelmúltban.

Fontos annak tudatosítása, hogy a beszédtechnológiai megoldások pénzügyi értelemben vett áttörése előre nehezen tervezhető. Már 1983-ban is azt jósolták komoly szakértők, hogy két éven belül exponenciális növekedés várható az angol nyelvű beszédfelismerés piacán. Ez nem következett be, helyette lineáris növekedés valósult meg, ami a finanszírozók jelentős részének kedvét szegte. Azóta is megfigyelhető 5–6 éves ciklusokban a marketing jellegű túlzó ígéretek és a valódi teljesítmény ellentétéből adódó figyelem-összpontosulás, majd -ellanyhulás. Azonban egyértelmű a hatalmas fejlődés, ha összehasonlítjuk a 80-as évek elejének és napjainknak a megoldásait. Szerencsére Magyarországon a beszédkutatásnak jelentős hagyománya van, ezért nem kell a nagy világcégek technológiáinak a mi viszonylag kis piacunkra történő honosítására várni, hanem hazai szellemi és anyagi erőforrásokra építve is versenyképes megoldásokat hoztunk létre.

Célközönségünk az egyetemek, főiskolák, valamint minden olyan oktatási hely, ahol informatikusokat képeznek. A könyv jó támogatást adhat távközlési fejlesztőknek és döntéshozóknak, a beszédtechnológiai fejlesztések szakembereinek, új tartalomszolgáltatási, egészségipari és rehabilitációs szolgáltatások tervezőinek. Azonban ennél szélesebb rétegnek is szeretnénk ajánlani könyvünket. Segítheti a humán területek oktatását is (fonetika, beszédelemzés, nyelvészet és a beszéd kapcsolata, beszédpszichológia, egészségügyi betegségmegelőzés, rehabilitáció, tájékoztatás). Ajánljuk továbbá a középiskolások felső tagozatának is, valamint mindenkinek, akit érdekel a témakör (például fizikusok, nyelvészek, rádiósok, televíziósok, filmesek, tudományos média szakemberei). Átfogó tartalma miatt hasznos információkat találhatnak benne a fenti szakmák művelői, a mérnököktől a bölcsészekig.

A munka a BME Távközlési és Médiainformatikai Tanszék kutatásai és fejlesztései köré épül, egyúttal kitekint a nemzetközi beszédkutatásra is. A szerzők a tanszéken jelenleg vagy korábban dolgozó oktatók és kutatók. A tartalom kidolgozása során figyelembe vettük a BME Villamosmérnöki és Informatikai Karán a témakör-

ben kidolgozott tárgyak, különösen a Beszédinformációs rendszerek tárgy oktatása során szerzett tapasztalatokat. Ebben az értelemben szerzőtársnak tekintjük a tárgyból sikeresen levizsgázott több, mint ezer hallgatónkat is.

Az alkalmazási példák is ehhez a körhöz kötődnek. Már csak terjedelmi okokból sem törekedhettünk a szerencsére ma már hazánkban is gyarapodó számú költségvetési és ipari kutatóhely eredményeinek teljes körű lefedésére. A témakör szélesebb kitekintésű – a nyelvtechnológiát is magában foglaló – hazai helyzetképe áttekintésének kiindulópontjául ajánljuk a Nyelv- és Beszédtechnológiai Platform honlapját: `http://hlt-platform.hu`. A könyvhöz tartozik a `http://magyarbeszed.tmit.bme.hu` honlap is, ahol oktatáshoz és kutatáshoz használható adatokat, adatbázisokat és programokat adunk közre.

Minden olvasónkat szeretettel várjuk a témakörben való elmélyülésre.

A könyvben közzétett kutatási és fejlesztési eredmények megszületéséhez az alábbi projektek támogatásai járultak hozzá.

Természetes beszédinformációs rendszerek: NKFP-2/034/2004
Beszélő mobiltelefon: GVOP-3.1.1-2004-05-0485/3.0
Ambiens intelligenciára épülő ipari alkalmazások kutatás-fejlesztése – BelAmi: ALAP2-00004/2005
Gyógyszervonal: GVOP-3.1.1 - 2004 - 05 - 0426 /3.0
VOXearch – digitális médiaarchívumok automatizált kategorizálása beszédfelismerés segítségével: GVOP – 3.1.1 – 2004 – 05 – 0385/3.0
Új vizsgálati és mérési módszerek kidolgozása korszerű távközlési szolgáltatások minőségének biztosítására : NKFP/002/015/2005
Teleauto: OM-00102/2007
Ember és informatikai rendszerek kapcsolatának új, etológiai modell alapú generációja: TÁMOP-4.2.2-08/1/KMR-2008-0007
Intelligens Multimodális Tudásközpont: KMOP – 1.1.1 – 07/1 – 2008 – 0034
Beszédfelismeréssel támogatott online multimédia menedzsment és médiatartalomra célzott hirdetési szolgáltatás kialakítása: KMOP – 1.1.3 – 08/A – 2009 – 0006

Szerkesztők – szerzők

Áttekintő szerkesztő:

Gordos Géza (1937) a beszéd mérnöke, a műszaki indíttatású beszédkutatás vezéralakja. A Budapesti Műszaki Egyetemen (BME) szerzett híradástecnikai szakos oklevelet 1960, dr. univ. 1966, kandidátus 1977, az MTA doktora 1995, dr. habil. 1995. 1960-tól a BME-n dolgozik. 1970-ben megalapította a Beszédfeldolgozási Laboratóriumot, amely 2010-ben is meghatározó szereppel bír a hazai beszédkutatásban és az alkalmazások fejlesztésében. Számos szakmai kitüntetés tulajdonosa, nemzetközileg elismert kutató. Kutatási területei: távközlés, beszédkódolás, beszédszintézis, beszédfelismerés.

Szerkesztők és szerzők:

Németh Géza (1959) villamosmérnök, híradástechnikai szakmérnök. A BME Villamosmérnöki Kar Híradástechnikai Szakán végzett (1983), a BME-n doktorált (dr. univ. 1987, PhD 1997). A BME TMIT Beszédtechnológiai Laboratórium vezetője. Kutatási területei: beszédtechnológia, szolgáltatásautomatizálás, többnyelvű beszéd- és multimodális információs rendszerek, mobil felhasználói felületek és alkalmazások. Egyik fejlesztője a ProfiVox magyar szövegfelolvasó szoftvernek, valamint számos mobilos beszédinformácós rendszernek.

Olaszy Gábor (1943) villamosmérnök, fonetikus. A BME Villamosmérnöki Kar Híradástechnikai Szakán végzett (1967), a BME-n doktorált (1985), a nyelvtudomány kandidátusa (1988), az MTA doktora (2003), habilitált (2004). Kutatási területei: a beszéd akusztikai szerkezete, fonetikai modellezés, szöveg-beszéd átalakító rendszerek tervezése, készítése, tesztelése. Egyik fejlesztője a ProfiVox magyar szövegfelolvasó szoftvernek, valamint számos beszédkutatási honlapnak.

Szerzők:

Abari Kálmán (1971) programtervező matematikus. A Debreceni Egyetem Pszichológiai Intézetében dolgozik, jelenleg PhD-hallgató. Fő érdeklődési területe a beszédfeldolgozás (a beszédjel akusztikai vizsgálata, beszédadatbázisok építése, interaktív beszédkutatási honlapok készítése), a matematikai statisztika pszichológiai alkalmazásai, statisztikai gépi tanulás és a mesterséges intelligencia (szakértői rendszerek, keresési eljárások).

Bartalis Mátyás (1981) műszaki informatikus. 2005-ben a BME Villamosmérnöki és Informatikai Karán, médiainformatika szakirányon végzett. Oklevele megszerzése óta a BME Távközlési és Médiainformatikai tanszékének Beszédtechnológiai laborjában dolgozik. Fő tevékenysége a beszédszintetizátorokon alapuló alkalmazások fejlesztésében való részvétel, valamint a beszédszintetizátorok adatbázisainak fejlesztése, javítása.

Bőhm Tamás (1980) műszaki informatikus. 2003-ban a Budapesti Műszaki és Gazdaságtudományi Egyetemen diplomázott. 2010-ben PhD-fokozatot szerzett. Vendéghallgatóként a University of New Hampshire-en és a Massachusetts Institute of Technologyn tanult. Jelenleg az MTA Pszichológiai Kutatóintézetében tudományos munkatárs, valamint a BME Távközlési és Médiainformatikai Tanszéken is kutató.

Campbell Nick (1948) beszédkutató. PhD-fokozatot kísérleti fiziológia témában kapott (1990) a University of Sussexen. Az IBM kutatója (1985), ahol beszédszintézis-algoritmusokat fejlesztett. At AT-T Bell Laboratories (1991) kutatója. Nyelvész tanácsadó az edinburghi egyetemen. 1991-től az ATR kutatója Japánban. Kutatási területei: természetes kommunikáció, spontán beszédadatbázisok készítése, nemverbális beszédkommunikáció, beszédszintézis, prozódiai modellezés.

Csapó Tamás Gábor (1985) informatikus. Diplomáját a BME Villamosmérnöki és Informatikai Karán szerezte 2008-ban. 2007-ben OTDK 1. helyezést szerzett beszédszintézis témakörű dolgozatával. 2008 óta a BME Beszédtechnológiai Laboratóriumának PhD-hallgatója. Kutatási témája az emberi beszéd prozódiájának modellezése és a beszédszintetizátor rendszerek természetesebbé tétele.

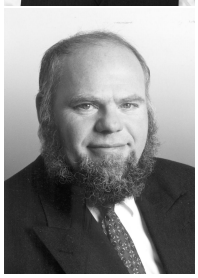
Czap László (1957) villamosmérnök, híradástechnikai szakmérnök. A BME Villamosmérnöki Kar Híradástechnika Szakán végzett (1980), híradástechnikai szakmérnök (1983), a BME-n szerzett doktori címet (dr. tech. 1987, PhD 2005). A Miskolci Egyetemen az Automatizálási Tanszék vezetője. Kutatási területe: audiovizuális beszédfelismerés és beszédszintézis.

Fegyó Tibor (1973) műszaki informatikus. A BME Villamosmérnöki és Informatikai Kar Műszaki Informatika Szakán végzett (1997). Kutatási területei: a beszéd számítógépes felismerése, akusztikai és nyelvi modellezése, beszédfelismerés alapú rendszerek tervezése és fejlesztése, valamint távközlési csatornák beszédminőségének mérése. Elévülhetetlen érdemeket szerzett a magyar gépi beszédfelismerés kutatásában. A PhD tudományos fokozat várományosa.

Kiss Géza (1975) villamosmérnök. 1997-ben végzett a BME Villamosmérnöki és Informatikai Karán, műszaki informatika szakon. 1999-től a BME TMIT beszédtechnológiai laborjában dolgozott, a beszédszintézis technológiákhoz kapcsolódó kutatási és fejlesztési munkákon. A ProfiVox magyar szövegfelolvasó rendszer egyik fejlesztője. 2009–2010 között Fulbright-ösztöndíjas, 2010-től PhD-hallgatóként a Center for Spoken Language Understanding (OHSU, USA) laborban végez kutatómunkát.

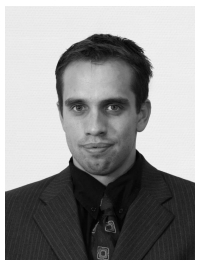
Mihajlik Péter (1975) okleveles villamosmérök. Diplomáját a Budapesti Műszaki Egyetemen (BME) szerezte 1999-ben. Azóta a BME Távközlési és Médiainformatiai Tanszékének munkatársa, elsősorban a magyar nyelvű beszéd gépi felismerésének kutatásával foglalkozik. A PhD tudományos fokozat várományosa.

Szaszák György (1979) villamosmérnök. Budapesti Műszaki és Gazdaságtudományi Egyetemen szerzett oklevelet 2002-ben, 2009-ben PhD-fokozatot. A BME Távközlési és Médiainformatikai Tanszékének munkatársa. Főbb kutatási területei a gépi beszédfelismerés, beszédjel-feldolgozás, prozódiai modellezés, beszédadatbázisok.



Takács György (1946) okleveles villamosmérnök, MBA, a műszaki tudomány kandidátusa, címzetes egyetemi tanár. A Posta Kísérleti Intézet kutatója, majd kutatási főmérnöke 1995-ig. Ezután az Ericsson Magyarország rendszertervezési főosztályvezetője és tanácsadója volt. A Hírközlési Főfelügyelet távközlési igazgatója 2002-ig. Azóta a Pázmány Péter Katolikus Egyetem Információs Technológiai Karának oktatója. Kutatási területe: beszédtechnológia, multimodális beszédfeldolgozás, mobiltelefon-szolgáltatások.

Tatai Péter (1941) okleveles villamosmérnök. 1964-ben végzett a Budapesti Műszaki Egyetemen VIK híradástechnikai szakon. 1964–1986: a Távközlési Kutató Intézet tudományos munkatársa, majd osztályvezetője. 1986–2007: a BME Távközlési és Médiainformatikai Tanszék oktatója, a Távközlési Jelfeldolgozási Laboratórium (TSP-Lab) vezetője, címzetes egyetemi docens. 2005-ben alapította az AITIA International Zrt.-t, amelynek elnök-vezérigazgatója.



Tóth Bálint (1980) okleveles villamosmérnök. Kitüntetett diplomával végzett a BME Villamosmérnöki Karán távközlési és telematikai szakirányon 2005-ben. PhD-tanulmányait rögtön a diplomázás után elkezdte beszédszintézis és multimodális felhasználói felületek témakörben. A beszédszintézis területén elsősorban rejtett Markov-modell alapú szövegfelolvasással foglalkozik, emellett pedig a multimodális felhasználói felületek mobil környezetben való alkalmazási lehetőségeit vizsgálja.

Vicsi Klára (1948) akusztikus. Az ELTE TTK-n tanári oklevelet szerzett (1971), doktorált (1982), a fizikai tudomány kandidátusa (1992), az MTA doktora (2005), BME habilitáció (2007). Kutatási területei: beszédakusztika, gépi beszédfelismerés, beszédadatbázisok készítése és pszichológiai akusztika. A fonetikai kutatások valamint beszédfelismerési munkák alapjául szolgáló magyar nyelvű beszédadatbázisok megteremtője. Úttörője a magyar nyelvű gépi beszédfelismerésnek. A multimodális beszédoktató és fejlesztő eljárások kidolgozásában nemzetközi hírnévre tett szert. Számos nemzetközi konferenciát, nyári egyetemet szervezett, vagy a szervezésben vett részt.

Viktóriusz Ákos (1985) okleveles mérnök-informatikus. 2009-ben OTDK 1. helyezett. Kutatási területei: intelligens mobileszközök, beszédinformációs alkalmazások fejlesztése mobilplatformokra: Symbian, Windows Mobile, iPhone, Android.

Zainkó Csaba (1976) 1999-ben végzett a BME Villamosmérnöki és Informatikai Kar médiainformatika szakirányon, és azóta a Távközlési és Médiainformatikai Tanszék Beszédtechnológiai laboratóriumában dialógusrendszerek és az ahhoz kapcsolódó komponensek kutatásával és fejlesztésével foglalkozik. Jelenleg a korpuszalapú beszédszintézis és a multimodális interfészek kutatásán dolgozik.

A könyv szerzői és fejezeteik a tartalomjegyzék szerinti fejezetszámmal

Abari Kálmán: 13.4.
Bőhm Tamás: 5.3.3., 13.5.
Campbell Nick–Németh Géza: 14.
Csapó Tamás Gábor: 9.13., 10.3.1.3, 13.1.
Czap László: 9.12., 10.3.1.6.
Fegyó Tibor: 12.2., 12.3.6., 12.3.8., 12.4.2., 13.3.
Kiss Géza: 1., 4.1., 10.3.2., 12.7., 13.2.
Németh Géza: 11.
Németh Géza–Kiss Géza–Bartalis Mátyás: 12.7.4.
Németh Géza–Zainkó Csaba: 12.3.1, 12.3.4.
Olaszy Gábor: 2., 3.1., 3.5., 4.2.–4.5., 5., 6., 7.1.4.1., 8.2.–8.3., 8.4.2., 8.4.4., 8.5., 10.–10.3.1.2., 10.3.1.4., 10.3.4.–10.3.5., 10.4., 12.11.
Olaszy Gábor–Abari Kálmán: 8.4.1., 8.4.3.
Olaszy Gábor–Bartalis Mátyás: 8.2.3.2., 12.3.5.
Olaszy Gábor–Németh Géza: 12.4.3.
Olaszy Gábor–Németh Géza–Kiss Géza: 10.3.6.
Szaszák György: 7.1., 9.10.
Szaszák György–Mihajlik Péter–Fegyó Tibor: 9.
Szaszák György–Vicsi Klára: 12.6.1.
Tóth Bálint–Németh Géza: 10.3.8., 12.10.
Tóth Bálint–Németh Géza–Kiss Géza: 12.3.3.
Takács György: 12.8.
Tatai Péter: 7.2., 12.1.
Vicsi Klára: 3.2.–3.4., 8.–8.1., 9.11., 12.9.
Viktóriusz Ákos–Németh Géza–Tóth Bálint: 12.7.5.
Zainkó Csaba: 4.6., 10.3.1.7., 10.3.7., 10.3.9.
Zainkó Csaba–Bartalis Mátyás–Németh Géza: 12.3.7.
Zainkó Csaba–Németh Géza: 10.3.3., 12.3.2., 12.4.1., 12.5.1.

Rövidítések jegyzéke

ACELP Algebraic-CELP
ACR Absolute Category Rating
ADC Analog-Digital Converter
ADPCM Adaptive DPCM
AMDF Average Magnitude Difference Function
AMR-WB Adaptiv MultiRate-WideBand
APC Adaptiv Predictive Coder
ASR Automatic Speech Recognition
ATIS Air Travel Information System
AaS Analysis and Synthesis
AbS Analysis by Synthesis
BEA Beszélt nyelvi adatbázis
BRI-ISDN Basic Rate ISDN
BME TMIT BME Távközlési és Médiainformatikai Tanszék
CALL Computer Aided Language Learning
CELP Code Excited LP
CIS Coverage Information System
CMN Cepstral Mean Normalisation
CPU Central Processor Unit
C Consonant
DAC Digital-Analog Converter
DCR Degradation Category Rating
DC Direct Current - egyenáram
DDE Dynamic Data Exchange
DECT Digital Enhanced Cordless Telecommunications
DEC Digital Equipment Corporation
DFT Discrete Fourier Transform
DM Delta Modulation
DOS Disk Operating System
DPCM Differential PCM
DSP Digital Signal Processor
DTMF Dual Tone Multi Frequency
DTW Dinamic Time Warping
EFR Enhanced Full Rate
EM Expectation Maximization
EPR Expert Pattern Matching
ELRA European Language Resources Association
ETSI European Telecommunications Standards Institute
FAP Facial Animation Parameter
FFT Fast Fourier Transform
FIR Finite Impulse Response
GB Gigabyte
GMM Gaussian Mixture Model
GPS Global Positioning System
GSM FR GSM Full Rate
GSM Global System for Mobile Communications
GSM HR GSM Half Rate
GUI Graphical User Interface
HCI Human-Computer Interaction
HMM Hidden Markov Model
HPS Harmonic Product Spectrum
HRTF Head Related Transfer Function
HTTP Hyper Text Transfer Protocol
ICT Information and Communication Technologies
IKT Információs és Kommunikációs Technológia
IPA International Phonetic Association
ITU International Telecommunications Union
ISDN Integrated Services Digital Network
IVR Interactive Voice Response
JSAPI Java Speech Application Interface
LDC Linguistic Data Consortium
LNRE Large Number of Rare Events
LP Linear Prediction
LPC Linear Predictive Coding vagy Linear Predictive Coefficients
LSP Linear Spectral Pairs
LTP Long Time Prediction
MAP Maximum A Posteriori
MCE Minimum Classification Error
MDCT Modified Discrete Cosine Transform

MFC Mel Frequency Cepstrum
MIPS Million Instructions Per Second
MLLR Maximum Likelihood Linear Regression
ML Maximum Likelihood
MMD Modified Mahalanobis Distance
MMI Maximum Mutual Information
MNSZ Magyar Nemzeti Szövegtár
MOS Mean Opinion Score
MP3 MPEG 1 - Layer 3
MPEG-4 MPEG Compression Standard Version 4
MPEG Motion Pictures Expert Group
MPE Minimum Phone Error
MQDF Modified Quadratic Discriminant Function
MRCP Media Resource Control Protocol
MSAPI Microsoft Speech Application Interface
MSD Morpho-Syntactic Description
MSISDN Mobile Station ISDN Number
NLP Natural Language Processing
NN Neural Network
O Observation
PCM Pulse Code Modulation
PC Personal Computer
PDA Personal Digital Assistant
PER Phoneme Error Rate
PESQ Perceptual Evaluation of Speech Quality
PET Pitch and Energy over Time
PLP Perceptual LP
PRE PRozódiai Egység
PRI ISDN Primary Rate Interface
PSOLA Pitch Syncronous OverLap-Add
PSQM Perceptual Speech Quality estimation Method
PSTN Public Switched Telephone Network
PbA Pronuncation by Analogy
RAM Random Access Memory
RPE Regular Pulse Excitation
RTF Real Time Factor
RTSP Real Time Streaming Protocol
SAMPA Speech Assessment Methods Phonetic Alphabet
SAM Speech Assesment Methods
SAPI Speech Application Programming Interface
SER Sentence Error Rate
SINAD Signal to Noise And Distortion ratio
SIP Session Initiation Protocol
SMS Short Message System
SNR Singal to Noise Ratio
SPECO SPEech COrrector
SQL Structured Query Language
STFT Short Time Fourier Transform
STP Short Time Prediction
SUI Speech User Interface
SVM Support Vector Machines
TTS Text to Speech
URL Uniform Resource Locator
VAD Voice Activity Detection
VODER Voice Operation DEmonstratoR
VTLN Vocal Tract Length Normalization
VUI Voice User Interface
V Vowel
VoIP Voice over Internet Protocol
WER Word Error Rate
WFST Weighted Finite State Transducer
WRR Word Recognition Rate
XML Extensible Markup Language

EMBER, NYELV, BESZÉD

1. fejezet
A beszéd és az információs társadalom

Kiss Géza

Az ember természetéből fakad a kommunikációs igénye. A beszéd a nyelvi kommunikáció legalapvetőbb formája. A 21. század elején elmondhatjuk, hogy a kommunikáció, azaz az információ továbbítása, kölcsönös áramoltatása, a hétköznapi életünk egyik központi szereplőjévé vált. Az információhoz való hozzájutás módjai is felgyorsultak.

Míg a 18. századi ipari forradalom előtt az emberiség évezredekig megközelítőleg azonos technikai színvonalon élt, alig két évszázaddal később, az informatikai eszközök térhódításával elérkezett a digitális forradalom. Napjainkat pedig már gyakran jellemzik olyan fogalmakkal, mint „információs korszak", „információs gazdaság", „információs robbanás", sőt az „információs túlterheltség" problémája is egyre több embert érint.

Ennek ellenére tévedés volna azt gondolnunk, hogy az információ jelentőségének felfedezése az elmúlt évszázad eredménye. Már az ókori görög filozófiákban is fontos szerepe volt a mai információfogalmunkkal sok szempontból rokon koncepcióknak. Az egyik szó, amelyet használtak rá, az *eidosz*; ezen Platón a dolgok „ideális képét", esszenciáját értette, amelyeken keresztül a világ megismerhető. Egy másik szó a *logosz*, amelynek köznapi jelentése gondolat, illetve mondás; a filozófiában Hérakleitosz a világmindenség eredetét és alapvető rendjét értette rajta; a keresztény Újszövetség pedig Jézus Krisztust a Logosz megtestesülésének tartja, aki által minden létrejött.

Az információ jelentőségének újrafelfedezése a 20. század eredménye. Ekkor alapozta meg Claude Shannon az információelméletet (Shannon 1948), amely azóta számos tudományágban óriási jelentőségre tett szert. Az információt ma az anyag és az energia mellett a világmindenség harmadik alapvető összetevőjének tekintik. Sőt, a kvantummechanika új eredményeinek összefoglalásaként a fizikus Jacob Bekenstein így fogalmazott: „A jelenlegi trend [a fizikában], amelyet a Princeton Egyetemen dolgozó John A. Wheeler indított el, hogy a fizikai világot tekintsük úgy, hogy infor-

mációból épül fel, melynek az energia és az anyag csak esetleges megnyilvánulásai" (Bekenstein 2003).

Az információról több nézőpontból beszélhetünk, és ennek megfelelően többfajta definíciója létezik. Az információelméletet Shannon a hírközlés matematikai elméleteként dolgozta ki, amely a lehetséges közlemények kódolt változatának, tömörítésének, átvitelének, tárolásának, védelmének és feldolgozásának problémakörével foglalkozik (Györfi et al. 2000).

A kommunikáció áthatja a természet egészét. Kommunikáció zajlik a Föld legegyszerűbb egysejtűi között is kémiai anyagok segítségével. A hangyák társadalmában feromonok, és újabb kutatások alapján valószínűleg ultrahang segítségével születik meg a boly összehangolt tevékenysége. A méhek táncukkal magyarázzák el fajtársaiknak, hogy milyen irányban és milyen távolságra keressék az újonnan fellelt nektárlelőhelyet, és hogy az mennyire gazdag élelemben. Még hosszan lehetne sorolni az állatvilág felmérhetetlenül változatos kommunikációs formáit, amelyek genetikailag kódoltak. Emellett kimutatták, hogy számos állatfaj tanult kommunikációs módokat is használ; ilyen például a bálnák közös víz alatti „éneklése". Számos énekesmadár is a fajtársaitól eltanult dallamokkal jelzi területét (Fischer 1999, 11–34. o.).

Az emberi kommunikáció egyik legfőbb eszköze a nyelv. Ennek egyik alapvető ismertetőjegye, hogy nem örökletes, hanem tanult viselkedés; egy másik, ezzel összefüggő tulajdonsága, hogy jellegében szimbolikus. A szimbolikus gondolkodás és kommunikáció azt jelenti, hogy szimbólumokhoz (jel, jelkép) többé-kevésbé önkényesen, konvencionálisan jelentést rendelünk, és ezt használjuk az egymás közötti kommunikációra. A nyelvvel történő kommunikációt „verbális" kommunikációnak nevezzük. Ennek fő eszköze a beszéd, de még sok másfajta nyelvi kommunikáció létezik, mint például az írás. A „verbális" kifejezésnek létezik egy szűkebb értelmű használata is, amely csak a hangzó beszédet érti rajta; mi a fentebbi, tágabb értelemben használjuk ezt a kifejezést.

Az emberi kommunikációnak jelentős része verbális közlési forma, de vannak nonverbális (nem nyelvi) elemei is. Ugyanis a nyelvet megelőzi a gondolat, a logika léte. Gondolkozni tudunk nyelv nélkül is, például képekben (Silver 2001, Bennett–Hacker 2003, 337–346. o.). Gondolatokat, érzéseket közvetíteni sem csak beszéddel tudunk, hanem például tekintettel, mimikával, mozdulatokkal, testtartással, sőt zenével, tánccal, rajzokkal, szobrokkal és más műalkotásokkal. Ezek mind a nonverbális kommunikációba tartoznak. A szemtől szembe zajló kommunikációban az egész testünkkel információt adunk magunkról, és az úgynevezett modalitásokon (érzékszervi csatornákon) keresztül információt szerzünk környezetünkről, a látás, hallás, tapintás, szaglás, ízlelés, hőérzet és egyensúlyérzet segítségével.

Míg nonverbális kommunikációban folyamatosan részt veszünk akaratunkon kívül is, a verbális kommunikációhoz szükséges a közlő fél szándéka. A nyelvi közlés leggyakoribb eszköze a hangzó beszéd és a vizuális írás, de történhet más módon is. A vakok és gyengénlátók például a Braille-írásjegyeket (kitapogatható domború

jeleket) használják. A siketek jelbeszédükben kézjelekkel közvetítik mondandójukat, a siket-vak emberek pedig taktilis kommunikációt használnak, amikor egymás arcizmainak állapotait tapintással észlelik. Egyes afrikai és amerikai népek füttynyelveket is használnak. Ahogy látjuk, a verbális kommunikációnak számos megjelenési formája van, de közös jellemzőjük, hogy mindegyik vagy konkrétan a beszédképző mechanizmusunkkal jön létre, vagy a beszéd valamilyen leképezése útján.

A különböző jelrendszerek vizsgálatával és rendszerezésével a szemiotika foglalkozik. Célja, hogy együtt tanulmányozza a kommunikációs rendszerek nyelvészeti, pszichológiai, filozófiai és szociológiai jellemzőit.

Az emberi beszéddel történő kommunikációnak számos olyan jellemzője van, ami az állati kommunikációra nem igaz, sőt olyan állati nyelv sem létezik, amelyikben ezeknek a jellemzőknek a többsége egyszerre szerepelne. Alább felsorolunk néhány ilyen fontos jellemzőt (Crystal 2003):

- a hangképzés és hallás használata (nem képi vagy tapintásos);
- minden irányban terjed, de lokalizálható a hangforrás (irányérzékenység);
- pillanatnyi akusztikai jel gyors eltűnéssel (szemben például az írással);
- felválthatóság (bárki képes bármely üzenetet elmondani, például nem kötődik nemhez);
- tökéletes visszajelzés (saját üzenetét észleli);
- specializáció (a beszédhangok csak a kommunikáció célját szolgálják);
- jelentésesség (a jelzés elemei jelentést hordoznak);
- önkényesség (nincs kapcsolat a jel és a jelölt valóság között);
- áthelyezés (térben és időben távoli eseményeket is lehetséges leírni);
- nyitottság (végtelen a kifejezhető jelentések száma);
- hagyományozó továbbadás (szemben például a genetikai örökléssel);
- kettős tagoltság (véges számú beszédhangot használ fel, amelyeknek nincs önálló jelentése, viszont ezekből épülnek fel a jelentéssel bíró szavak).

Becslések szerint a Földön jelenleg mintegy 7000 az élő emberi nyelvek száma; nagyon változatosak hangkészletükben, szókincsükben és nyelvtanukban, de egyik sem primitív: nincs olyan nép, amelynek nyelve minden szempontból nagyon egyszerű vagy kis szókincsű volna (John 1981, 27–33. o.). Az 1.1. táblázatban az Ethnologue (Lewis 2009) alapján megadjuk a legtöbb anyanyelvű beszélő által beszélt nyelvek első húszas listáját, kiegészítve néhány földrajzi vagy történelmi szempontból hozzánk közel álló nyelv adataival. A különbségek mellett számos hasonlóságot is lehet találni: olyan jellemzőket, amelyek minden emberi nyelvre jellemzőek. A nyelvtudomány számos ilyen nyelvi univerzálét fedezett fel, amelyek mindegyik nyelv szemantikájára, hangtanára vagy épp nyelvtanára érvényesek, illetve amelyek általában igazak.

Az ilyen egyezések teszik lehetővé, hogy amikor a beszédet vizsgáljuk, ne kelljen minden esetben egy konkrét nyelvről beszélni, hanem olyan megoldásokat találhas-

1.1. táblázat. Legalább 3 millió anyanyelvi beszélővel rendelkező nyelvek sorrendi besorolása (a válogatás az Ethnologue 2009 adatai alapján készült, a dőlt betűvel jelzett tételek nyelvcsaládokat jelölnek, a []-ben levő hárombetűs rövidítések az ISO 639-3 szabványt követik)

Sorszám	Nyelv	Elsődleges ország	Összes ország	Beszélők száma (millió)
1	*Kínai [zho]*	*Kína*	*31*	*1213*
2	Spanyol [spa]	Spanyolország	44	329
3	Angol [eng]	Egyesült Királyság	112	328
4	*Arab [ara]*	*Szaúd-Arábia*	*57*	*221*
5	Hindi [hin]	India	20	182
6	Bengáli [ben]	Banglades	10	181
7	Portugál [por]	Portugália	37	178
8	Orosz [rus]	Oroszország	33	144
9	Japán [jpn]	Japán	25	122
10	Német (standard) [deu]	Németország	43	90,3
11	Jávai [jav]	Indonézia	5	84,6
12	*Lahnda [lah]*	*Pakisztán*	*8*	*78,3*
13	Telugu [tel]	India	10	69,8
14	Vietnami [vie]	Vietnam	23	68,6
15	Maráthi [mar]	India	5	68,1
16	Francia [fra]	Franciaország	60	67,8
17	Koreai [kor]	Dél-Korea	33	66,3
18	Tamil [tam]	India	17	65,7
19	Olasz [ita]	Olaszország	34	61,7
20	Urdu [urd]	Pakisztán	23	60,6
21	Török [tur]	Törökország	36	50,8
23	Lengyel [pol]	Lengyelország	23	40,0
27	Ukrán [ukr]	Ukrajna	27	37,0
40	Román [ron]	Románia	20	23,4
55	*Szerbhorvát [hbs]*	*Szerbia*	*28*	*16,4*
67	Bajor [bar]	Ausztria	4	13,3
73	Magyar [hun]	Magyarország	14	12,5
81	Cseh [ces]	Cseh Köztársaság	12	9,5
82	Lombard [lmo]	Olaszország	3	9,1
102	Német (svájci) [gsw]	Svájc	5	6,5
123	Szlovák [slk]	Szlovákia	12	5,0
124	Finn [fin]	Finnország	7	5,0

sunk, amelyek az összes nyelvre vagy legalább a valamilyen szempontból hasonló nyelvekre érvényesek. Ilyen egyezések például, hogy minden nyelvben találunk zöngés és zöngétlen hangokat, egyszerű és összetett szerkezetű beszédhangokat stb.

A mai korban „információs társadalomban” élünk. De mit is jelent ez? Egy definíciója szerint: nagyfokú információsűrűség a legtöbb polgár mindennapi életében, a legtöbb szervezetben és munkahelyen; közös vagy kompatibilis technológia használata személyes, közösségi, oktatási és üzleti tevékenységek széles körében; olyan használati eszközök, amelyek segítik az információ gyors átvitelét, fogadását és cseréjét különböző helyszínek között, a távolságuktól függetlenül.

Ez a definíció ma már igaz minden fejlett országra, sőt számos fejlődő országra is. A feltételei nem csak megvalósulnak, hanem egyre nagyobb mértékben jellemzőek

a társadalom minden területére a fejlett technológia, a szabványos megoldások, az információtechnológia és a telekommunikáció révén. Ezeket a technológiákat együttesen infokommunikációnak (Information and Communication Technologies, ICT) nevezzük. Az ICT az információátalakítás, -tárolás, -védelem, -feldolgozás, -átvitel és -kinyerés területét foglalja magában.

Ahogy a társadalmak információs társadalommá váltak, illetve ebbe az irányba haladnak, számos szükségszerű változáson mentek és mennek át. Új szakmák jönnek létre, az új eszközök megjelenésével együtt új életmódok alakulnak ki. Ezek a változások érintik a nyelvet, nyelvhasználatunkat és nyelvi kompetenciánkat is. Változik a nyelv, mivel az új eszközök és életmódok leírására új szavak jönnek létre. Jóval többet használunk kommunikációs eszközöket, mint elődeink, és legtöbb esetben a számítógépes írást használjuk az üzenetközvetítésre (elektronikus levél, SMS, chat stb.). A valóságos, szemtől szemben történő beszéddialógus mintha visszaszorulna, így a nonverbális kommunikáció értelmezésében való gyakorlatunk is csökkenhet (Fischer 1999, 88., 110., 172–176. o.).

Az infokommunikációban azonban a beszédnek kitüntetett szerepe van, mind az emberek közötti információáramlás segítésében (lásd például mobiltelefonok, IP-telefónia, videotelefonálás), mind az ember-gép kapcsolatban (Human-Computer Interaction, HCI).

A beszéd használata az ember-gép kapcsolatban, azaz amikor gépekkel beszélünk, része annak a törekvésnek, hogy a számítógépeket (és más információs rendszereket) többféle bemeneti és kimeneti eszközön keresztül tudjuk elérni, úgynevezett multimodális interakcióval (multimodal interaction): például vizuális, taktilis és audio be- és kimenetek használatával. Többek között ezekkel a kérdésekkel foglalkozik a beszédtechnológia tudománya.

2. fejezet
A beszéd komplex szerkezete

Olaszy Gábor

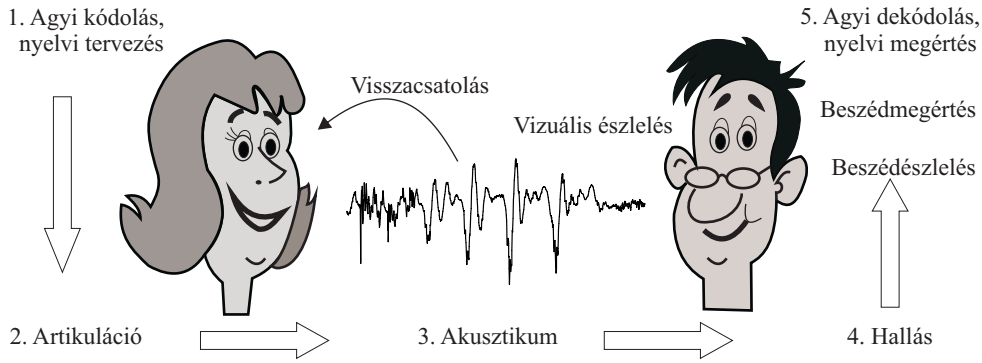
A beszélni tudás kizárólag az ember képessége. A beszéd hanghullámok révén (akusztikai úton) közvetíti az információt. A beszéd összefoglaló neve mindannak, amit egy nyelvi közösség tagjai, vagyis az ugyanazon nyelven beszélő emberek szóbeli érintkezésük során hangos közlésként mondanak. A beszéd létrejöttéhez három dolog egyidejű megléte szükséges. Ezek: a nyelvi rendszer, ami meghatározza az akusztikai elemeket, a beszélő ember ép biológiai szervei a beszédképzéshez és a beszédjel felfogásához, valamint a közvetítő közeg, amely a hanghullámot továbbítja. Ha e három elem közül bármelyik hiányzik (részlegesen vagy teljesen), a beszéd információközvetítési hatásfoka csökken (akár nullára is). A beszéd – mint akusztikai jel – komplex szerkezetét a fenti három elem határozza meg folyamatosan, a beszélés minden pillanatában. A nyelv véges számú, egymástól megkülönböztethető elemet használ, ezek a fonémák. A nyelv fonémái minimális számosságú halmazt képeznek, segítségükkel a nyelv minden szava jelentéshelyesen, de csak egyféleképpen állítható elő agyi szinten. A témáról részletesebben lásd Siptár (2006a) munkáját. Már egyetlen fonéma megváltoztatása is jelentésváltozással jár (*bál, bár; sok, sokk, sók; barom, karom; elfogad, elfogat*). Alapvetően a fonémák fizikai realizációi jelennek meg a beszédhangokban. Egy-egy beszédhangot minden ember kicsit másképpen ejt ki, mégis bizonyos határokon belül ugyanannak a fonémának soroljuk be. A fonémáknak a hangkörnyezettől független realizációit alaprealizációnak nevezik (Gósy 2004b). Ha a realizáció függ a hangkörnyezettől, akkor a fonéma variánsáról (allofón) beszélünk, azért variáns, mert nincs jelentésmegkülönböztető szerepe a nyelvi rendszerben. A beszéd fizikai rezgése mögött tehát mindig ott van a nyelv rendszere, amely meghatározza a beszéd tartalmi dekódolhatóságát. A beszédet az ember a hangképző szerveivel hozza létre. A beszéd tehát személyhez is kötött, az egyén is meghatározza a keletkezett beszéd jellemzőit, hangzását, egyrészről a saját fiziológiai jellemzőivel (például az arcméretek, a hangszalagok hossza, tömege, az egészségi állapot), másrészt a tudati döntéseivel (kifejezésforma, saját stílus, hangszínezet, saját beszédjellemzők). A beszédet továbbító közeg elsődlegesen a levegő.

A beszéd továbbításához vagy tárolásához esetlegesen használt technológiák szerint lehet más közeg is. A beszéd lényeges tulajdonsága a redundancia is, ami azt jelenti, hogy a megértést segítő komponensekből sokkal több van jelen a hullámformában, mint ami a tényleges beszédészleléshez és -megértéshez feltétlenül kellene. Ez is hozzájárul ahhoz, hogy a beszédet akkor is megértjük, amikor más, zavaró akusztikai tényezők (utcazaj, mások beszéde) is eljutnak a hallási rendszerünkhöz.

A beszéd a maga három alkotóelemével tehát komplex információtovábbító. A tudományos kutatások egyik célja, hogy feltárják ennek az információhalmaznak a sajátosságait, és a beszédjel komponenseiből kinyerjék az egyes részinformációkat. Ezért a beszéddel foglalkozó tudományok műveléséhez szorosan hozzátartozik sok határtudomány ismerete is. A nyelvtudomány írja le a nyelvek szerkezeti tulajdonságait, a nyelvi tartalom szerkesztési módjait. A fonológia a nyelvtudomány részeként a nyelvek hangrendszerét és azok sajátosságait vizsgálja. A nyelv agyi feldolgozásának modelljeivel a pszicholingvisztika (más néven pszichológiai nyelvészet) foglalkozik, amely a kognitív pszichológia egyik ága. Témái között szerepel például annak vizsgálata, hogy hogyan tárolja agyunk a nyelvi ismereteket (szavakat, szabályokat), és hogyan értjük meg a nyelvi tartalmat. A beszédszerveink működését, a beszédprodukció részletes folyamatait, a beszéd megértésének mechanizmusait nyelvi szempontból a fonetika és a pszicholingvisztika, fiziológia oldalról pedig az orvostudomány vizsgálja (ideggyógyászat, foniátria, audiológia). A hang fizikai terjedésével és az azt befolyásoló környezeti tényezőkkel az akusztika foglalkozik. Ennek egyik ága a pszichoakusztika, amely a fizikai hanginger és a szubjektív hangészlelés közötti kapcsolatokat tárja fel. A beszédjel fizikai feldolgozásával, átvitelével kapcsolatos problémakör műveléséhez matematikai, villamosmérnöki, informatikai és digitális jelfeldolgozási ismeretekre van szükség. A természetes beszédlánc (emberi dialógus) egyes elemeinek gépi megvalósítása a beszédtechnológia tudományához kapcsolódik (gépi beszédszintézis, -felismerés, -megértés, -azonosítás, -módosítás, beszédtömörítés). Az átviteli közeg térbeli és időbeli kiterjesztésével (telefonálás, hangfelvétel készítése és lejátszása) a híradástechnika foglalkozik.

2.1. A beszéd körfolyamata, az emberi dialógus

Kommunikációról információelméleti értelemben akkor beszélhetünk, ha létezik egy adó, egy vevő, egy összekötő csatorna és egy egyezményes jelrendszer. Az adó az üzenetét a jelrendszer segítségével kódolja és átküldi a csatornán, amelyet a vevő a jelrendszer segítségével dekódol. A beszéddel történő kommunikációnál az adó az éppen beszélő fél, a vevő a hallgató, a csatornát a levegő jelenti mint rezgésátviteli közeg, a jelrendszer pedig a közös nyelv. A beszéddel történő kommunikáció folyamatát öt fő részre oszthatjuk fel (2.1. ábra).



2.1. ábra. A beszéd körfolyamata az adó és vevő (dialógus) kapcsolatban

Az első a nyelvi kódolás szintje, amelyben a beszédtervezés során agyunk a közölni kívánt gondolatokat nyelvi formába önti: megkeresi a szavakat, valamint a felhasználni kívánt nyelvi szerkezeti formákat, meghatározza a szavak sorrendjét, és megtervezi az akusztikai megjelenítés részleteit, a közlendő dallamszerkezetét, a hangsúlyokat és minden mást, ami a megszólaltatáshoz szükséges nyelvi szintű elem (például gúnyos ejtést akar produkálni).

A második szakasz fiziológiai. A megtervezett nyelvi formát a motoros agykérgi területek a hangképző és artikulációs szerveket vezérlő parancsokká alakítják, majd a tüdő, a légcső, a hangszalagok, a garat, a nyelv, az állkapocs és az ajkak összehangolt aktivizálásával létrehozzuk az akusztikus formát, a beszédet.

A harmadik szakaszban (akusztikus terjedés) a közvetítő közeg, a levegő játssza a főszerepet, amely továbbítja az akusztikus jelet, egyrészről a hallgató, másrészről a beszélő füléhez. Ezzel létrejön a kommunikációs kapcsolat a két fél között, ugyanakkor a beszélő azonnali visszajelzést is kap arról, hogy mi hangzott el a saját szájából. Ez a visszacsatolás kisgyermekkorban a beszédtanulás időszakában igen fontos, a köznapi beszédtevékenység során pedig segíti az esetleges korrekciót. A keletkezett hang nem csak a levegőben terjed, hanem magában az emberi testben is. A legjelentősebb a csontvezetés, ami szintén visszacsatolásként működik. A levegő mint átviteli közeg térbeli kiterjesztését műszaki megoldások segítik (távközlés, hangrögzítés stb.). Így a beszéd körfolyamati képe kiszélesedik egymástól távol lévő személyekre is.

A negyedik szakasz (hallás) újból fiziológiai, amelyben a levegő rezgéseit a fül bonyolult rendszere idegi jelzésekké transzformálja. Itt lép be a folyamatba a vizuális csatorna, ami segíti az egymással közvetlenül beszélgető partnerek beszédmegértését. A szájmozgás, a mimika, a gesztusok mind hozzájárulnak a jobb beszédmegértéshez. A vizuális csatornának is létezik térbeli kiterjesztése, ez a videotelefonálás.

Az ötödik, záró szakasz ismét nyelvi: agyunk a beszéd agyi reprezentációjában felismeri a beszédhangokat, a szavakat (beszédészlelés), az azok közötti viszonyt, és

a jelet nyelvileg értelmezi, jelentést rendel hozzá (beszédmegértés). Ezzel megvalósult a beszélő által tervezett nyelvi információ eljuttatása a hallgatóhoz.

Minden szakasz között van bizonyos átfedés: a tervezés művelete spontán beszédprodukcióban egyidejűen zajlik a hangképzéssel; részben ezért tartunk időnként szüneteket, és ezért kell esetenként korrigálnunk magunkat. Ugyanígy a dekódolás is a hallással szinte egyidejűleg zajlik, közel 0,2 másodperc késleltetéssel, ami egy szótag átlagos hossza (Field 2003, 107. o.). Az ötödik szakaszban (szemtől szembe zajló kommunikáció esetén) a látásunk is segíti a megértést (szájról olvasás, mimika érzékelése, gesztikulálás, testmozgás értelmezése).

Az agy szerepe a fenti vezérlésekben csak részben tisztázott. A beszédképzés vezérlését döntően a bal agyfélteke irányítja. Az artikulációért 80%-ban a bal félteke úgynevezett Broca-területe a felelős. A fizikai beszédjel feldolgozásáért a bal félteke Wernicke-területe felel. A legújabb kutatások szerint a jobb félteke is részt vesz a beszéd vezérlésében. A nyelvmozgás koordinálásában például mindkét félteke egyformán aktív, az ének létrehozásánál a jobb agyfélteke jut nagyobb szerephez (Gósy 2004b).

Dialógusnak vagy párbeszédnek nevezzük azt a folyamatot, amikor a fenti körfolyamat váltva zajlik két, illetve több ember között: a beszélő közlésére a hallgató válaszol, majd újra és újra felcserélődnek a szerepek. Agyunk a beszédet nem önmagában dolgozza fel, hanem az összes érzékszervünkből kapott információt kombinálja és értelmezi, sőt ehhez a korábban szerzett összes tudásunkat is felhasználja. Segíti a nyelvi dekódolást a látás is: a verbális tartalom megértését rossz hallási viszonyok között a szájmozgás látványa támogathatja. A dialógus szereplőinek beazonosítását segíti a sztereóhallás (betájoljuk a hangforrás irányát). Közvetlen dialógusban értelmezzük a nonverbális elemeket is (a gesztusokat, a mimikát), és tudat alatt ellenőrizzük, hogy ezek összhangban vannak-e a nyelvi tartalommal. A dialógushelyzetre jellemző, hogy a felek között létezik egy hallgatólagos megállapodás, mely szerint a beszélő csak olyan tudásra épít, amivel feltételezhetően rendelkezik a hallgató is, a hallgató pedig feltételezi, hogy a beszélő olyan üzenetet fogalmazott meg, amelynek ő képes kikövetkeztetni a jelentését. A folyamatos üzenetváltásban a dialógus szereplőinek viselkedési formája szorosan hozzátartozik a beszéd körfolyamatához (reakciók, fordulatok, válaszformák stb.). A beszédkutatás témaköreibe a beszéd körfolyamatának minden eleme beletartozik, beleértve a dialógust is.

2.2. A nyelvi tudás szintjei

Ahhoz, hogy nyelvtanilag helyes mondatokat tudjunk kimondani és megérteni, többszintű tudásra van szükségünk. Az adott nyelv hangjaira vonatkozó agyi beidegződés az artikulációs és percepciós bázis. A lexikai tudásbázis a szavak ismeretét jelenti.

A szintaktikai ismeretek halmaza határozza meg, hogy a szavakat hogyan köthetjük egymáshoz, hogy értelmes mondatokat alkossanak. A szemantikai szint a jelentés értelmezésében segít. A szavaknak önálló jelentésük van (lehet több is), de az akusztikai forma dekódolása sok esetben nem elég egy mondat pontos értelmezéséhez. A szavak együttes hatásukon keresztül hozzák létre a mondat jelentését. Ez az összefüggés kiterjed a mondatok közötti térre is, mivel egy szintaktikailag helyes mondatnak néha több jelentése lehet. Például egy ilyen egyszerű mondatnak is, mint *Laci és Mari barátai jelentek meg az esküvőn.* lehet több jelentése. A említett barátok lehetnek mindkettejük barátai, vagy csak Mariéi. A helyes jelentést csak a szereplők ismeretében tudjuk megállapítani, illetve szóban érzékeltethetjük a két eset közötti megkülönböztetést azzal, hogy használunk-e szünetet a *Laci* szó után. A mesterséges intelligencia kutatások egyik fő iránya a nyelvek szemantikai szintű modellezése. A pragmatikai szint is fontos szerepet játszik az értelmezésben. Az emberi kapcsolatokban a nyelvhasználat funkciója gyakran olyan információ kérésére irányul, amihez a társadalmi környezet ismeretére, pragmatikai tudásra van szükségünk. A pragmatika tárgya az, hogy a kontextus (a párbeszéd szereplői, feltételezett szándékaik, a beszélgetés helyszíne, a társadalmi környezet ismerete stb.) hogyan járul hozzá a jelentéshez. Például egy dialógusban elhangzik a következő mondat: *Hol van az Eiffel-torony?* A válasz a pragmatikai tényezőktől (például a szituációtól) is függ. Lehet az, hogy: *Ha kimész a háztömb mögül, meglátod.*, vagy az is, hogy: *Párizsban.* Egy mondat tényleges jelentését sok esetben tehát a világról való ismereteink alapján próbáljuk meghatározni, figyelembe véve a szövegkörnyezetet és a szituációt, amelyben elhangzik.

A lexikális, a szintaktikai, a szemantikai és a pragmatikai tudásunk nyelvi kompetenciánk részét képezik, amit általában akaratunktól függetlenül használunk.

2.3. Beszédformák

A beszédnek számos megnyilvánulási formája lehetséges. Ez két szinthez kötődik, a nyelvi közösséghez (társadalmi szint) és a beszélő személyhez (egyéni szint). Társadalmi szinten a sokféle egyedi megvalósulás jellemzői a közösség számára oly módon jelennek meg, hogy a közös anyanyelvet beszélők megértik egymást. A beszédnek ilyenformán olyan invariáns jegyei vannak, amelyek társadalmi szinten fejezik ki az egységes szerkezetét, az anyanyelvi formáját. A beszélőfüggetlen gépi beszédfelismerésnél ezeket az invariáns jegyeket igyekeznek kinyerni a beszélők hullámformáiból.

Az egyén szintjén a beszéd a társadalmi szintű, sokféle realizációból kiragadott egyik egyedi forma. Ugyanakkor az egyén beszéde is sokféle megvalósulásban jöhet létre, ez a beszédet létrehozó biológiai rendszerből fakad. A beszédkutatás tehát az-

zal néz szembe, hogy a nyelv hangzó formájának megvalósulása számos paraméter függvénye (szinte végtelen a variációk sokasága). Csak a legfontosabbakat soroljuk fel: tájegység, nem, kor, egyéni tényezők, szituációk, érzelmek, témakörök. Mindmind más beszédet jelentenek. A beszélőfüggő gépi beszédfelismerőknél az egyén hangjára jellemző paramétereket használják fel az algoritmusok.

Egy másfajta osztályozást eredményez, ha a produkciós biológiai mechanizmus oldaláról közelítjük a témát. Ebből a szempontból kétféle csoportosítás lehetséges, amelyek át is fedik egymást. Ezek a spontán megnyilatkozás, ami általában dialógusszituációban jön létre, valamint a felolvasás, amikor előre meghatározott szöveget alakítunk beszéddé. A spontán beszéd a legalapvetőbb beszédforma, amit naponta használunk családunkban, munkahelyünkön, társas kapcsolatainkban. A felolvasást is számos helyen használják (hírolvasás, esti mese, vízállásjelentés stb.). A két beszédforma között lényeges különbség van. Társadalmi szinten mindkettő létezik és mindkettőnek a fizikai elemzése már a 20. század közepén elkezdődött (Shannon 1948). A beszéd fizikai átvitele (távközlés) során a spontán beszéd mint jelfolyamat volt a kutatás tárgya. A részletező fonetikai elemzésekhez pedig inkább felolvasásból származtatott, előre megtervezett beszédmintákat használtak. A 21. század elején jutott el oda a fonetika tudománya, hogy a spontán beszéd részletező elemzése is előtérbe került (Gósy–Horváth 2007).

Ebben a könyvben a felolvasásos beszéddel kapcsolatos eredmények vannak túlsúlyban. Ennek oka, hogy a spontán beszéd kutatása még csak most van felfutóban. Ugyanakkor azt is látni kell, hogy az alapkutatási eredmények – amikre támaszkodhatunk – a felolvasásos beszéd elemzéseiből állnak rendelkezésünkre. Beszédtechnológiai szempontból mindkét beszédforma fontos, és a jövő kutatásaiban bizonyosan nagyobb hangsúlyt fog kapni a spontán beszéd is (például személyre szabott hangú beszédszintetizátorok tervezése, szituációk gépi azonosítása a beszéd alapján stb.).

2.3.1. Spontán beszéd

A természetes helyzetű dialógusban másfajta beszédtervezés történik, mint egy szöveg felolvasásánál. A mondandó gondolati megformálása és a fiziológiai beszédképzés (artikuláció, hangképzés) szimultán zajlik (Gósy 2005), egyszerre két dologra figyelünk. Mi ennek a következménye? Az, hogy akaratunk ellenére hibázunk, pontatlanok vagyunk, és az esetek többségében ezt észleljük és korrigáljuk is. A jelenség oka, hogy gondolataink időnként gyorsabban változnak, mint ahogy a beszéd fizikai megvalósítása történik. Ezért lesz ténylegesen spontán ez a beszéd mind a mondanivaló, mind az akusztikai megvalósítás szempontjából. A mondanivaló tekintetében ez abban tükröződik, hogy gyakran változtatjuk meg a kimondás során a gondolatunk tartalmát (esetleg csak egy szót, egy toldalékot). Ilyenkor jönnek létre a megakadási

jelenségek, vagyis a beszélő beszédtervezési és beszédkivitelezési bizonytalanságai (Gósy 2002). A beszédképzés (a fizikai jel) szempontjából pedig az jellemzi a spontán beszédet, hogy sokkal lazább az artikuláció, a lazítási folyamatok hatására kimarad(hat)nak beszédhangok, gyakoriak a megszakítások (megakadások, korrekciók), gyakori a hezitálás, a szünettartás. Jellegzetes az is, hogy több mondat egybefolyik a kiejtésben, sok esetben nincs egyértelműen érzékelhető (és mérhető) mondathatár.

A spontán beszédnek különféle megvalósulási formáit különböztetik meg a beszédkutatásban. Ezek mind a beszédtervezési folyamat működési szintjeivel vannak összefüggésben. A teljesen spontán dialógusban minden tervezési szint működik. Vannak olyan szituációk, amikor bizonyos szintek kikapcsolhatók, illetve csak részben működnek. Ezekben az esetekben félspontán beszéddel van dolgunk. Ez valósul meg, amikor riportot készítenek. Ilyenkor a riporter kérdése már segíti a riportalanyt a válaszban. Hasonló a helyzet, amikor valamilyen konkrét tárgyról, képről kell beszélnünk, annak témaköre szűkíti a beszéd tárgyát, jobban tervezhető a mondanivaló.

A spontán beszéd módszeres kutatása csak az utóbbi évtizedekben gyorsult fel, részben a beszédtechnológiai követelmények hatására (társalogni akarunk a gépekkel). Elmondható, hogy a spontán beszéd képviseli a legbonyolultabb beszédformát a verbális kommunikációban; vizsgálata, feldolgozása is bonyolult, mivel a kutató csak a beszédhullámra támaszkodhat (nincs semmi előzmény arra vonatkozóan, hogy a beszélő mit fog mondani, csak a pillanatnyi szituáció és a beszédhelyzet alakítja a produkciót). A kutatásokat egyrészről a megfigyeléses adatgyűjtés jellemzi (megakadási jelenségek hallásalapú gyűjtése), másrészről hangzó anyagok nyelvi és fonetikai vizsgálata (Gósy–Beke 2010).

A spontán hangzó anyagok gyűjtését akadályozza, hogy a kutatónak nehéz hozzáférnie az igazi spontán beszédhez mint rögzített vizsgálati anyaghoz, ugyanis jó minőségű felvételeket nehéz spontán módon, terepen készíteni. A nehézség kiküszöbölésére kialakultak olyan módszerek, amelyekkel közelíteni lehet a spontán szituációt stúdióban is. Ilyen például, amikor olyan beszélgetést rögzítenek, amelyben a felvétel tervezői tudatosan irányítják a beszélőket (de azok nem tudnak róla, tehát verbális reakciójuk spontánnak tekinthető). Társalgási hangfelvételek anyagai szintén jól elemezhetők. A kutató segítségére lehet a média is, ahonnan szintén rögzíthetők riportok, beszélgetések. A mérési lehetőségek szélesítésére és egységesítésére már készítenek spontánbeszéd-adatbázisokat, amelyekről később szólunk. Összefoglalva elmondhatjuk, hogy a spontán beszéd kutatása egyre szélesedik világszerte, hiszen a kihívás nagy. A következőkben bemutatunk néhány kutatási területet, témát a magyarral kapcsolatosan.

A megakadási jelenségek a spontán beszéd jellegzetes elemei. A magyar megakadási jelenségek gyűjtése a 21. század első éveiben kezdődött meg. Az első magyar, jegyzeteléses technikával gyűjtött Nyelvbotlás-korpusz 2004-ben jelent meg (Gósy 2004a), és 5139 tételt tartalmazott. A munka a megakadások fajtáit rendszerezte, több megakadási kategóriát is definiált. Ilyenek például: freudi elszólás; morfológi-

ai, szintaktikai hiba, ezen belül az újraindítás nélküli morfológiai változtatás vagy az ismétlés toldalékjavítással; összeolvasztás (kontamináció); téves szótalálás; téves szókezdés; téves szóhasználat (malapropizmus); „a nyelvem hegyén van" jelenség; változtatás; újraindítás; egy korábbi elem nem odaillő beépülése (perszeveráció); későbbi hangzó(k) korábbi megjelenése (anticipáció); hang, szótag, szó felcserélése (metatézis); egyszerű nyelvbotlás; többféleképpen osztályozható jelenség. Ezekről részletesen a fent idézett kötetben olvashatunk, ahol a megfigyelt jelenségeket kategóriák szerint csoportosították. A jelenség szöveges megadása mellett külön oszlopban olvashatjuk a szándékolt közlést is, így jól érzékelhető a megakadás konkrét története is. A gyűjtés folyamatos, minden évben közreadják az új adatokat az MTA Nyelvtudományi Intézete által évente kiadott *Beszédkutatás* c. tanulmánykötetekben (2005: 761 további tétel; 2006: 388; 2007: 244; 2008: 444; 2009: 117).

A spontán beszéd elemzésének problémás témaköre a tagolás is. Ebben a beszédformában elvész a klasszikus értelemben vett mondatszintű tagolódás. Gósy (2003, 20. o.) szerint: „Gyakorlatlan beszélők a verbális nyelv tagolásainak tudatos megvalósítására nem képesek, azt fiziológiai tényezők, például a légzés, sokkal inkább szervezik, mint a szöveg tartalmi és formai egységeinek tudatos jelzései." Ő vezeti be a virtuális mondat fogalmát, amellyel a spontán beszéd tagolódási rendszerét vizsgálja. Kísérleti eredményei szerint a virtuális mondatok léteznek, mind a beszélő spontán beszédében, mind a hallgató percepciójában. A virtuális mondatok szerinti tagolódást jellegzetes akusztikai paraméterek jelenléte biztosítja, azonban lényeges szerepe van a szintaktikai és a szemantikai szinteknek is.

A spontán beszéd egyik jellegzetes eseménye a szünet is (lehet hangos hezitálás is), ami segít időt nyerni a beszélő számára, hogy folytassa a mondandóját. A szünetnek számos megjelenési fajtája van.

A spontán beszéd lejegyzése nem annyira egyértelmű, mint a felolvasásé. A spontán beszéd artikulációja sok egyszerűsítést tartalmazhat (a felolvasott beszédhez képest), ez téma- és szituációfüggő. A lejegyzési mód tervezésekor szembesülünk ezekkel az egyszerűsödésekkel, ami dilemma elé állítja a kutatót. Mit jegyezzek le (amit mondott, vagy amit jelentene)? Hogyan jegyezzem le (például a redukált hangokat)? Mindezek mellett számolnunk kell a beszédjelhez társuló egyéb hangjelenségekkel is, ezek lejegyzése több szempontból is fontos lehet. Egyrészről azért, hogy szét lehessen választani a mondanivalót és az azon kívüli akusztikai részeket (hümmögés, krákogás, cuppantás), amikről a beszélőnek esetleg nincs is tudomása. Másrészt fontos lehet abból a szempontból is, hogy tanulmányozzuk a spontán beszéd teljes akusztikai szerkezetét és az egyénre vonatkozó beszédviselkedési formákat.

2.3.2. Felolvasásos beszéd

A kísérleti fonetikai kutatások zöme – már a kezdetektől fogva – tudatosan megtervezett beszédmintákkal dolgozik. Ez azt jelenti, hogy a szöveget előre összeállítják, majd felolvastatják. A beszédkutatásban és a fejlesztéseknél sok esetben szükséges a tudatos tervezés, hogy a megfelelő beszédanyag álljon rendelkezésre. Ezt egy adott szöveg megtervezése jelenti. Ilyen módszert ma is széles körben alkalmaznak az alapkutatástól egészen az alkalmazott fejlesztésekig. A felolvasott beszédnek több előnye van a spontánnal szemben. Az egyik, hogy előre meg lehet határozni a témakört, a szöveg pontos tartalmát és ezzel az elhangzó hangsorozatot. A másik előny, hogy a várható akusztikai produktum tisztább lesz artikulációs szempontból, mivel a beszélőnek nem kell a nyelvi tervezési fázist elvégeznie, készen kapja a szöveget, amit beszéddé kell alakítania. Hátrányként sorolható fel, hogy a beszéd hangteste nem tükrözi a hétköznapi, egymás közötti kommunikáció laza ejtésformáit. A felolvasott beszédet illetően kutatásfejlesztési szempontból kétféle szintet kell megkülönböztetnünk: szólistafelolvasás, illetve szövegfelolvasás. A szólistákból általában szegmentális információkat lehet kinyerni (hangok, hangkapcsolatok akusztikai tulajdonságai), míg a szövegfelolvasásból szupraszegmentális tényezőket is (dallam, ritmus, hangsúlyozás, hangszínezet). Néhány példán keresztül megpróbáljuk érzékeltetni a felolvasott beszéd kutatási terét.

Az elsőben azt mutatjuk be, hogy milyen módon változik a szó időtartama, egyrészről a szótagszám, másrészről a beszédforma függvényében (2.2. ábra). A felol-

2.2. ábra. A szó időtartamának változása a szótagszám függvényében különböző beszédprodukciókban

vasásos beszéd adatai Olaszy (2006b), a spontán beszédé Gósy (2004b) munkájából származnak. Látható, hogy a szavak hosszának alakulása a szótagszám függvényében közel lineáris emelkedést mutat. A szövegfelolvasásnál és a spontán beszédnél a szóidőtartamok között nincs nagy eltérés. Érdekes, hogy a két görbe keresztezi egymást. Az egy szótagú szavak a spontán beszédben hosszabbak, mint a folyamatos felolvasásban, a két szótagúaknál megegyeznek a szóhosszúságok, majd a szótag-

szám növekedésével a felolvasás javára egyre nő az eltérés. Feltételezésünk, hogy spontán beszédben az egy szótagú szavakat bizonytalanabbul, lassabban mondjuk, a több szótagúaknál pedig a szó elindítása után igyekszünk minél előbb túl lenni a szó kimondásán, nagyobb tempót valósítunk meg, mint a felolvasásnál. Ez a feltételezés tudományosan még nincs igazolva.

A szavakból álló listák felolvasása teljesen más kategória, mint a szövegfelolvasásé, itt a bemondó szinte csak a szegmentális szerkezet megvalósítására koncentrál, nem kell értelmezési, hangsúlyozási, ritmikai feldolgozást végeznie. A szólisták felolvasásánál lassabb artikulációs tempót valósítanak meg a bemondók, mint a szövegfelolvasásnál. Ez is mutatja, hogy a beszédprodukciókban az időszerkezeti elemek igen széles határok között mozoghatnak (például egy szöveg felolvasásakor a felsorolásban lassulhat a beszédtempó).

A második példánkban tegyük fel, hogy a hármas mássalhangzó-kapcsolódások artikulációs és akusztikai részleteit akarjuk tanulmányozni. Mivel az ilyen hangkapcsolódások ritkán fordulnak elő a beszédben, nem célszerű rábízni a véletlenre, hogy mikor találunk ilyen hangkapcsolódást, tehát nem célszerű sok mondatot felolvastatni, illetve ilyen anyagban keresni (hírek, regény). Célzott szóanyag gyűjtésével sokkal egyszerűbben képezhetünk kutatási nyersanyagot (lásd a `http://fonetika.nytud.hu/cccc` szólistáját).

A harmadik példa egy fejlesztési feladathoz kapcsolható. Ha készítenünk kell egy olyan beszédszintetizátort, amelyik számokat képes felolvasni emberi hangszínezetű, jó minőségű beszéddel, akkor a feladat megoldására olyan hullámforma-adatbázist kell létrehozni, amelyikben minden számelem szerepel hangsorként is és dallam, hangsúly, ritmus szempontjából is. Kérdés, hogy mit olvastassunk fel? Célszerű speciális szöveganyagot tervezni, amely fonetikai szempontok szerint összeállított több számjegyű számok listájából áll. Ezt kell felolvastatni (lásd a 8.2.1. fejezetet).

3. fejezet
Fiziológiai, fizikai alapok

A beszéd mint a nyelv hangzó formája biológiai rendszerek között működik. A beszédkeltés, a beszéd felfogása, megértése és értelmezése az emberhez kapcsolódó tevékenység. A beszéd ugyanakkor megvalósulási formájában egyfajta összetett levegőrezgés, tehát fizikai jellemzőkkel leírható. Ebben a fejezetben sorra vesszük a beszéddel kapcsolatos fiziológiai jellemzőket, majd a beszédjel fizikai leírását. Külön alfejezetet szánunk a kettő összekapcsolásának, vagyis annak, hogy a fizikai jel milyen érzetet vált ki a hallási rendszerünkben.

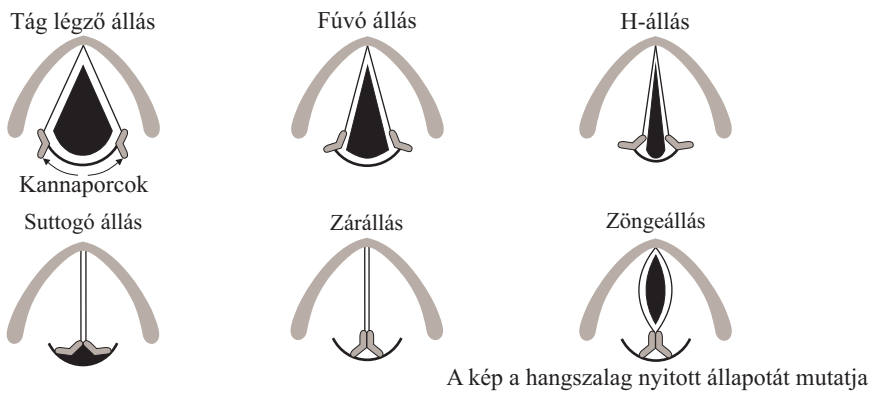
3.1. A beszédképzés folyamata

Olaszy Gábor

A beszédkeltés alapvető fiziológiai szervei a tüdő, a légcső – szubglottális tér –, a gége – glottális tér –, a garat, a száj- és orrüreg – szupraglottális tér. Működésüket az agy irányítja. A tüdő biztosítja a kifelé irányuló levegőáramot, a gégeműködés határozza meg, hogy zöngés vagy nem zöngés hangot ejtünk. A gége az elsődleges hangképzési terület, vagyis itt van a beszéd forrása. A gégében keletkezett hang tovább halad a gége feletti üregekbe (száj, orr), ahol rezgésformája módosul. Ez a módosulás a passzív és aktív artikulációs szervektől függ. A passzív szervek a fogak és a szájpad, az aktívak a nyelv, a nyelvcsap (uvula), az állkapocs és az ajkak. Az artikulációs mozgásokat az aktív szervek egymástól független mozgatásával végezzük, ezzel változtatjuk a szájüreg pillanatnyi formáját, keresztmetszeti tényezőit. A nyelvcsap irányítja a levegőáramlás útját, hogy csak a szájüregen, csak az orrüregen, vagy esetleg mindkettőn keresztül haladjon a levegő. A beszédhangok az artikulációs csatornában kapják meg a végleges hullámformájukat.

3.1.1. Gégeszintű hangképzés

A tüdőből kipréselt levegőáramlás a szubglottális útvonalon (légcső) a gégébe jut. Itt dől el, hogy a levegőáramlás milyen formában halad tovább a száj- és orrüreg felé. A hangszalagok állása határozza meg a lehetőségeket (3.1. ábra). A gége működését a fonetikai szakkönyvek részletesen tárgyalják (Gósy 2004b), itt csak vázlatosan tekintjük át. A működési mechanizmust az ideális hangképzés szempontjából tárgyaljuk, amikor a szervek a rendeltetésük szerint működnek. A speciális, illetve rendellenes, szokatlan működésekről (glottalizáció, pillanatnyi működési zavar, rekedtség, suttogás) csak a legfőbb jellemzőket mondjuk el. Alapvetően a kannaporcok állása határozza meg a gége pillanatnyi működési formáját, vagyis a hangszalagok állását.



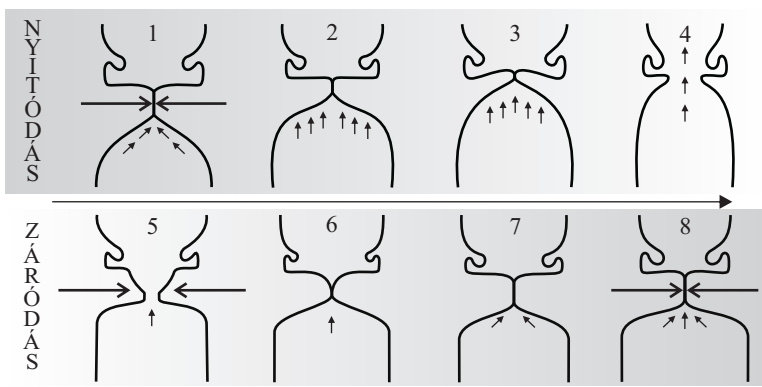
3.1. ábra. A hangszalagok jellemző állásai lélegzésnél és beszéd esetén a gége keresztmetszeti síkjára merőleges irányból szemléltetve

A beszédtevékenység során a fúvó és a zöngeállás váltakozik, vagyis zöngétlen és zöngés hangokat ejtünk (kivétel a H-állás, ami egyetlen beszédhanghoz köthető).

A beszéd indítása minden esetben a tág lélegző állásból indul, vagyis levegőt veszünk. Ezután az esetek többségében vagy a fúvó állásra, vagy a zöngeállásra váltunk át. Például a *sás* szó ejtésekor az első hangnál fúvó állásban, a másodiknál már zöngeállásban működnek a hangszalagok, a harmadiknál ismét fúvó állásban. Zöngeállásnál a hangszalagok a nyomásviszonyok változásának a függvényében kváziperiodikusan szétnyílnak, majd összezáródnak (nem rezegnek). Ennek a mozgásnak a kialakításában döntő szerepe van a hangszalagokat összezáró izmoknak. E nyomásváltozás létrehozásában a hangszalagok tehát kényszerített mozgást végeznek. A zöngehang tehát egy kváziperiodikus levegőnyomás-változás, amely a gégében keletkezik, a tüdő és a hangszalagok közreműködésével, és a hallási rendszerünk számára hallható frekvenciával rendelkezik. Az összezáró izmok pillanatnyi erőha-

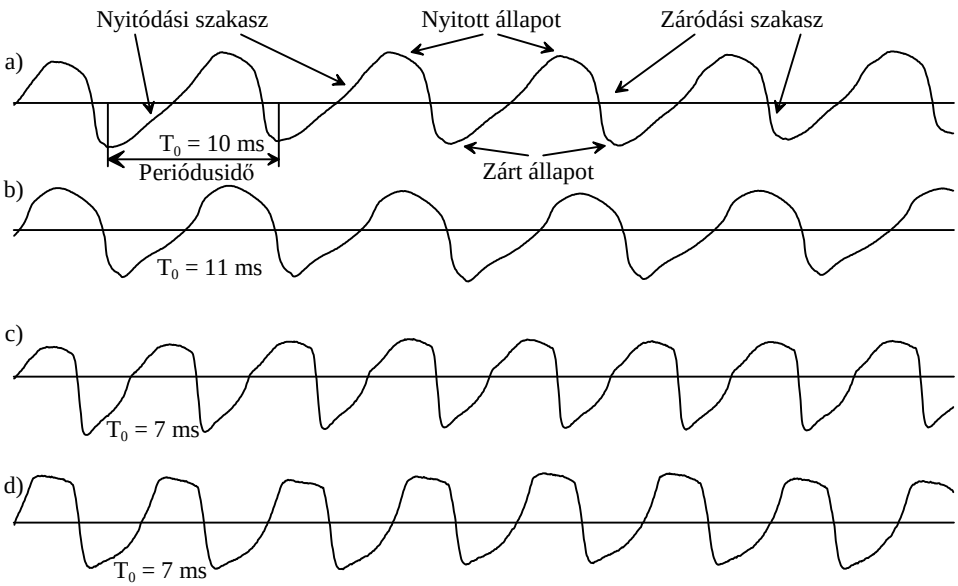
tása határozza meg a rezgés frekvenciáját (ezért tudjuk a hangmagasságot tudatosan változtatni).

A zönge (fonáció) kialakulásának folyamata két alapvető lépésre osztható, a nyitódás és a záródás folyamatára. Egy teljes periódus eme két szakaszát a 3.2. ábrán mutatjuk be 4–4 részelemre bontva. Az (1) jelű állapotba a hangszalagokat összezárjuk (lezárjuk a levegő útját). Az összezáródás pillanatában a nyomásviszonyok egyenlőek a hangszalagok felett és a szubglottális üregben. Mivel a tüdőből levegőt préselünk a légcsőbe, a nyomás növekedni fog a hangszalagok alatti (szubglottális) térben (2. kép), és a hangszalagok fokozatosan felfelé mozdulnak el, de még nem nyitódnak szét (3. kép). Amikor a nyomás eléri az összezáró izmok erőhatását, akkor a hangszalagokat a levegő függőlegesen felfelé szétfeszíti (4. kép), a hangszalagok kicsit felfelé és kicsit vízszintesen elmozdulva szétnyílnak (ekkor a levegő kiáramlik egy löket formájában). Amikor kiegyenlítődik a nyomás, akkor kezdődik a záródási szakasz (5. kép). A hangszalagokat az izmok ismét összezárják (6. kép), a zárás feszítettsége fokozódik (7. és 8. kép). Ez a folyamat ismétlődik egészen addig, amíg



3.2. ábra. A hangszalagok mozgásának folyamata egy zöngeperiódusban. A vízszintes nyilak jelzik az összezáró izmok erőhatását

zöngés hangot ejtünk. Az ismétlődés ideje a periódusidő (T_0). Ez férfiaknál 8–12 ms, nőknél 4–6 ms, gyermekeknél pedig még ennél is rövidebb. A zönge rezgésformáját sok tényező befolyásolja. Elsősorban a gége méretei, a hangszalagok hossza és rugalmassága, az izmok feszítő tartása, a hangszalagok nyálkássága, valamint a beszélő akarata, pillanatnyi érzelmi állapota. Ebből következik, hogy alapvetően más képet mutat a gyermekek, a nők és a férfiak zöngerezgése is, de az egyéni zöngerezgési kép is más és más ezen kategóriákon belül is. A képzett beszédhang is befolyásolja a zöngeképződést (3.3. ábra). A fentiekből következik, hogy folyamatos beszéd során sohasem regisztrálhatjuk ugyanazt a zöngerezgésformát. Ugyanez vonatkozik a periódusidőre is. A hangfolyamban két egymás melletti zöngeperiódusnak az időtartama csak megközelítőleg lesz egyforma, ezért a beszéd zöngés szakaszaira



3.3. ábra. Példa a hangszalagoknál létrejövő hangnyomás időfüggvényének változatosságára. Férfi beszélő zöngéje a *á* és *i* betű kiejtése során a) és b); női beszélő zöngéje ugyanezen két hangnál, c) és d)

a kváziperiodikus jelzőt használják. Többek között ez az állandó ingadozás teszi emberi hangzásúvá (nem gépiessé, monotonná) a beszédet. Ezt a kváziperiodikusságot fontos megvalósítani a mesterséges beszéd-előállításnál is.

A zönge frekvenciája adja a beszéd alapfrekvenciáját (alaphangját), az amplitúdója pedig a hangerősséget. Az alaphang elsősorban a beszélő egyén hangfekvését határozza meg (mély, magas), de ezenfelül több szerepe is van a beszédprodukcióban. Változtatásával hozzuk létre a beszéd dallamát, fontos szerepe van a hangsúlyozásban, a mondat modalitásának kialakításában, az érzelmek kifejezésében. A beszélő széles tartományban (50 Hz–500 Hz) tudja változtatni az alapfrekvenciát. A beszélő neme jó közelítéssel meghatározható az alapfrekvenciájának átlagából. A férfiakra 100 Hz-es, a nőkre a 180 Hz-es átlag jellemző, a gyermekeké pedig több száz Hz-es is lehet.

A zönge amplitúdójától függ, hogy valaki hangosan, vagy halkan beszél. Az egészséges zöngehang az alapja a tiszta, karakteres beszédnek. A zöngeképzést számos tényező zavarhatja (betegség, ellazuló ejtés, akaratlagos változtatás, nem kellő nedvességellátás, lepedékképződés stb.). Zavar esetén a hang fátyolos, recsegős, levegős lehet, illetve az egyes zöngétlen mássalhangzók zöngésekké válhatnak, például öregkori motoros vezérlési gyengeségből adódóan. Jellegzetes jelenség a beszédben, amikor a zönge irregulárissá válik, ami abban nyilvánul meg, hogy a periódussorozat amplitúdója nem egyforma, amplitúdóingadozás lép fel, esetleg egyes periódusok ki is maradhatnak, vagy nagyon kicsi lesz az amplitúdójuk. Az ilyen zöngeképzést

glottalizációnak nevezik (Bőhm et al. 2008), általa rekedtessé válik a beszéd néhány szótagja. Jellemző előfordulása a mondatok befejező szakaszában van, amikor már a levegő fogyóban van, és a beszélő már nem tud telt zöngét képezni. A folyamatos rekedt ejtés is a zöngeképzés speciális esete, de ez a teljes beszédfolyamra jellemző, állandósult állapot. Ilyenkor a hangszalagok nem tudnak kváziperiodikusan működni, csak esetlegesen nyílnak ki, csapódnak össze 20–25 ms-onként. A zöngeképzés rezgésképének vizsgálatával betegségek is diagnosztizálhatók.

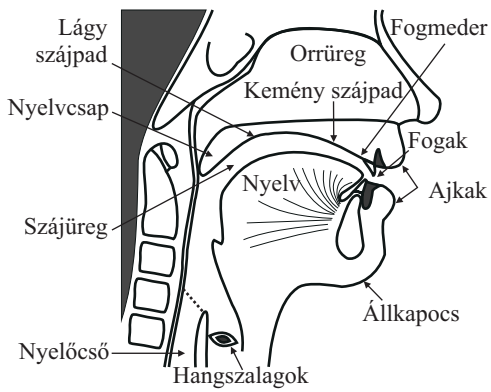
A zöngés hangképzés speciális esete, amikor lezárjuk az orális és nazális csatornát, de zöngeképzés van (*baba*). A példaszó mássalhangzójában a zöngehang létrejön, azonban az artikulációs csatornából a hang időtartama alatt nincs levegőkiáramlás. Az így keletkezett hangot nevezik fojtott zöngének. A fojtott zönge létezésének időben korlátot jelent, hogy a lezárt szájüregben meddig lesz kisebb a nyomás a szubglottálisnál. Amennyiben a két nyomás egyenlővé válik (a zár miatt), a fojtott zönge megszűnik, a hangszalagok nyíló-záródó mozgása leáll. A fojtott zönge tehát csak rövid időtartamú lehet (max. 150 ms), hosszabb ideig nem tudjuk képezni a nyomásviszonyok és a szájüregi méretek korlátozó tényezői miatt. Hogyan halljuk meg akkor a fojtott zöngét? Az arcüreg átveszi a belső szájüregi rezgést és így megrezegteti a külső levegőt. Ezt veszi fel a mikrofon, illetve ezt halljuk közvetlen dialógusban. Az előbbiekből következik az is, hogy a fojtott zönge amplitúdója kicsi. A fojtott zönge frekvencia-összetevőire az artikulációs csatorna nincs lényeges befolyással, mivel nincs hangkiáramlás (közel áll a zönge rezgésformájához).

Abban az esetben, amikor nincs fonáció (zöngétlen hangok ejtésekor), a fúvó állásnak megfelelően a gége nem jelent akadályt a tüdőből kipréselt levegőnek. Akadályokat az artikulációs csatornában hozhat létre a beszélő akár zár, akár rés formájában. Ilyenkor turbulens áramlások, néma szakaszok, illetve lökéshullámok alakulhatnak ki az aktív artikulációs szervek mozgáskombinációinak eredményeképpen. A suttogó beszédnél a hangszalagok a suttogó állást veszik fel, zöngeképzés nincs, a gégében keletkezett turbulens áramlás adja az eredetileg zöngés hangok alapját is. Ennek a végleges frekvenciaszerkezetét ugyanúgy alakítja ki az artikulációs csatorna, mint a zöngés hangoknál (ezért ismerjük fel a suttogott magánhangzókat).

Összefoglalva tehát azt mondhatjuk, hogy a beszéd hangképzésében a hangszalagok többfunkciós feladatot látnak el. Normál beszéd esetén 3–3 paraméterrel számolhatunk, ha a zöngés, illetve a zöngétlen hangkategórára vetítjük a képzést. A zöngés hangokhoz tartozik alapként a zöngeállás, valamint annak működéséből fakadó további két fizikai jellemző, a zönge alapfrekvenciája és amplitúdója. A zöngétlen hangok ejtésekor az idevonatkozó három hangszalagállás valamelyike valósul meg a képzés során, a fúvó állás (a zöngétlen réshangok képzéséhez), a zárt helyzet (a zöngétlen zár- és zár-rés hangok zárszakaszában), illetve a h-állás.

3.1.2. Az artikulációs csatorna

A hangszalagok és az ajkak közötti hangképzési területet nevezik artikulációs csatornának, toldalékcsőnek. Három üregrészből áll a garat (pharinx), a szájüreg (cavum orale) és az orrüreg (cavum nasale). A toldalékcsövet szupraglottális üregrendszernek is szokták nevezni.



3.4. ábra. A hangképző és az artikulációs szervek

Az artikulációs csatorna hossza jellemzően rövidebb a nőknél, mint a férfiaknál (14, illetve 17 cm átlagosan), térfogata átlagosan 130 cm^3, illetve 170 cm^3 (Gósy 2004b). Az artikulációs csatornában formálódnak ki a beszédhangok. A passzív és az aktív artikulációs szervek együttese határozza meg a pillanatnyi beszédhangokra jellemző artikulációs konfigurációkat (3.4. ábra). Az aktív artikulációs szervek közül a nyelvnek van a legnagyobb szerepe. A nyelvhegy, -hát és -perem külön-külön szerepet kap az artikulációban, A nyelv akadályokat is képezhet a mássalhangzók esetében. Az aktív artikulációs szervek mozgásformái a következők. A nyelv előre-hátra, illetve fel-le mozog, akadály nem jön létre a szájüregben. Ezek a mozgások főleg a magánhangzók képzésénél játszanak szerepet. Amennyiben a nyelv akadályt, illetve rést képez, akkor mássalhangzót ejtünk. Az ajkak kerekítésének, illetve széthúzásának a magánhangzók kialakításánál van szerepe. Az ajak egymással való összezárása, illetve a fogakkal való érintkezése a mássalhangzókhoz kapcsolható. Az állkapocs nyitódási foka a magánhangzók képzésében játszik szerepet.

A mássalhangzók jellemzéséhez képzési helyeket és képzési módokat adnak meg a fonetikában. Mindkét kategóriát az artikulációs szerveinkkel hozzuk létre. A magyar mássalhangzók képzési helyei a következők: két ajak összezáródása (bilabiális képzési hely), alsó ajak-felső fog érintkzése (labiodentális), felső fogmeder-fog érintése a nyelvheggyel, illetve résképzés (dentialveoláris), a fogmeder érintése a nyelvheggyel, illetve résképzés (alveoláris), a kemény szájpadlás érintése a nyelvháttal,

illetve résképzés (palatális), a lágyszájpad és a nyelvhát hátsó részének érintkezése, illetve résképzés (veláris), a garatüreg a hangképződés helye, az artikulációs csatorna teljesen nyitott (faringális). A képzési módok a következők: amennyiben a nyelvvel rövid időre lezárjuk a szájüreget, akkor zárhangot képezünk, ha valamilyen rés alakul ki az artikulációs csatorna bármely részén, akkor réshangot ejtünk. A kettő kombinálódása a zár-rés hang, amelyik az előbbi két artikulációs mozzanat egymásutániságából jön létre. A pergetett hang a nyelvhegy pergetésével alakul ki a dentialveoláris területnél. A magánhangzószerű mássalhangzókat – szerkezetük miatt – közelítő hangoknak nevezzük.

Fontos szerepe van egyes beszédhangok kialakításában az orrüregnek is, itt jönnek létre a nazális hangok. Az orrüreg méretei nem változnak, ezért ezt passzív szervnek tekintjük. A nyelvcsap zárt, félig zárt, illetve nyitott állapota határozza meg, hogy az orrüreget mennyire vonjuk be a hangképzésbe. Az artikuláció során mind az artikulációs csatorna, mind pedig a szubglottális tér mint rezonátor vesz részt a hangképzésben. A szubglottális terület is hatással van a hangképzésre, bár szerepe sokkal kisebb, mint a toldalékcsőé (Csapó et al. 2009).

A toldalékcső keresztmetszeti és alakváltozásai azt eredményezik, hogy a zönge viszonylag egyszerű rezgéséből a toldalékcső végére (ajkak) bonyolult rezgésformák jönnek létre, mindig az adott artikulációs konfigurációnak megfelelően más és más. Mit tekintünk egy beszédhangra jellemző hullámformának? Vizsgáljuk itt csak a hosszan tartható hangokat. Egy-egy jellemző artikulációs állásból származó keresztmetszethez és alakhoz tartozó, az ajkaknál kisugárzott hanghullám felel meg egy-egy ilyen beszédhangnak (izolált ejtésben). A magyarban másodpercenként 10–20 esetben jönnek létre ilyen jellemző konfigurációk (folyamatos artikulációs mozgással). Ez azt jelenti, hogy 10–20 beszédhangot ejtünk másodpercenként (a köznapi beszédre a 13–14 hang/s a jellemző érték).

3.1.2.1. A koartikuláció

Az artikuláció a beszédképzés során folyamatos. Két beszédhangot a képzés szintjén a koartikuláció köt össze, vagyis az a mozgássor, ami a két hangra jellemző artikulációs konfiguráció között zajlik le. Ebből következik, hogy a beszédhangokra jellemző rezgésformák folyamatos átmenettel csatlakoznak egymáshoz, ezeket a beszédszakaszokat nevezik hangátmeneteknek. A hangátmenet tehát a koartikuláció következménye. A hangátmeneti részek legalább olyan fontosak, mint a hangokat képviselők. A két elem együttesen biztosítja percepciós rendszerünknek a megfelelő nyelvi információt. A koartikuláció elméletéről és vizsgálati módszereiről Hardcastle–Hewlet (1999) munkájában olvashatunk.

A koartikulációnak két szintjét különbözteti meg Gósy (2004b), a fonológiait (például a zöngésségi hasonulás) és a fonetikait (például, amikor egy beszédhang képzési

helye eltolódik). A beszédtechnológiában mindkettővel kell foglalkozni, bár mindig az adott fejlesztés határozza meg, hogy milyen mélységben. A koartikulációs mozgások végrehajtása időt igényel. Ha bonyolult mozgásokat kell végrehajtani, hogy az egyik hang artikulációjából eljussunk a másikig, akkor az eltelt idő hosszabb, mint ellenkező esetben. Ez kimutatható az egyes beszédhangokra jellemző időtartamokban, amit a specifikus időtartamok hordoznak (5.1.1.1. fejezet). Az egyes fonémák hangalakja megváltozhat a koartikuláció függvényében, ez azonban csak néhány artikulációs pozícióhoz köthető. A bilabiális képzési helyű [m] megváltozik labiodentális képzési helyűre [ɱ], például a *ha**m**vas, ká**m**for* szavakban. Ez azt jelenti, hogy a labiodentális réshang artikulációja előrefelé hat, és kissé magához igazítja a bilabiális képzést. Ez minden esetben így van. A koartikulációs hatásokról részletes fonetikai adatok találhatók Gósy (2004) munkájában. A magyar hangkapcsolódások koartikulációs mozgásaiból létrejövő frekvenciaszerkezeti változásokat bemutatjuk a függelékben minden mássalhangzóra C1VC1 kapcsolatokban.

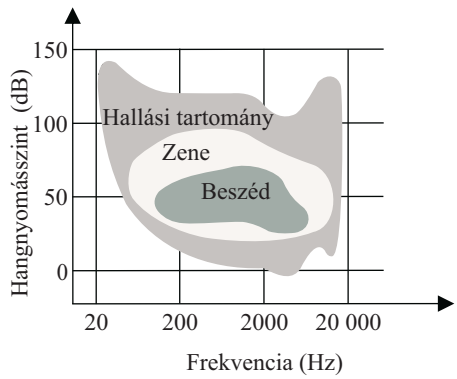
3.1.2.2. A gégeműködés és az artikulációs csatorna viszonya

A gégeszintű jel és az artikulációs csatorna egymástól függetlenül is működnek (gondoljunk arra, hogy a gégeeltávolításon átesett ember is tud beszélni egy külső, a gégéhez szorított, rezgést létrehozó készülékkel, ami hasonló hatást érhet el, mint a zöngeképzés rezgése, noha nem emberi hangszínezettel). A gége szintjén létrejött levegőáramlás által keltett hang a beszédképzés alapkomponense, nélküle nincs természetes beszéd. Az artikulációs csatorna hatása erre a gerjesztő hangra épít. A beszéd az artikulációs csatornában nyeri el végleges rezgésformáját. A két komponens ugyanakkor függetlennek tekinthető egymástól. A kisugárzott beszédhangok a jellemző frekvenciakomponenseiket az artikulációs csatorna üregrendszerében kialakuló rezonanciafrekvenciáknak, valamint a létrehozott szűkületek helyének és alakjának köszönhetik. Ha megváltozik a gége szintjén a levegőáramlásból keletkezett hang, akkor annak hatása jelentkezni fog a kisugárzott beszédhangban is (például rekedt hang, levegős hangadás, suttogás). Mindezek a tények fontosak a beszéd fizikai feldolgozása és bizonyos beszédtechnológiai elemzési módszerek szempontjából. Látni fogjuk, hogy a gerjesztett szűrő modelljében (3.3. fejezet) a gerjesztést egy külön áramkör képviseli, az artikulációs csatornát imitáló hangolható szűrőrendszer pedig egy másik áramkör.

3.2. A hallási folyamat

Vicsi Klára

Az emberi hallórendszer komplex akusztikai, mechanikai, hidrodinamikai elektromos jelátalakító, idegvezetési és agyi szerkezet. Nemcsak számos ingerre reagál, hanem a beszédhangot és az alaphangot (hangmagasságot, hangfekvést), sőt, a hangforrás irányát is precízen beazonosítja. A hallási funkciók nagy részét a fül végzi el, ám a legutóbbi kutatások kihangsúlyozták, mennyire függ a hallás attól az adatfeldolgozástól is, amely a központi idegrendszerben történik.



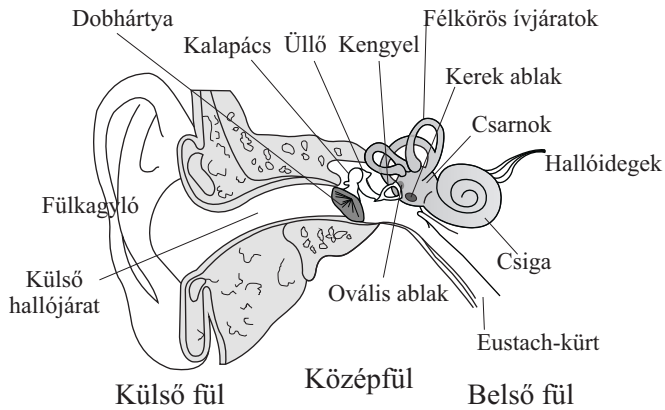
3.5. ábra. A zene és a beszéd érzékelési tartománya a teljes hallási tartományon belül

A hallási funkciók. A hangnyomásingerek azon tartománya, amelyre a fül reagál, igen széles. Egy különösen hangos hang energiatartalma körülbelül millió milliószor (10^{12}) nagyobb, mint a leggyengébb, de még hallható hangé. Bizonyos hangfrekvenciákon a dobhártya kimozdulása kisebb, mint 10^{-8} mm, ami körülbelül egytizede a hidrogénatom átmérőjének. Becslések szerint a belső fülben található nagyon finom hártya, az alaphártya rezgéseinek amplitúdója – amely a hallóidegeknek továbbítja az ingert – még ennél is közel százszor kisebb (Békésy 1960).

A hallás frekvencia- és intenzitásbeli érzékelési tartományát a 3.5. ábra mutatja. Az ábrán látható, hogy a beszédhullámok intenzitás-frekvencia területe lényegesen szűkebb, mint maga a hallható hanghullámok érzékelési tartománya. A hallás frekvenciatartománya egyénileg változik; ritka az olyan személy, aki a teljes 20–20 000 Hz-es hallási tartományt képes hallani. A fül viszonylag érzéketlen az alacsony frekvenciájú hangokra; például 100 Hz-en durván 1000-szer kisebb az érzékenysége, mint 1000 Hz-en. A magas frekvenciájú hangok érzékenysége kisgyermekkorban a legnagyobb, és az élet folyamán fokozatosan csökken, tehát egy felnőtt nehezen hallja meg a 10 000, vagy 12 000 Hz-nél magasabb hangokat.

3.2.1. A fül szerkezete

A fület a működési funkciók függvényében három részre szokás osztani: külső fül, középfül, belső fül (3.6. ábra).



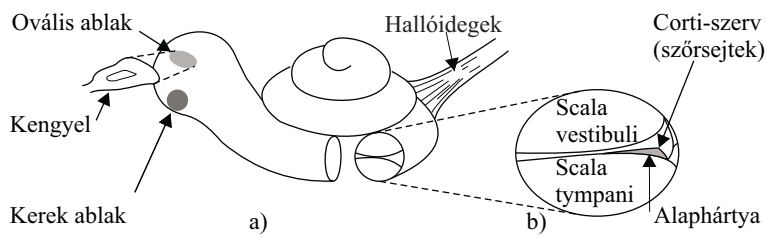
3.6. ábra. A fül vázlatos metszete

A *külső fül* a fülkagylóból és a hallójáratból áll, amelyet a dobhártya zár le. A fülkagyló segít a hangok összegyűjtésében, és hozzájárul azon képességünkhöz, hogy meghatározzuk a hangforrások irányát. A külső hallójárat rezonátorcsőként működik, növeli a hallás érzékenységét a 2000–5000 Hz-es tartományban.

A *középfül* a dobhártyából és a hallócsontokból (kalapács, üllő és kengyel) áll. A dobhártyát, amely kör alakú, és sugárirányú rostokból épül fel, megfeszítve tartják a feszítőizmok. A dobhártya a bejövő hanghullámok hatására elmozdul a nyomásingadozás függvényében, és ezt a mozgást a hallócsontocskák továbbítják a belső fülbe, a hártyaszerű ovális ablakon keresztül. A dobhártya szintjén megjelenő nyomásváltozás felerősödve érkezik az ovális ablakhoz. Egyrészről a hallócsontocskák emelőrendszerként működnek, mintegy 1,5-szörös erőmegsokszorozást hoznak létre. Másrészről mintegy húszszoros nyomásnövekedést okoz a dobhártya és az ovális ablak területe közötti különbség (kisebb felületre ugyanannyi erő jut, nagyobb nyomást eredményezve). A csontocskák másik funkciója, hogy védjék a belső fület a nagyon erős hangoktól és a hirtelen nyomásváltozástól (robbanás, üstdob). A nagy erejű hang kétféle izomzatot aktivizál: az egyik a dobhártyát szűkíti, a másik a belső fülben elhúzza a kengyelt az ovális ablaktól. Ezt az erős hangokra való reakciót akusztikus reflexnek nevezik. Minthogy a dobhártya légmentesen lezárja a külső- és a középfül közötti részt a külvilágtól, szükség van némi nyomáskiegyenlítés biztosítására, hiszen a dobhártya csak akkor tud rendeltetésszerűen működni, ha a külső és belső fülben a nyomás ugyanakkora. A külső légnyomás megváltozásakor (például gyors magasságváltozás esetén csökken a légnyomás) a középfül nyomásviszonyait

is ehhez kell igazítani. Ez az Eustach-kürtön keresztül történik, amely összeköti a középfület a garatüreggel. Figyelemre méltó, hogy mindezek a középfülbeli funkciók térben mindössze akkora helyet foglalnak el, mint egy kis méretű kockacukor.

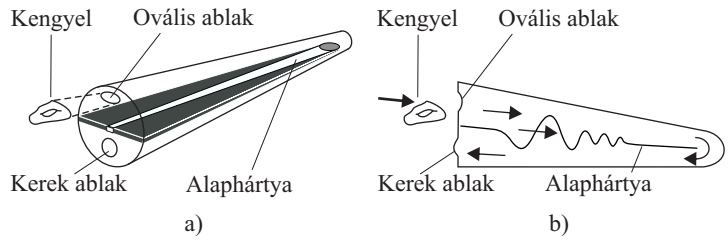
A bámulatosan összetett *belső fül* félkörös járatokból és a csigából áll. A félkörös járatok nagyon kevéssé, vagy egyáltalán nem működnek közre a hallásban; ezek a test vízszintes-függőleges detektorai, amelyekre az egyensúlyozáskor van szükség. A hang feldolgozásában a csigának van fontos szerepe, itt a mechanikus rezgések megfelelően kódolt idegi impulzusokká alakulnak át. A csiga kiterített hossza 3–4 cm. A csiga jelbemenete az ovális ablak hártyája. A kengyel erre a hártyára adja át a rezgéseket. A csiga folyadékkal van tele. Az ovális ablak hártyájának rezgései továbbterjednek a folyadékban. Mivel a folyadékok összenyomhatatlanok, gondoskodni kell arról, hogy a nyomáshullám terjedhessen a csigában. Erre szolgál a kerek ablak rugalmas hártyája. Amikor az ovális ablak hártyája befelé mozdul el, a kerek ablaké kifelé.



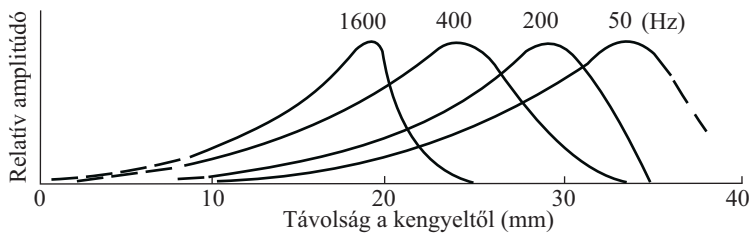
3.7. ábra. A csiga és a csigából kivágott rész sematikus diagramja

Az alaphártya a csigát két részre osztja a *scala vestibuli* és a *scala tympani* járatokra (3.7. ábra). Az alaphártya teljes hosszában nyugszik a finom és komplex Corti-szerv. Ez a „hallás székhelye”, amely több sor apró szőrsejtből áll.

Annak érdekében, hogy megértsük, hogyan rezeg az alaphártya, nézzük meg a csiga kiterített és egyszerűsített változatát a 3.8. ábrán. A csiga itt egy kúposan elvékonyodó hengerként jelenik meg, amelyet két részre oszt az alaphártya. A henger vastagabb végénél van az ovális és kerek ablak, amelyeket az alaphártya térben elválaszt. A csiga keskeny végén található egy lyuk, ami összeköti a felső és alsó üreget, és szabad áramlást biztosít a folyadéknak a kerek ablak felé. Amikor a kengyel az ovális ablak felé mozdul el, hidraulikus nyomáshullámok kerülnek továbbításra a scala vestibuli kamrában, hullámokat indukálva az alaphártyában. A magas frekvenciájú hangok az alaphártya legnagyobb amplitúdójú kimozdulását az ovális ablak közelében okozzák, ahol az alaphártya a legkeskenyebb. Az alacsony frekvenciák a legnagyobb amplitúdójú hullámokat az alaphártya másik végénél hozzák létre, ott ahol az alaphártya széles és laza (lásd a 3.9. ábrát). Így jön létre a kezdeti, még nem nagy felbontású frekvenciaanalízis a csigában.

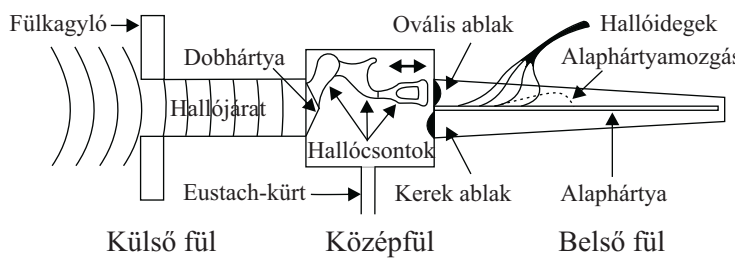


3.8. ábra. (a) Egy kiterített csiga sematikus diagramja, amely az alaphártyát és az ovális és kerek ablakot szemlélteti. (b) Amikor a kengyel az ovális ablakhoz nyomódik, a nyomás lüktetése továbbterjed a csiga folyadékában és eljut a kerek ablakig, miközben hullámokat kelt az alaphártyában



3.9. ábra. Az alaphártya hely szerinti kimozdulási amplitúdója a frekvencia függvényében (Békésy 1960)

Amikor az alaphártya kimozdul, a Corti-szerv szőrsejtjeinek szőrcsillói a hullámzás hatására elhajolnak, ezáltal idegi impulzusokat hoznak létre, amelyek az idegpályákon az agyvelőbe továbbítódnak. A keletkező impulzusok sűrűsége főleg az intenzitástól függ, de kevésbé annak frekvenciájától is. A teljes hallási mechanizmus sematikus felépítését a 3.10. ábra illusztrálja.



3.10. ábra. A fül sematikus reprezentációja, amely a teljes hallási mechanizmust illusztrálja. A külső fülből érkező hanghullámok mechanikus rezgéseket okoznak a középfülben, és végül idegi impulzusokká alakulva továbbítódnak az agyba

A hangok egy része nem a levegő rezgésével, hanem a koponya, az arccsont rezgéseivel jut el a belső fülbe. Ezt nevezik csontvezetéses hallásnak. A csontvezetés általi hallás fontos szerepet játszik a beszédben, főleg a saját beszédben. A zümmö-

gő hangok, vagy a fogak koccanása szinte teljesen csontvezetés által hallhatóak. (Ha ujjunkkal befogjuk a fülünket, így állva útját a levegőnek, a zümmögés hangosabban fog szólni.) Amikor beszélünk, vagy énekelünk, két különböző hang jut el a hallószervhez, az egyik a csontvezetés, a másik a levegővezetés útján. Ismert, hogy a saját hangját hangfelvételről visszahallgató személy más hangot hall, mint amit a saját megszokott hangjának tart. Ez attól van, mert a mikrofon csak a levegő útján érkező hangot veszi fel.

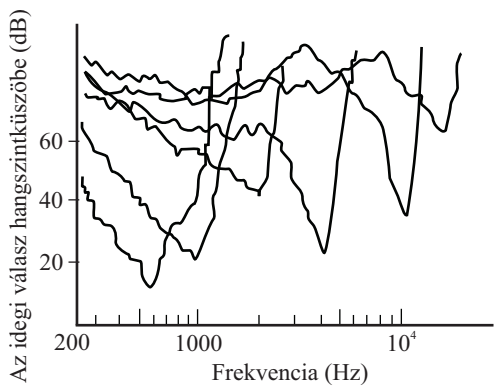
Békésy György magyar kutató nagymértékben járult hozzá, hogy a hallási folyamatot jobban megértsük. Állati és emberi holttestek fülében található csigák működését vizsgálta (Békésy 1960). Ezért a kutatásért 1961-ben Nobel-díjat kapott. Saját tervezésű mikroeszközökkel képes volt feltárni az alaphártya egy részét és azt mikroszkóp alatt vizsgálta. A csigafolyadékot elvezette, és sós oldattal, valamint porrá tört szén- és alumíniumszuszpenzióval cserélte fel. Amikor a csiga hangingert kapott, fényt látott felvillanni a szuszpenzált porból, így fedezte fel a hullámmozgást az alaphártyában. Az alaphártya rezgésének illusztrálására Békésy a csiga számos mechanikus modelljét építette fel.

3.2.2. Jelfeldolgozás a hallórendszerben

A jelfeldolgozás a hallórendszerben két részre osztható: az egyik egység a periférikus hallórendszer (maga a fül), ahol a levegőben és a csontokban terjedő mechanikai rezgések elektromos impulzusokká alakulnak át, a másik a hallási idegrendszer az agyban.

Periférikus hallórendszer. A fül által érzékelt hangnyomásváltozásból keletkezett idegi impulzusokat a hallóidegrendszer különböző szakaszai tovább alakítják. Kísérletileg igazolt, hogy egy igen vékony elektródát helyezve a csigából az agy felé tartó hallóidegbe, a hallóidegrost egy egyedülálló ideg szálában haladó elektromos jel felvehető (Tasaki 1954). Minden hallóidegszál egy bizonyos hangnyomás- és frekvenciatartományon belül reagál. Minden egyes idegszálnak van tehát egy karakterisztikus frekvenciája (CF, characteristic frequency), amelyen maximális érzékenységet mutat. Az idegszál hangolási görbéje aszimmetrikus, a CF-nél magasabb frekvenciákon meredek, ellenkező irányban kevésbé az. A mért elektromos jel egy impulzussorozat, mindegyik impulzuscsúcs megfelel az alaphártyához kapcsolódó szőrsejt kisülésének. A keletkező impulzussorozatok szorosan korrelálnak az alaphártya mechanikus rezgésformájával, vagyis az alaphártya kimozdulási amplitúdóival mintegy 4000–5000 Hz-es frekvenciáig. Az alaphártya kimozdulása folytonos, tehát egy adott színuszos gerjesztésre a maximális kimozdulás a frekvencia függvényében az alaphártya egy adott pontján történik, de a maximális

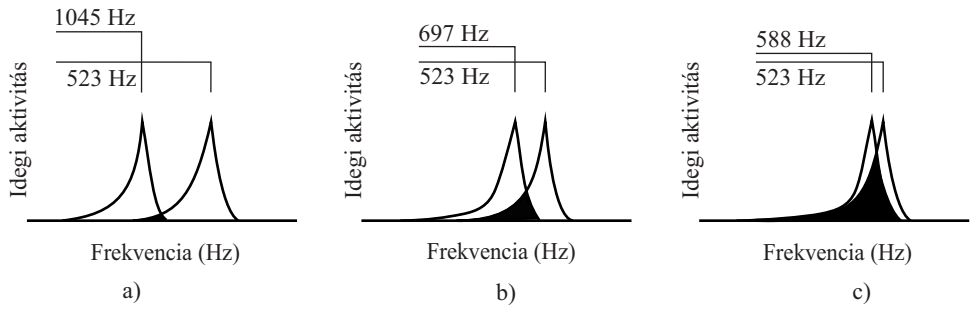
kimozdulási pont környezete is kimozdul, gerjesztve a környező szőrsejteket is. A 3.11. ábra hat különböző idegszál *hangolási görbé*it, vagyis a frekvenciaváltozásra adott válaszait mutatja be egy macska hallószervében (Kiang–Moxon 1974). Ha 500 Hz-es és 90 dB-es gerjesztést adunk a fülbe, akkor mind a hat idegrostban válaszimpulzusok jelennek meg. A maximális érzékenység a gerjesztőjel által az alaphártyán okozott maximális kitérési helynek megfelelő szőrsejtben és az onnan kiinduló idegszálban van.



3.11. ábra. Hat különböző idegszál hangolási görbéi egy macska hallószervében

Az olyan kifinomult technikák, mint a lézerfénnyel való próbálkozás (Khanna–Leonard 1982) és a Mössbauer-effektus használata (Johnstone–Boyle 1967) esetén kiderült, hogy az élő állatok esetében az alaphártya-kimozdulás sokkal élesebb frekvenciaválaszt eredményez, mint az elpusztult állat 3.9. ábrán bemutatott elmozdulásán láthattuk. Rhode–Robles (1974) úgy találták, hogy a halál beállta után néhány órával az alaphártya mechanikai reakciója 10–15 dB-t csökken, a kimozdulási maximum frekvenciája csökken, és a reakciógörbe kiszélesedik. Tény, hogy az alaphártya mechanikus frekvenciaválasza élő csigánál összehasonlítható az idegrostokban észlelt hangolási görbékkel. Van azonban bizonyíték arra is, hogy az idegi hangolási görbék élesednek, ahogy az ingerület az idegpályákon az agy felé tart. Ha megfigyeljük az idegszálon szinuszos gerjesztés esetén keletkezett impulzuskisüléseket, akkor észrevehetjük, hogy ha impulzuskisülések vannak, akkor azok mindig a szinuszhullám kitérési maximumánál vannak, de nem minden periódusban. A kisülések közötti idő lehet egy, két vagy több periódusnyi. A helyzet egy kicsit még bonyolultabb, amikor az inger egy összetett hang, mégis úgy találjuk, hogy az idegimpulzusok mintái a hallószervben pontos információkat szállítanak az ingerhang frekvencia-spektrumáról. Vegyünk egy ingert, amely tiszta C_3 (523 Hz) és C_4 (1046 Hz) hangokból áll, oktávnyi távolságban azonos intenzitásban egymástól. Idegi hangolási görbéik (vagy frekvenciaválaszgörbéik), amelyeket a 3.12(a) ábra mutat be, nagyon kis mértékű át-

fedést mutatnak, tehát nagyon kevés szőrsejt reagál mindkét frekvenciára egyszerre, mivel az alaphártya kimozdulási amplitúdói, amelyek a szőrsejteket gerjesztik, távol esnek egymástól. Így az egyik komponens feldolgozása az agyban csak nagyon kevéssé függvénye a másik jelenlétének.



3.12. ábra. Idegi frekvenciaválasz-görbék azonos intenzitású tiszta szinuszos hangpárokhoz. Ha a frekvenciaintervallum csökken, a görbék átfedése nő

Ahogy a két komponens közötti intervallum csökken, a helyzet megváltozik. Az alaphártya kimozdulási amplitúdói mind több és több átfedést mutatnak, tehát a szőrsejteket mind nagyobb számban ingerli mindkét komponens (3.12. ábra (b) és (c) részei).

Kritikus sávok. Az emberi fül felbontása a frekvenciatartományban az úgynevezett kritikus sávokkal írható le. Amikor két tiszta hang frekvenciában olyan nagyon közel áll egymáshoz, hogy jelentős átfedés jelenik meg az alaphártya kimozdulási amplitúdógörbéin, akkor ugyanazon kritikus frekvenciasávon fekszenek. A kritikus sávok megfeleltethetőek a csiga frekvenciafelbontó képességének, és fontos szerepet játszanak a percepcióban: ha ugyanis fülünket egyszerre több hang éri, és ezek egy kritikus sávon belül vannak, akkor intenzitásuk a fizikai törvényszerűség szerint összegződik, és nem észleljük őket különálló hangokként. A kritikus sávok frekvenciahatárai kimérhetők úgy, hogy egy keskeny sávú zaj frekvenciahatárait fokozatosan szélesítjük. Kezdetben azt tapasztaljuk, hogy az észlelt hangosság növekedése a fizikai törvényszerűség szerint történik. A kritikus sáv határához érve viszont a hangosságérzet hirtelen megnő. A sávok szélessége függ a sávközép-frekvenciától! A sávszélesség 500 Hz alatt közel állandó, 100 Hz. 500 Hz fölött a sávszélesség a sávközép-frekvencia növekedésével nő, megközelítőleg a sávközép-frekvencia 20%-a. Az emberi hallásra 24 kritikus sáv jellemző, a 20 Hz és 15 500 Hz közötti hallástartományban. Értékei a 3.1. táblázatban láthatók (Zwicker–Fastl 1990). A hallási érzékelés leírásában a kritikus sáv koncepció igen jelentős, sok modell és hipotézis alapja. Éppen ezért a kutatók létrehoztak egy, a hallási frekvencia érzékelésre jellemző kritikus sávszélesség arányú skálát, az úgy-

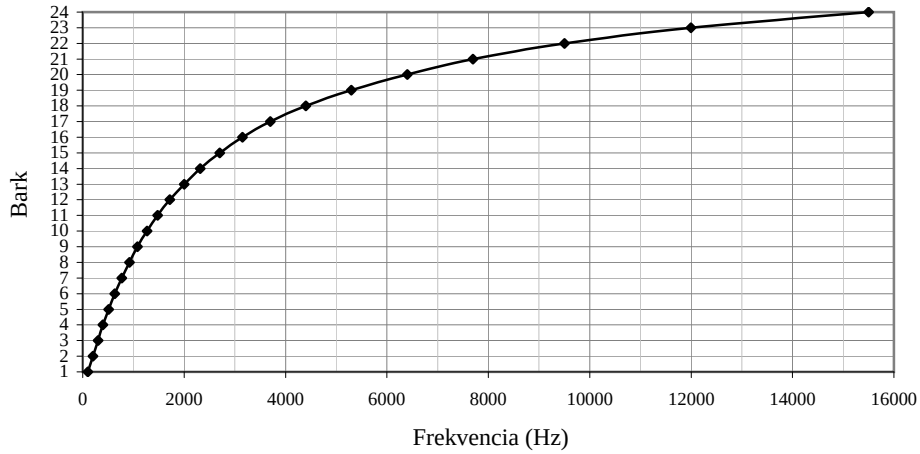
3.1. táblázat. A kritikus sávok jellemző adatai: kritikus sávarány (mértékegysége a [Bark]), annak megfelelő alsó és felső frekvenciahatár, sávközép-frekvencia és sávszélesség

Kritikus sávarány [*Bark*]	Frekvenciahatár [*Hz*] alsó	felső	Sávközép-frekvencia [*Hz*]	Sávszélesség [*Hz*]	Kritikus sávarány [*Bark*]	Frekvenciahatár [*Hz*] alsó	felső	Sávközép-frekvencia [*Hz*]	Sávszélesség [*Hz*]
1	0	100	50	100	13	1720	2000	1850	280
2	100	200	150	100	14	2000	2320	2150	320
3	200	300	250	100	15	2320	2700	2500	380
4	300	400	350	100	16	2700	3150	2900	450
5	400	510	450	110	17	3150	3700	3400	550
6	510	630	570	120	18	3700	4400	4000	700
7	630	770	700	140	19	4400	5300	4800	900
8	770	920	840	150	20	5300	6400	5800	1100
9	920	1080	1000	160	21	6400	7700	7000	1300
10	1080	1270	1170	190	22	7700	9500	8500	1800
11	1270	1480	1370	210	23	9500	12000	10500	2500
12	1480	1720	1600	240	24	12000	15500	13500	3500

nevezett kritikus sávarányú skálát. Ez a skála azon a tényen alapszik, hogy a hallási rendszerünk a széles sávú zajokat kritkus sávszélességben elemzi. Amennyiben van egy olyan kritikus sávszélességű skála, ahol a sávhatárok illeszkednek, mint ahogy az a 3.1. táblázatban látható, akkor 24 sávval a teljes hallás átfedhető. A kritikus sáv határpontok tehát adott frekvenciáknak felelnek meg, amint azt a 3.13. ábra mutatja. A határpontok nem azt jelentik, hogy a kritikus sávok csak a két határpont között léteznek, inkább azt, hogy ezek a sávok képesek folyamatosan eltolódni egy olyan skála mentén, amelyet a találkozási pontok határoznak meg. Azt a skálát, amelyet így hozunk létre, *kritikus sávarányú skálának* nevezik, értéke 0-tól 24-ig tart a hallástartományban, egysége a bark, amelynek nagysága a sávhatárpontok sorrendjének felel meg.

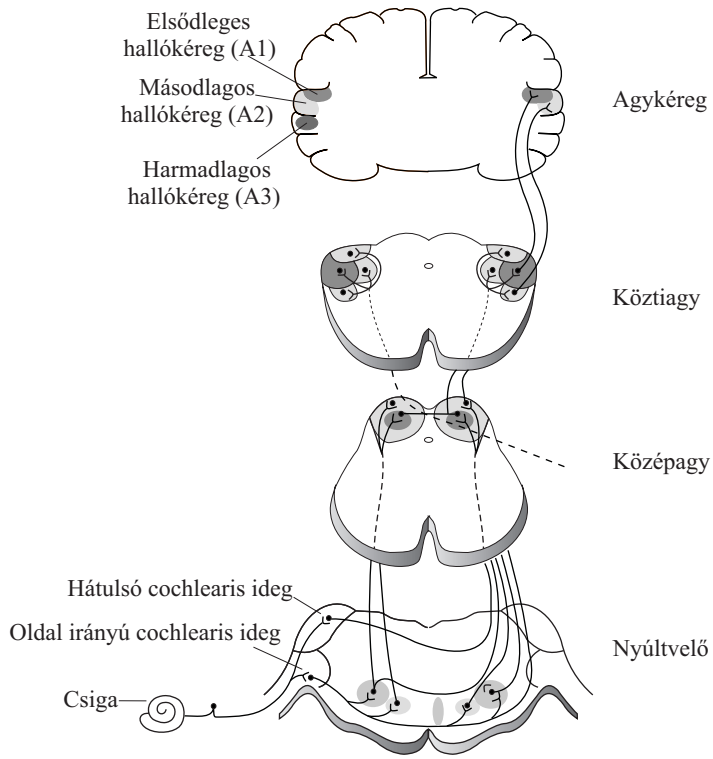
A kritikus sávoknak, valamint a csiga tonotópiás felépítését tükröző kritikus sávarányú hangmagasságskálának igen fontos szerepe van az emberi beszédpercepciós folyamatok leírásában, hiszen – ellentétben a frekvencia- és térhangalapú mértékegységekkel – kifejezésre juttatják azt a tényt, hogy az emberi hallás érzékenysége frekvenciafüggő (Fletcher 1940, Plomp 1976, Zwicker et al. 1957). Minden egyes kritikus sáv adatgyűjtési egységnek számít az alaphártyán. Egy kritikus sáv 1,3 mm-es hosszúságú, és kb. 1300 neuront tartalmaz (Scharf 1970).

A kritikus sávarány szoros kapcsolatban van a hangmagasság érzetoldali skálájával, a *melodikus hangmagasság (mel) skálával*, amelyet a 3.13. ábrán mutatunk be, és részletes leírását a következő fejezetben adjuk meg. Most annyit jegyzünk meg, hogy abban az esetben, ha az ingeroldali (lineáris) frekvenciaskálát és az érzetoldali, kísérleti úton létrehozott melodikus hangmagasságskálát 131 Hz-en 131 melnek feleltetjük meg, akkor az így rögzített mel-skála esetén az emberi hallást 2400 mel fedi le. A kritikus sávarányú skála esetén az emberi hallást 24 bark fedi le, így 1 bark a kritikus sávszélességen 100 mel.

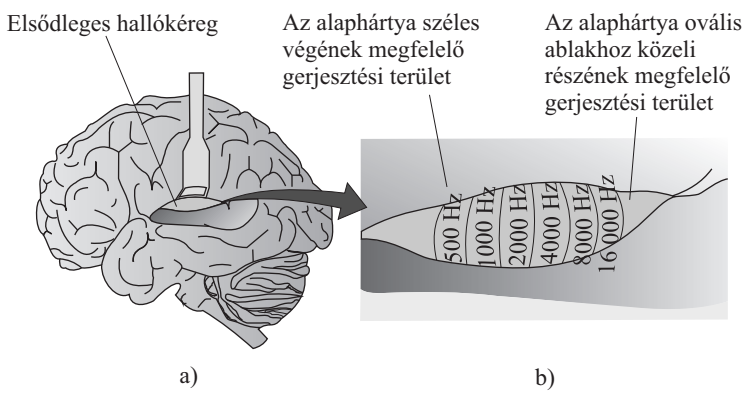


3.13. ábra. A kritikus sávarány (bark) a frekvencia függvényében

A hallási idegrendszer. A csigából kilépve egy belső szőrsejtből 20 idegszál indul el. Idegi kapcsolódásokon keresztül halad az információ az idegszálakban az agykéreg felé. Az idegszálakban az impulzussűrűség nyugalmi helyzetben kb 1 és 100 Hz közötti, gerjesztőjel esetén a gerjesztőjel intenzitásának megfelelően 800 Hz körüli érték mentén ingadozik. Egy impulzusszerű kisülés 1 ms időtartamú, néhány μV nagyságú. A hallóideg körülbelül 30 000 afferens (felfelé vezető) és efferens (leszálló) idegszálból áll, amelyeknek körülbelül 6%-a efferens, azaz a központi idegrendszerből a perifériás idegrendszerbe továbbít információt. Az efferens sejtek által kiváltott izomösszehúzódás aktívan befolyásolja az alaphártya viselkedését, például növelni tudja az alaphártya egyes frekvenciákra való fogékonyságát, az afferensek pedig receptorsejtekként működnek. Az idegszálak összeköttetéseken keresztüli útját,vagyis a hallópályát a csigától az agykéregig a 3.14. ábra mutatja. Mind az afferens, mind az efferens ágak a nyúlt agyvelő és a középagy szintjén is keresztezik egymást. Jobbkezesek esetén a bal fülbe beadott jel 80–85%-a jobboldali agyfélteke homloklebenyébe fut és viszont. A két homloklebeny nem szimmetrikus: a bal féltekén főként időbeli megfejtés, beszédfeldolgozás történik, a jobb félteke főként a térbeli információ, a színképi megfejtés helye (zene). Elektronikus és mágneses agyi válaszok vizsgálata alapján ma azt tartjuk, hogy az egy féltekei hallókéreg 3 elkülönült részre (Auditory szint, A) van felosztva: Az elsődleges hallókéreg (A1): tonotopikai szerveződésű, ami azt jelenti, hogy e hallókéreg különböző területei más-más frekvenciára érzékenyek (lásd 3.15. ábra). Ma a kutatók azt tartják, hogy az agynak ez a része az, ahol az alapfrekvencia és a hangosság meghatározása történik. A másodlagos hallókéreg (A2): a dallam és a ritmus feldolgozásáért felelős, valamint a beszédben a beszédhangok feldolgozásáért (Wernicke-terület). A harmadlagos hallókéreg (A3): a teljes zenei összbenyomásért felelős.



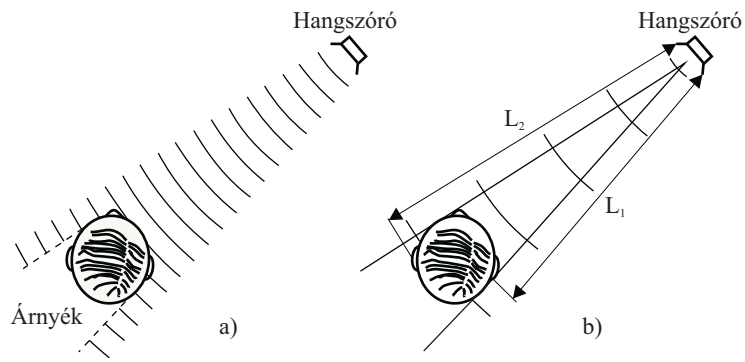
3.14. ábra. A hallópálya felépítése



3.15. ábra. Az elsődleges hallókéreg tonotopikai szerveződése

Kétfülű hallás és lokalizálás. A kétfülű hallás legfontosabb előnye a hangforrás érzékelésének képessége. Bár a lokalizáció bizonyos mértékig egyfülü hallás útján is lehetséges, a kétfülű hallás nagymértékben megnöveli a hangforrás irányát meghatározó képességünket.

1876-ban Rayleigh kísérleteket mutatott be, hogy meghatározza az ember hanglokalizáló képességét különböző frekvenciák esetében. Úgy találta, hogy az alacsony frekvenciájú hangokat nehezebb lokalizálni, mint a magas frekvenciájúakat. Rayleigh magyarázata szerint a fej egyik oldaláról jövő hang intenzívebb hangzást eredményez a hangforrás felöli oldalon lévő fülben, mint a másikban, mert a fej „hangárnyék”-ot vet a magas frekvenciájú hangok esetében (3.16. ábra a) része). 1000 Hz-en a hangszint mintegy 8 dB-lel nagyobb a hangforráshoz közelebbi fülben, ám 10 000 Hz-nél a különbség 30 dB is lehet. Alacsony frekvencián azonban az árnyékeffektus kicsi, mert a nagy hullámhosszú hanghullámok elhajlanak a fej körül.



3.16. ábra. Egy hangforrás irány szerinti lokalizálása. (a) 4000 Hz feletti frekvenciáknál a lokalizálás a két fül közötti intenzitáskülönbség szerint történik. (b) 1000 Hz alatti frekvenciáknál a lokalizálás az L1 és L2 hangútvonalak közötti terjedési időkülönbség alapján megy végbe

Az alacsony frekvenciájú hangok valamivel kisebb pontossággal lokalizálhatóak, mint a magasabb frekvenciájúak. 1907-ban Rayleigh egy második elméletet ajánlott a lokalizálást illetően az alacsony frekvenciájú effektusok magyarázatára. Az egyik oldalról jövő hang az egyik fület a másik után éri el, vagyis fáziskülönbség van köztük, amint azt a 3.16. ábra b) része mutatja. Rayleigh óta számos kísérlet bizonyította a tényt, hogy az 1000 Hz körüli és ennél mélyebb frekvenciák számára a lokalizáció főleg a két fül közötti fáziskülönbség érzékelésén keresztül jelenik meg, vagy pedig a beérkezési idő különbségének detektálása alapján. 4000 Hz felett az intenzitáskülönbség szerint történik a lokalizálás. 1000 és 4000 Hz között a lokalizáció pontossága csökken, nagy hibaaránnyal 3000 Hz körül, demonstrálva, hogy a két mechanizmus nem fedi át egymást jelentősen. Magas frekvenciáknál (5000 Hz körül, illetve afölött) a fülkagyló segít a hang lokalizálásában, különösen az elölről, vagy hátulról jövő hang megkülönböztetésében, mert egy kicsit nagyobb hatékony-

sággal vesz fel hangokat elölről. Néhány állatnak megvan az a képessége, hogy a hang felé tudja fordítani a fülét, az embernek viszont az egész fejét el kell fordítani ahhoz, hogy megváltoztassa a fülkagyló orientációját.

A hang lokalizációjának egy fontos következménye az úgynevezett elsőbbségi effektus (néha Haas-effektusnak is nevezik), amelyet a hangok szobában való lokalizálására használ az ember. Ha hasonló hangok érkeznek körülbelül 35 ms-on belül (0,035 s), a hangforrás nyilvánvaló iránya az az irány, ahonnan az első érkező hang jön. A fül automatikusan feltételezi, hogy ez direkt hang, és az egymást követő hangok egyszer vagy többször visszaverődnek.

A külső fül átviteli függvénye (a hangforrástól a dobhártyáig) függ a dobhártya impedanciától, a hallójárattól, a fülkagyló és a fej együttes hatásától. Tehát az átviteli függvény a hang beesési irányától függően a fülkagyló, a fej, a váll frekvenciafüggő hatása miatt változik. Azt az átviteli függvényt, ami leírja az átvitelt különböző beesési irányokból a szabadtérből a hallójárat tetszőleges pontjáig (a dobhártyáig), a *külső fül komplex átviteli függvényének* nevezzük. Ezen HRTF (Head Related Transfer Function) függvényeket a fejhez rögzített koordináta-rendszerben mérik (Blauert 1997).

A szelektív térbeli hallásnak, valamint az efferens idegpályák működésének köszönhető az a képességünk, hogy bizonyos irányból érkező hangokra fokozottan tudunk koncentrálni (szelektív térbeli hallás). Egy nagyobb társaságban a magas alapzaj ellenére képesek vagyunk meghallani, mit mond beszélgetőpartnerünk, mert a környezetből érkező hangingereket képesek vagyunk kikapcsolni. Ezt a jelenséget koktélparti-effektusnak is szokás nevezni. Jól bizonyítja a szelektív térbeli hallás szerepét az a tény, hogy egy, a partiról készült monohangfelvételen nem tudjuk ugyanazokat a beszédfoszlányokat kiszűrni, azaz a környezeti zajt elnyomni.

3.3. A beszéd fizikai jellemzése

Vicsi Klára

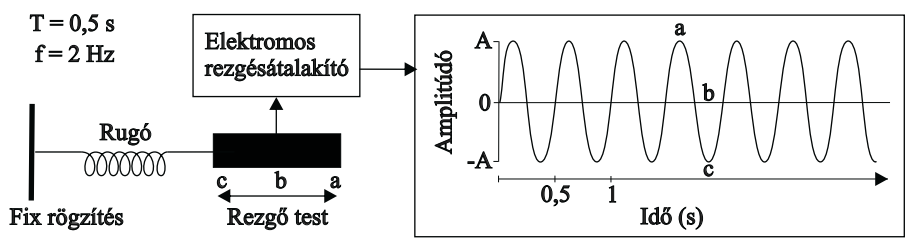
A beszéd az ajkakat elhagyva tekinthető kész jelnek. Ahogy kisugárzódik a közvetítő térbe, rezgése terjedni kezd. Ilyen szempontból ugyanolyan rezgés, mint bármely más akusztikai jel (zene, sziréna, zörej), tehát alkalmazhatjuk rá a fizikai leírási formákat. Ehhez tekintsük át a rezgésekkel kapcsolatos alapvető ismereteket.

3.3.1. A rezgőmozgás, a hang keletkezése

A hang rezgés révén jön létre. A rezgő húr, gépalkatrész, hangvilla a rezgését átadja a környező levegő molekuláinak úgy, hogy a környező térben a molekulák sűrűsö-

dése és ritkulása jön létre. Ezeknek a sűrűsödéseknek és ritkulásoknak hatására a környező levegőben folytonos nyomásingadozások alakulnak ki, amelyek a levegő molekuláinak a segítségével, a molekulák egymás közötti rezgési energiájuk átadásával, hanghullámok formájában a levegőben továbbterjednek. Ahhoz, hogy a hangként megjelenő rezgés jellemzőit megérthessük, meg kell határozni a rezgés fogalmát. Azokat a fizikai folyamatokat nevezzük rezgéseknek, amelyek meghatározott időközönként újra meg újra ugyanazt az állapotot érik el, vagy ugyanazon állapoton haladnak át (Tarnóczy 1984). Ez az oszcillálás lehet periodikus vagy rendezetlen, véletlenszerű, azaz aperiodikus.

Harmonikus rezgőmozgás. A legegyszerűbb periodikus rezgés a harmonikus rezgőmozgás. Vegyünk egy rugót. Egyik végét rögzítsük, a másik végére helyezzünk egy m tömeggel bíró testet (egy elméletileg súrlódásmentes felületre), a 3.17. ábra szerinti elrendezésben. Alaphelyzetben a test nyugalomban van, ez az ábrán a b) pont, a rezgésképen pedig az eredeti nyugalmi helyzet időpillanata. Mozdítsuk ki a testet úgy, hogy széthúzzuk a rugót F erőkifejtéssel nyugalmi helyzetéből, az a) pontba, és magára hagyjuk. A test ellenkező irányba, a b) egyensúlyi helyzeten áthaladva a c) pontig kitér, majd visszafelé kezd el mozogni és kitér az a) pontig (csillapítatlan esetben). Ez a mozgás elméleti energiavesztés-mentes esetben periodikusan ismétlődik végtelen ideig, tehát egyenlő időközönként kerül a test ugyanabba az állapotba. A maximális rezgési kitérés, vagyis a rezgés amplitúdója (A) nem változik. Ez csillapítatlan rezgés. A kitérés (s) időbeli változása szinuszgörbét ad, ami a következő képlettel írható le.



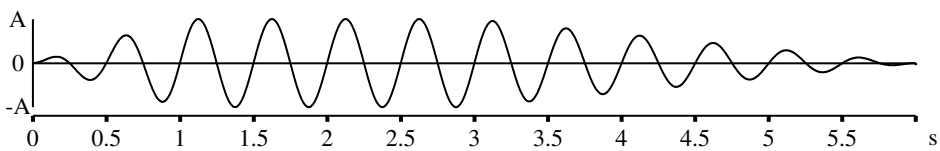
3.17. ábra. Egy csillapítatlan rezgő test mozgása és kitérési grafikonja az idő függvényében

$$s = A \sin \omega t; \quad \omega = \frac{2\pi}{T}; \tag{3.1}$$

Egy periódus ideje (T) az az idő, amíg a rezgő test először kerül újra ugyanabba az állapotba, amelyben a periódus elején volt. Az egy másodperc alatti ismétlődések, periódusok száma a frekvencia (f), mértékegysége a hertz [Hz].

$$f = \frac{1}{T}[Hz] \quad vagy \quad T = \frac{1}{f}[s] \tag{3.2}$$

Ha két teljes periódus lezajlik 1 s alatt, vagyis a periódusidő 0,5 s, akkor a rezgés frekvenciája 2 Hz, ha 20 teljes periódus zajlik le 1 s alatt, akkor a periódusidő 0,05 s, tehát a rezgés frekvenciája 20 Hz. Tehát minél nagyobb a rezgésszám másodpercenként, annál kisebb a periódusidő. A szinuszgörbével leírható harmonikus rezgések, például a hangvilla hangja, úgynevezett tisztahang érzetét keltik. A tisztahangok csak amplitúdójukban és frekvenciájukban különbözhetnek egymástól és a természetben nemigen fordulnak elő. Valójában a környezetünkben az egyensúlyi helyzetéből kimozdított és magukra hagyott rugalmas testek csökkenő amplitúdójú csillapodó rezgést végeznek. A rezgő test energiájának egy része a súrlódás révén hővé alakul át, a másik része pedig a levegő részecskéit hozza mozgásba, vagyis hangjelenség formájában kisugárzódik. A rezgés csökkenő amplitúdójú, változatlan frekvenciájú szinuszos rezgés lesz (3.18. ábra). Az egyensúlyi helyzetéből kimozdított és magukra hagyott rugalmas testek (például egy megpendített húr, egy megkoccintott üvegpohár stb.) ilyen csökkenő amplitúdójú, úgynevezett szabad rezgést végeznek, és a rezgés frekvenciája, a testre jellemzően mindig ugyanaz. Ezt nevezik a test természetes vagy sajátfrekvenciájának, amely függ a test méretétől, anyagi állandóitól. A sajátfrekvencia szoros kapcsolatban áll a rezonanciával.



3.18. ábra. A berezgés, az állandósult állapot és a lecsengés folyamata. Lineáris folymatoknál a berezgés és lecsengés idején csak az amplitúdó változik, a frekvencia nem

3.3.2. A hang terjedése a levegőben

A levegő elemi részecskéi nyugalmi állapotban állandó, rendezetlen mozgásban vannak, de úgy, hogy minden részecskének van egy átlagos „stabil” mozgási állapota, meghatározott távolsága a többi részecskétől. Tengerszinten ekkor $p_0 = 1$ atm nyomás mérhető. Ha valami a részecskéket ebből az állapotból kimozdítja, olyan erők keletkeznek, amelyek igyekeznek a részecskéket az egyensúlyi helyzetükbe visszahelyezni. Amikor egy test (a hangforrás) rezeg, a szomszédos levegő részecskéit a nyugalmi állapotból kimozdítja, és velük együtt rezeg, kimozdítva némi késéssel a távolabbi szomszédos részecskéket is azok nyugalmi helyzetéből. Vagyis a stabil állapotból való kimozdulás hatása terjed tova, a részecskék csak az egyensúlyi helyzetük körül rezegnek, átadva a zavarás hatását a szomszéd részecskéknek. A hanghullámterjedés tehát a zavar mozgásának a terjedése a hangot közvetítő közegben, például

levegőben úgy, hogy maguk a részecskék nem haladnak együtt a hullámmozgással. A hanghullámok a levegőben úgy terjednek, hogy a részecskék a hullám terjedési irányában rezegnek. Ezek az úgynevezett longitudinális hullámok. A víz felszínén terjedő hullámoknál a részecskék le-föl mozognak merőlegesen a terjedés irányára. Ezek a transzverzális hullámok. Itt is csak le-föl mozognak a vízrészecskék, és nem utaznak a hullámmal együtt. A vivőközeg, amely valamilyen mechanikai rezgés hatását közvetíti, lehet légnemű, cseppfolyós vagy szilárd.

A hang terjedési sebessége (c):

$$c = \frac{\lambda}{T} = \lambda f [m/s]. \tag{3.3}$$

A továbbiakban a levegőben terjedő hanghullámokkal foglalkozunk, amelyek terjedési sebessége c = 331,5 [m/s], 0 °C-on és 1 [atm] (100 000 [Pa]) nyomáson. A hang hullámhossza (λ) a hanghullám (T) periódusidő alatt megtett útja. A hang hullámhossza és a hang frekvenciája (f) fordított arányban állnak egymással. Egy 20 Hz-es hang hullámhossza 16,6 m, egy 20 000 Hz-es hang hullámhossza 1,66 cm a levegőben. A hanghullámok terjedésénél, mint minden hullámformánál, általában előfordulnak visszaverődések és elhajlások.

Hangnyomás és hangteljesítmény. A hang terjedésekor a részecskék sűrűsödése és ritkulása egy adott pontban $p_{hang}(t)$ nyomásingadozást eredményez. Ez a nyomásingadozás igen kicsi és a légköri (sztatikus) nyomás értékére szuperponálódik, vagyis annak hangfrekvenciás ingadoztásában nyilvánul meg. A nyomás időbeli változása tehát

$$p_{légköri} + p_{hang}(t) \tag{3.4}$$

alakban jelentkezik. Maga a $p_{hang}(t)$ függvény tartalmazhat periódusosan és statisztikusan ingadozó (nem periódusos) elemeket, de az úgynevezett alapzajtól eltekintve véges ideig tart, és rendszerint berezgési és lecsengési elemekkel is rendelkezik. Tehát matematikailag rendkívül bonyolult függvény. Ezért a megismerés formája rendszerint nem az időbeli lefolyás rögzítése, hanem valamilyen időbeli átlag, leggyakrabban a négyzetes középérték, az úgynevezett effektív érték megállapítása. Ennek értéke a *hangnyomás*:

$$p_{eff} = \sqrt{p^2(t)} = \sqrt{\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} p^2(t) dt} \quad [N/m^2]. \tag{3.5}$$

A négyzetes középérték mérése nemcsak fizikai ok, hanem egyben biológiai tapasztalat is. A fül ugyanis az úgynevezett effektív értéket érzékeli. Tárgyalásaink során *hangnyomás*on mindig a hangnyomás-ingadozás effektív értékét értjün [N/m^2]-ben,

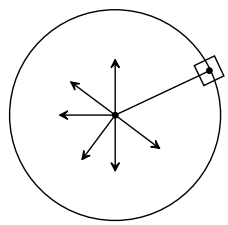
vagy [*Pa*]-ban, és (p)-vel jelöljük. Néhány nyomásérték összehasonlításként:
a sztatikus nyomás 10^5 Pa (vagy 1 atm, vagy 10^5 N/m^2)
a beszéden belüli, beszédhangok közötti átlagos nyomásingadozás 10^{-2} – 10^{-1} Pa
a beszéd dinamika tartomány (halktól a kiabálásig) 10^{-3} – 10^{-1} Pa
a hallásküszöb nyomásértéke $2 \cdot 10^{-5}$ Pa
a fájdalomküszöb nyomásértéke 20 Pa

A hangforrás elsődleges adata a *hangteljesítmény* (*P*): ami a hangforrás körül képzett gömbfelületen időegység alatt átáramlott összes energiamennyiség wattban.

$$P = \frac{p^2}{\rho c} S[W], \tag{3.6}$$

ahol *S* a felület [m^2], ρc a közegre jellemző akusztikai keménység, *c* a hangsebesség [m/s], ρ a közeg sűrűsége [kg/m^3]. A hangteljesítmény „mennyiségi", tehát összegező adat: az elemi értéknek, a felületegységre eső teljesítménynek, vagyis az intenzitásnak a sugárzó körüli teljes gömbfelületen vett integrálja.

A *hangintenzitás* (I): egységnyi felületen merőlegesen időegység alatt átáramlott hangenergia [W/m^2] (3.19. ábra).



3.19. ábra. A hang intenzitása az egységnyi felületen merőlegesen időegység alatt átáramlott energia

$$I = \frac{p^2}{\rho c} [W/m^2], \tag{3.7}$$

Szintérték – a dB fogalma. Az a legkisebb hangintenzitás-érték amelyet még épp meghallunk, vagyis az úgynevezett hallásküszöb hangintenzitás-értéke $I_0 = 10^{-12} W/m^2$, azaz 0,000 000 000 001 W/m^2. Egy nagy teljesítményű repülőgép zaja 10 m távolságban kb. $1 W/m^2$, ami az emberi fájdalomküszöbhöz közeli érték. Ez annyit jelent, hogy a hangintenzitás értéke az emberi hallható tartományban 12 nagyságrendet fog át, vagyis a repülőgép zajának a hangintenzitása az éppen meghallható hang intenzitásának 10^{12}-szerese. Olyan esetekben, amikor a kezelt mennyiségek mértéke több nagyságrendet átfog, célszerű szintértékként logaritmikus viszonyszámot használni. A Ilyen viszonyszám a decibel (dB), ami az adott, teljesítmény jellegű mennyiségek arányának 10-es alapú logaritmusa, 10-zel szorozva:

$$X_{dB} = 10\lg\frac{X}{X_0} = 10\lg X - 10\lg X_0. \tag{3.8}$$

Az akusztikában és a vele kapcsolódó tudományágakban, mint a fonetika, pszicholingvisztika, digitális beszédfeldolgozás stb. a hangintenzitás és a hangnyomás kezelésére dB-szintértéket használnak, és a viszonyítási alap a hallásküszöbintenzitás, illetve hangnyomás értéke.

A hangintenzitás szintértéke:

$$L_{dB} = 10\lg\frac{I}{I_0} = 10\lg I - 10\lg I_0, \tag{3.9}$$

ahol a viszonyítási alap a hallásküszöb intenzitásértéke, vagyis $I_0 = 10^{-12} W/m^2$.

A hangintenzitás szintje tehát a hallásküszöbnél:

$$L_{dB} = 10\lg 1 = 0\,dB, \tag{3.10}$$

a hangos beszédnél:

$$L_{dB} = 10\lg\frac{10^{-6}}{10^{-12}} = 10\lg 10^6 = 60\,dB, \tag{3.11}$$

egy nagy teljesítményű repülőgép zaja esetén

$$L_{dB} = 10\lg\frac{10^{0}}{10^{-12}} = 10\lg 10^{12} = 120\,dB. \tag{3.12}$$

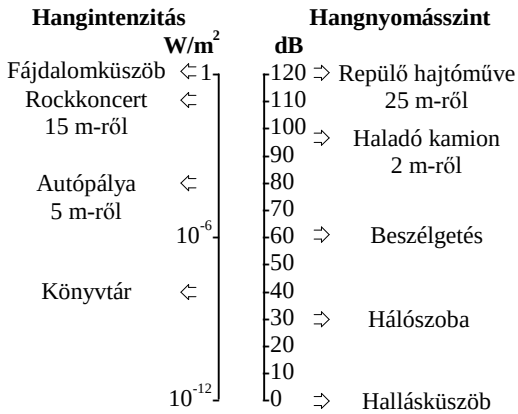
A hangintenzitás a hangnyomás négyzetével arányos. Szintben kifejezve:

$$L_{dB} = 10\lg\frac{I}{I_0} = 10\lg\frac{p^2}{p_0^2} = 20\lg\frac{p}{p_0}. \tag{3.13}$$

A hangnyomás szintértéke:

$$L_{dB} = 20\lg\frac{p}{p_0} = 20\lg p - 20\lg p_0, \tag{3.14}$$

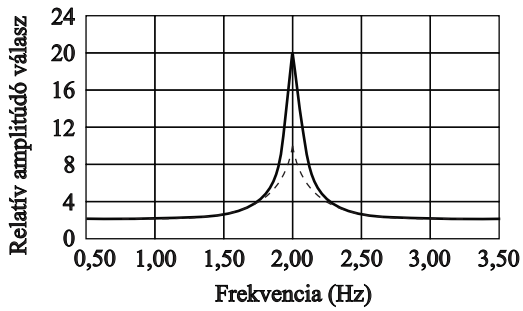
ahol a viszonyítási alap a hallásküszöb hangnyomásértéke, vagyis $p_0 = 2\cdot 10^{-5} Pa$. Ezt nevezik *akusztikai decibel*nek. Ha tehát hangnyomásarányokkal számolunk (például beszédhangok amplitúdó-időfüggvényeinek összehasonlításakor) dB-ben, akkor a hangnyomásarányok logaritmusának húszszorosát kell vennünk. Így akár hangnyomás-, akár intenzitásszint-értékekkel számolhatunk, a szintértékek nagysága egyenlő. A teljes hallástartomány dinamikai skálája a 3.20. ábrán látható szemléletes példákkal.



3.20. ábra. A teljes hallástartomány dinamikában

3.3.3. Kényszerrezgés, rezonancia

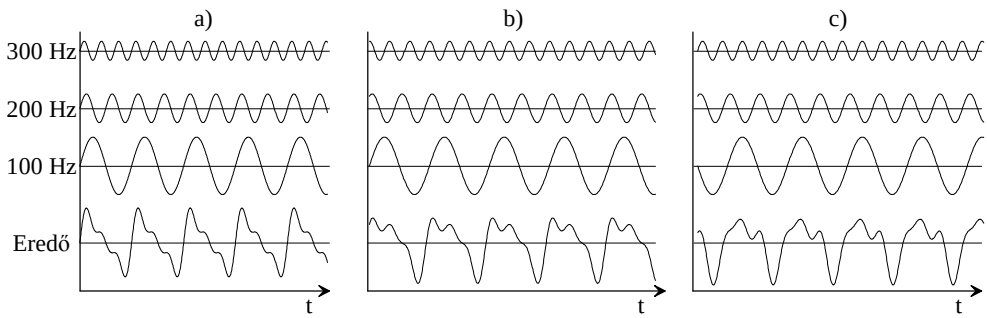
A rezgés folyamatossá tétele csak külső energia bevitelével lehetséges, az energiaveszteséget pótolni kell. Ha az egyszerű rugóra rögzített testet külső erővel előrehátra mozgatjuk, akkor a testre kényszerítő erőt fejtünk ki, és a test kényszerrezgést fog végezni. A kényszerrezgés frekvenciáját a kényszerítő erő (gerjesztő erő) frekvenciája szabja meg. A kényszerítő erő hatására létrejövő mindenkori kitérési amplitúdó több tényezőtől függ. Elsősorban a kényszerítő erő amplitúdójától, utána a test sajátfrekvenciájától és végül a kényszerítő erő frekvenciájától. Az ily módon kényszerrezgést végző test kitérési amplitúdója akkor lesz a legnagyobb, ha a kényszerítő (gerjesztő) frekvencia megegyezik a kényszerített (gerjesztett) rendszer sajátfrekvenciájával. Ezt nevezik rezonanciajelenségnek, a frekvenciát pedig rezonanciafrekvenciának. Ilyenkor együtt rezeg a kényszerítő rendszer a kényszerítettel. A frekvencia függvényében felvett rezgésamplitúdó-görbét rezonanciagörbének nevezzük. A rezonanciafrekvencián a kényszerrezgés amplitúdója a kényszerítő rezgés amplitúdójának sokszorosa lehet. A rezonanciagörbe alakja függ a csillapítástól. Például, ha nagy a súrlódás, akkor az amplitúdó növekedése kisebb lesz, és fordítva. A 3.17. ábrán bemutatott m tömegű test kényszerrezgésekor, a kényszerítő erő frekvenciájának függvényében kialakuló rezonanciagörbét látjuk a 3.21. ábrán két különböző csillapítás esetében. A függőleges tengelyen a konstans amplitúdójú gerjesztéshez viszonyított rezgésamplitúdó látható. A szaggatott vonal mutatja a test sajátfrekvenciáján mérhető amplitódót, nagyobb csillapítás esetén.



3.21. ábra. A 2 Hz sajátfrekvenciájú test kényszerrezgésekor kialakuló két rezonanciagörbe a kényszerítő frekvencia függvényében. A szaggatott görbe nagyobb csillapítás esetében alakul ki

3.3.4. Összetett rezgések

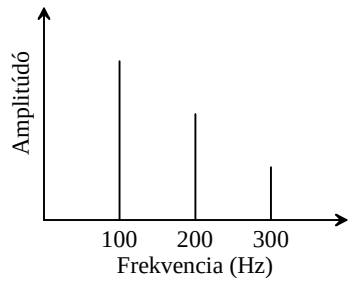
A szinuszos formájú harmonikus rezgés által keltett tisztahang ritka jelenség a hangforrások világában. A hangvilla rezgése szolgáltat tisztahangot, vagy igen ügyes fütyüléssel egyesek képesek közel tisztahangot létrehozni. A természetben előforduló rezgések azonban összetett rezgések. Több, egymástól különböző rezgőmozgást egy anyagi részecske nem végezhet egy időben. Érvényesül a lineáris szuperpozíció elve, vagyis egyazon pontra ható rezgések egyszerűen összeadódnak. Például zárt térben egy hangforrás keltette hanghullám rezgéséhez a falról visszaverődő hanghullámok rezgései hozzáadódnak. A tiszta szinuszos rezgések párhuzamos összetételéből egyszerű és összetett, azaz nem tiszta szinuszos periódusos rezgések vagy ezeknek különleges esetei származhatnak. Az egyszerűség kedvéért a következő példában egy három frekvenciaösszetevőből álló összetett rezgést vizsgálunk meg. Az összetevő rezgések amplitúdója nem egyenlő, frekvenciájuk aránya egész szám a legalacsonyabb frekvenciájú rezgéshez viszonyítva. A 100 Hz, 200 Hz és 300 Hz frekvenciájú rezgések esetén a rezgések frekvenciájának hányadosa egész szám, 200/100 = 2, 300/100 = 3. Az ilyen elemek össztétele ismét periodikus rezgést eredményez, melynek frekvenciája megegyezik az összetételben szereplő legkisebb frekvenciával, alakja azonban nem szinuszos, hanem erősen függ az összetevő rezgések frekvenciájától és kezdőfázisától. A 3.22. ábrán háromféle fázisbeállításban (azaz egymáshoz képest három különböző időeltolásban) mutatjuk be ugyanazokat a részrezgéseket és a keletkezett összetett rezgést (Tarnóczy 1982). Látható, hogy az egyes részhangok fázisbeállításától mennyire függ a végeredményként keletkező rezgés alakja. Az eredő rezgésalakok helyességét ellenőrizhetjük, ha adott időpillanatokban az összetevő rezgések amplitúdóit egyszerűen, grafikusan összeadjuk. A fenti példából két dolog következik. Az egyik, hogy a fáziseltolás nem befolyásolja az összetett hang érzeti hangzását (annak ellenére, hogy más-más a hullámforma alakja), hiszen a frekvenciakomponensek nem változnak.



3.22. ábra. Részrezgések összegzése. Ugyanazon komponensekből, de más fázisértékekkel összerakott jel végső rezgésformája más. A hallási rendszerünk azonban ezt ugyanolyan színezetű hangnak érzékeli

Összetett rezgések frekvenciaelemzése. Ahogy már említettük, a természetben előforduló rezgések összetett rezgések, ezeket érzékeljük a fülünkkel, és ilyen jeleket veszünk fel mikrofonokkal hangfelvételkor. A szuperpozíció elvének tárgyalásakor láttuk, hogy egymással egész számú viszonyban álló frekvenciájú szinuszos rezgések párhuzamos összetétele periodikus rezgést eredményez. Ez a tétel megfordítható, tehát a periodikus rezgések részelemei szinuszos rezgések. A rezgések vizsgálatakor az egyik célkitűzés, hogy megállapítsuk az összetett rezgések frekvenciakomponenseit. Azt a folyamatot, amikor egy összetett rezgést (akár periodikus, akár nem) frekvenciakomponensekre bontunk, frekvenciaelemzésnek nevezzük. Fourier, francia matematikus a 19. század elején kimutatta, hogy lineáris rendszerekben bármely összetett rezgés időfüggvénye felbontható különböző frekvenciájú, amplitúdójú és fázisú harmonikus komponenseire (szinuszhangok sokaságára).

Amikor frekvenciakomponensekre bontjuk az adott összetett rezgés hangnyomásidőfüggvényét (analóg jel az időtartományban), akkor azt frekvenciatartományban értelmezhető függvénnyé alakítjuk át. A gyakorlatban a hang spektruma lehet nyomásamplitúdó-, teljesítmény- vagy energiaspektrum, attól függően, hogy az adott időpontban a frekvencia-összetevők a nyomásamplitúdó, a teljesítmény vagy az energia eloszlását adják meg. A periodikus összetett rezgésekre jellemző egy alap ismétlődési periódus (alapfrekvencia, alaphang), amely az összetett hangot felépítő összes frekvencia-összetevő közül a legalacsonyabb frekvenciájú, általános jele f_0 (a beszédben F_0 jelet használnak). Férfiak beszédében az alaphang az F_0 jellemzően 100 Hz körüli, nőknél ennek közel kétszerese. A beszédben az alaphang adja meg a beszélő személy úgynevezett hangfekvését (mély hangú, magas hangú beszélő). A hangfekvés jellemző a beszélő személyre. A összetett hang többi összetevőjét felhangoknak ($\mathrm{f}_1\ \mathrm{f}_2\ \ldots \mathrm{f}_n$) nevezzük. A beszédben a felhangok a legalacsonyabb frekvenciájú alaphang (f_0) egész számú többszörösei (például a magánhangzókban). A periodikus rezgések tehát úgynevezett *vonalas sepktrum*mal rendelkeznek (csak a felhangok frekvenciáin vannak spektrumösszetevők). Fontos megjegyezni, hogy a



3.23. ábra. A 3.22. ábra legalsó sora szerinti periodikus összetett rezgések elméleti amplitúdóspektruma. A kép ugyanaz mindhárom rezgésformára, mivel azok csak fázisukban térnek el

felharmonikusok egymás közötti távolsága lineáris a frekvenciatengelyen, és hogy két felharmonikus között a távolság megegyezik az alapfrekvencia értékével (lásd a 3.23. ábrát). A hallórendszerünk képes arra, hogy egy összetett periodikus hangból a különbségi hangokat kihallja. Ezt a hallási tulajdonságot használják ki az orgonatervezők, amikor nem készítik el a mély alaphangnak megfelelő nagy méretű sípokat. Ebből a hallási tulajdonságból adódik az is, hogy egy beszélő személy alaphangját akkor is meg tudjuk ítélni, ha a tényleges F_0 komponens nincsen benne a beszédjelben (például a telefonon hallott hangban a férfi alaphang nincs benne az átviteli rendszer szűrése miatt).

A vonalas spektrum elméleti képe a gyakorlatban azonban kissé más. A vonalak környezetében is vannak frekvenciakomponensek, mert a vizsgált jel általában nem stacionárius, ezért a jelet csak korlátozott időtartományon belül (ablakolással) tudjuk vizsgálni, és az ablakolás torzítja jelet (lásd a magánhangzók spektrumképeit az 5.9 ábrán). Továbbá a jel még az ablakon belül is változhat, ezzel további összetevőket hoz be a szinképbe. Például a beszéd esetében a hangszalagok nyitódási-záródási bizonytalanságából adódó alapfrekvencia-ingadozás (kváziperiodikus jelleg).

A nem periodikus rezgés esetén, mint például olyan zörej, amelynek frekvenciaösszetevői minden frekvencián egyenlő intenzitással megtalálhatók (fehérzaj) vagy az impulzus jellegű, gyors lefolyású hangok (zárfelpattanás), az összetevő frekvenciakomponensek között nincs olyan szabályosság, mint ami a periodikus hangoknál volt. A nem periodikus rezgések végtelen sok frekvenciájú szinuszos összetevőből állnak, és ezek a frekvenciatartomány bármely pontján lehetnek (vagyis az összetevők nem meghatározott frekvenciáknál koncentrálódnak). A nem periodikus jelek ezért *folytonos spektrum*mal rendelkeznek.

A frekvenciatartományban értelmezett függvényeknek a gyakorlatban főként két ábrázolási módja van. Az egyik a spektrum típusú ábrázolás, ahol egyetlen időablakhoz rendelt spektrális komponenseket két dimenzióban ábrázolják, a vízszintes tengelyen a frekvenciát, a függőleges tengelyen az intenzitást tüntetik fel. A másik ábrázolási mód a spektrogram típusú ábrázolás, amikor az időablakot folyamatosan

csúsztatják az időtengelyen mutatva az ablakban mért spektrum időbeli változását. Ez 3-dimenziós ábrázolási mód: a vízszintes tengelyen az idő, a függőlegesen a frekvencia található, a harmadik dimenzió pedig a spektrális összetevők amplitúdójának értékét mutatja, amelyet általában szürke árnyalatos skálával (3.24. ábra) vagy színkódokkal érzékeltetnek.

3.3.5. A beszédjel elemzése

A beszédjel összetett rezgés, amely időben folyamatosan változó, különböző rezgésmódok kombinációja. A beszédjel elemzése bonyolult feladat, különösen két szempontból.

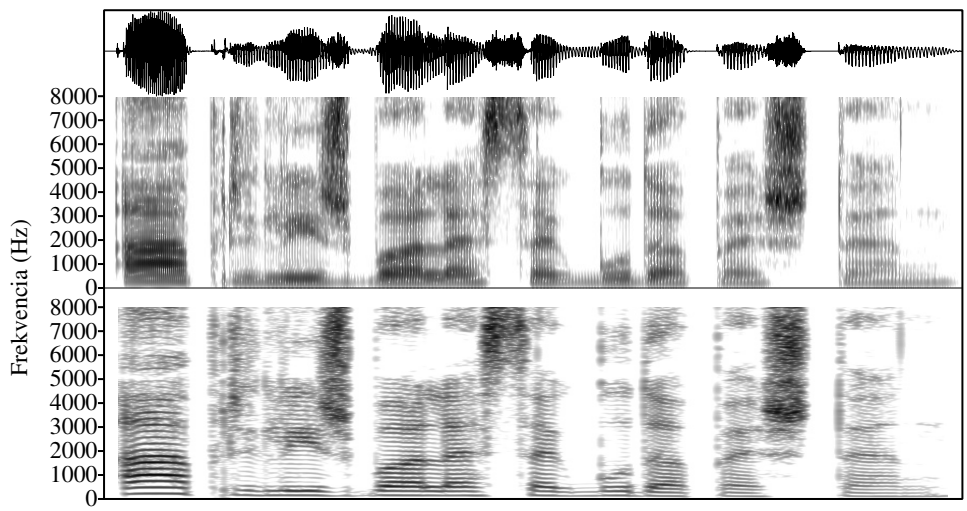
1. Egyrészről a szabályosság nem teljesül, hiszen a beszéd biológiai produktum, ahol a beszédjel időfüggvényének egyes megvalósulásai a biológiai rendszer pillanatnyi állapotától függnek. Például a hangszalagok nyitódásának, záródásának rendszeres ismétlődése sem tekinthető szabályosnak, az ismétlődések kicsit eltérnek egymástól. Ezért a beszédben a zöngejel alapfrekvenciáját kváziperiodikusnak tekintik, és ez a kváziperiodicitás fontos eleme az emberi hang jellemző hangzásának. Például még ugyanazon személy kitartott magánhangzójának az időfüggvénye is más és más periódusokat tartalmaz (nem determinisztikus). Az ilyen típusú, de időben állandó (stacionárius) jeleknél a hosszabb időre vett átlaguk hasonló, így egyetlen realizáció időátlagából vonunk le következtetéseket. Ez a következtetés azután más realizációk időbeli átlagára is jó közelítéssel érvényes lesz. Így, leggyakrabban a teljesítményszint vagy intenzitásszintsűrűség-spektrumot (a Fourier-transzformált négyzete) szokás kiszámítani, vagyis egy meghatározott sávszélességre eső teljesítmény- vagy intenzitásszintet dB/Hz-ben (O'Shaughnessy 1987). Valójában meghatározott sávszélességben szűrjük a jelet, és a meghatározott sávszélességbe eső teljesítményt vagy intenzitást számoljuk. Gyakran a jellemzőt sok mérés utáni átlagszámításból adjuk meg. A teljesítményszint-spektrum, illetve intenzitásszint-spektrum a jel meghatározott időintervallumában a frekvencia-összetevők teljesítményszint-, illetve intenzitásszint-eloszlását adja meg.

2. A beszéd-előállítás időben változó folyamat. A beszédképzés során az ember folytonos és időben változó jelet állít elő, amelyben időben változó (tranziens), időben állandósult (stacioner) és időben gyorsan változó, impulzusszerű jelek váltakozva követik egymást. Az ilyen összetett jelek matematikai kezelése bonyolult. A beszéd egyes szakaszai azonban, korlátozott időtartományban közel állandónak vehetők, és így a beszédrészlethez tartozó időablakban az elemzés elvégezhető (7.1.2. fejezet). Az ilyen közel állandó beszédszakaszokra a kvázistacionárius jelzőt alkalmazzák. A beszéd ilyen kvázistacionárius részek sorozatának tekinthető. A spektrumelemzés a kvázistacionárius részeken úgy történik, hogy az egymás

után kijelölt mérési pontokban (például 10–20 ms-os időosztás szerint) mindig egy meghatározott időablakban (szélessége például 25 ms) mérjük a teljesítményspektrumot. Az időablakot végiggörgetve a vizsgált beszédszakaszon megkapjuk a gördülő teljesítményspektrumot, vagyis az idő függvényében változó dinamikus teljesítményspektrogramot (ezt a fonetikában szonogramnak, az általános beszédkutatásban dinamikus hangspektrogramnak nevezik). Ez megmutatja a frekvencia-összetevők teljesítményszint-eloszlásának időbeli változását a mért intervallumban (például a mondatban). Az ilyen regisztrátumokon a beszéd frekvencia-összetevői láthatóvá válnak. Az ilyen képeket nevezték már az 1940-es években látható beszédlenyomatnak.

Frekvenciaelemzésnél a meghatározott időintervallum, vagyis az elemzési ablak szélessége (ΔT) meghatározza az elemzés frekvenciafelbontását (Δf), tehát azt, hogy milyen részletességgel kapjuk meg a spektrumösszetevőket. A kettő szorzata ($\Delta T \Delta f$) állandó érték. Ahhoz, hogy időben gyors változásokat le lehessen olvasni a spektrum képéről, az ablakszélességet rövidre kell választanunk (ilyenkor válnak láthatóvá a zárfelpattanások rövid időtartamú zörejeinek frekvenciakomponensei). Ennél a formánál láthatók jól a nagy energiájú felhangcsoportok, amelyeket formánsoknak neveztek el. Ezek szélesebb (300 Hz-nyi) frekvenciasávban terülnek el, több felharmonikust is magukban foglalnak. A finom frekvenciafelbontáshoz hosszú időablakra (például 30 ms) van szükség. Ilyenkor láthatóvá válnak a zöngés hangok felhangjai, viszont az időben gyorsan lejátszódó események elkenődnek az időtengelyen. A beszédelemzési technikában tehát az elemzés célja határozza meg, hogy milyen időablakkal végezzük az elemzést (a szokásos időablakok 5 ms és 50 ms közötti értékűek). A 3.24. ábrán egy mondat időfüggvényét, valamint a kétféle felbontással készült spektrogramot mutatjuk be (fenn rövid, és lenn hosszú időablakkal végezve az elemzést). Az ábra vízszintes tengelyén az időt mutatjuk másodpercben, a függőleges tengelyen a frekvenciát tüntettük fel 8 kHz-ig. Az adott időponthoz tartozó intenzitásszint nagysága arányos a feketedés mértékével. Minél feketébb a kép, annál nagyobb a hangrész frekvenciaösszetevőjének az intenzitása. A rövid időablakú elemzéssel kapott felső spektrogramon az artikuláció folyamán bekövetkező változásokat, zárfelpattanásokat hűen tudjuk követni, de a frekvenciafelbontás elég rossz. A felhangtartalom összemosódik, ennek következtében az energiakoncentrációk jobban leolvashatók. A hosszabb időablakú elemzésnél a gyors változások nem követhetők jól, de a frekvenciafelbontás sokkal jobb, mint a felső spektrogramnál. Itt a vízszintes csíkok feketedési szintjei az egyes felhangok erősségének időbeli változását szeparáltan mutatják. A hangspektrogram jól használható vizuális elemzésre is, hiszen a beszéd komponensei láthatók rajta. A spektrogram úgynevezett olvasásakor a rajta látható információhalmazból a beszédre vonatkozó számos információ kiolvasható. Ha a feketedések váltakozását az időtengely mentén vizsgáljuk, akkor látható, hogy kisebb és nagyobb energiájú hangrészek váltják egymást. A zöngés-zöngétlen szakaszok is elválaszthatók egymástól. A rövid időablakkal végzett elemzéseknél a

hangspektrogram időtengelyén láthatók a beszélő zöngés hangperiódusai (függőleges bordázat), ebből kiszámítható a pillanatnyi alaphang magassága is. Ha ismerjük a hangidőtartamokra jellemző értékekeket (lásd a későbbi fejezetekben), akkor megállapíthatjuk, hogy hány hangot mondott a beszélő másodpercenként. A hangok típusaira is következtethetünk az energiaeloszlásokból. Ha a frekvenciatengely szerint vizsgáljuk a komponenseket, akkor azt láthatjuk, hogy szintén kisebb-nagyobb energiájú részek követik egymást adott időponthoz kötve a frekvenciatengelyen. Az energiakoncentrációk frekvenciahelyeiből következtethetünk a hangok típusára. A réssel képzett zöngétlen hangok például csak 1500 Hz fölötti frekvenciákon tartalmaznak nagy energiájú frekvencia-összetevőket.

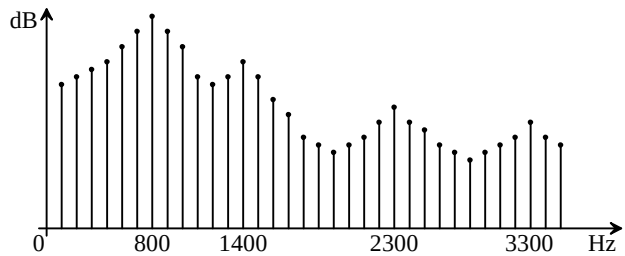


3.24. ábra. Az *Áprilisban leszek Budapesten.* (1,7 s) mondat amplitúdó-időfüggvénye (fenn), spektrogramja rövid, 5 ms-os elemzési ablakkal (középen) és spektrogramja hosszú, 30 ms-os elemzési időablakkal (lenn) női ejtésben

3.3.5.1. Formáns, zörejgóc

Hogyan alakul ki a beszéd jellegzetes, változatos spektrális tartalma? A zöngés elemek frekvenciaszerkezete a gégeszintű hangforrás és az artikulációs csatorna (mint rezonáló üregrendszer) együttes tulajdonságaiból alakul ki, és a pillanatnyi (minden zöngeperiódusnak megfeleltetett) spektrummal jellemezhető. Ezt nevezzük az artikuláció akusztikus vetületének (lásd később). Ha van fonáció, vagyis hangszalagrezgés, akkor a hangszalagrezgéssel előállított zönge rezgésformája gerjeszti az artikulációs csatornát. A zönge tartalmazza az alaphangot és annak felharmonikusait, egészséges esetben mintegy 5000 Hz-ig. A vonalas zönge spektrumában a felhangok

amplitúdói átlagosan 12 dB/oktáv csökkenéssel vannak jelen. A legnagyobb amplitúdóval az alaphang rendelkezik. A felharmonikusok frekvenciái az alaphang frekvenciájának egész számú többszörösei. Ez a zöngehang kerül az artikulációs csatornába, ahol a felharmonikusok bizonyos csoportjai a pillanatnyi rezonanciafrekvenciákon és környékükön felerősödnek. A spektrumban mérhető ilyen felerősödött felhangcsoportokat a fonetikai szakirodalomban formánsnak nevezik. Minél mélyebb a beszédhang, annál több felharmonikus vesz részt a formánsok kialakításában. Minden zöngés hangnak megvannak a jellemző formánsai, általában 4–5 formáns mérhető. Jelük: F1, F2, F3, F4, F5, a frekvencia növekedésének függvényében. A formánsokhoz definiálnak formánssávszélességet is (a rezonenciafrekvencia-görbe maximumától visszaszámolt –3 dB-es pontok közötti frekvenciatávolság). Ezeket B1, B2, B3, B4, B5-tel jelölik (3.26. ábra). Minél magasabb a formáns sorszáma, annál távolabb



3.25. ábra. Férfi ejtésű magánhangzó elméleti vonalas spektruma. A formánsok a vonalakra helyezett burkológörbe csúcsainál vannak

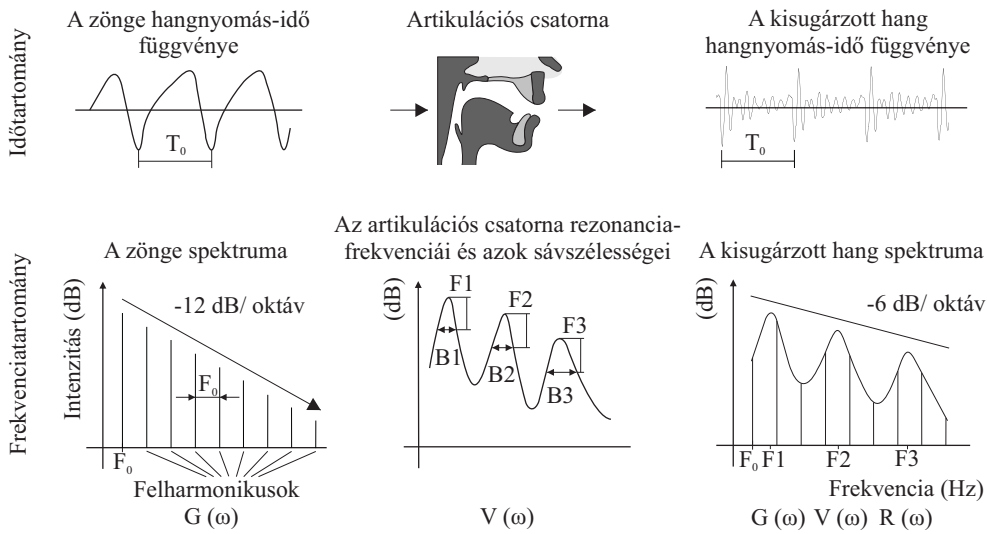
van az alaphangtól, és annál szélesebb a sávszélessége. A formánsok és azok sávszélességei határozzák meg a szinképben a maximum és minimum helyeket, amelyek az egyes hangokat jellemzik. A magánhangzókat az első két formánsuk és azok sávszélessége már jellemzi. A zöngés hangokban a formánsok mozognak az artikuláció függvényében. Formánsok csak akkor mérhetők, ha van a száj- vagy orrnyíláson keresztül hangkisugárzás. Zárhangok zárszakaszában (lásd később) tehát nem mérhető formáns, még akkor sem, ha zöngeképzés van. Abban az esetben, amikor nincs fonáció (zöngétlen hangot ejtünk), az artikulációs csatornában kialakított akadályoknak megfelelően turbulens áramlások, zárfelpattanási zörejek (lökéshullámok) gerjesztik a hangképző csatornát, a képzés helyének megfelelő zörejgócokat kialakítva a hang színképében. Esetenként hangmentes szakaszok (néma fázis), alakulnak ki a nyelv által meghatározott artikulációs mozgások és hangadások szerint. A beszédben a zöngés és zöngétlen hangok esetleges váltakozása adja meg azt a nagyfokú variálhatóságát, ami biztosítja, hogy bármilyen tartalmú üzenetet ki tudunk fejezni. Fontos tudni, hogy a beszédhangok nem egymástól elkülönült elemei a beszédnek, hanem hangátmenetek kötik össze őket (lásd a 3.1. fejezetet), a koartikuláció következtében.

Annak ellenére, hogy a hangképző szervek komplex akusztikai rendszert alkotnak, egyszerűsített modellek segítenek bennünket abban, hogy megértsük a különböző beszédhangok előállításának módozatait. Az artikulációs csatorna működésének modellezésére számos elmélet látott napvilágot. Két ilyen modellt ismertetünk.

A beszédképzés gerjesztett szűrő modellje. A létrehozott beszéd akusztikai tulajdonságait lényegében három tényező határozza meg: a hang előállításának a módja vagyis a hangforrás, a hozzá kapcsolodó hangképző üregrendszer rezonanciatulajdonságai és a hangtérbe való sugárzás milyensége (3.26. ábra). Így a beszéd alapvetően e három komponens által képzett függvény szorzataként áll össze a 3.15. egyenlet szerint (Fant 1960, Kent–Read 1992), ahol $S(\omega)$ jelenti a beszéd eredő színképét (Speech), $G(\omega)$ a hangszalag rezgésével létrehozott hangforrás színképét (Glottal), $V(\omega)$ a hangképző üregek átviteli függvényét (Vocal tract), $R(\omega)$ pedig az ajak kisugárzásánál a sugárzási ellenállás átviteli függvényét (lip Radiation). A 3.26. ábrán látható a három összetevő hatása a beszédjelre.

$$S(\omega) = G(\omega)V(\omega)R(\omega). \tag{3.15}$$

Magánhangzók és zöngés mássalhangzók képzésénél, a hangszalagműködésből keletkező zönge hangnyomás-időfüggvénye közel fűrészfog jellegű, T_0 alapperiódussal. Ennek szinképi összetevőit képviseli $G(\omega)$ függvény, amely leírja az alaphang (F_0) és a felhangok együttesét. A felhangok intenzitása az ilyen fűrészfog típusú időfüggvények esetén átlagosan 12 dB/oktáv meredekséggel csökken a frekvencia növekedésével. Ezt a színképet befolyásolja a változó méretű artikulációs csatorna, amely egy üregrendszer több rezonanciafrekvenciával (Kent–Read 1992), ez a $V(\omega)$ függvény. A rezonanciafrekvenciákon és azok környezetében a felhangok intenzitása megnő, más helyeken elnyomódik. Így alakulnak ki a zöngés hangok formánsstruktúrái. A végső, a hangtérbe jutó hang színképét még az ajak kisugárzásánál a sugárzási ellenállás $R(\omega)$ átviteli függvénye befolyásolja. Ez azt jelenti, hogy az akusztikus energia kicsatolása frekvenciafüggő. Magasabb frekvenciákon a sugárzási ellenállás kisebb, mint az alacsony frekvenciákon. Ez a színkép felső frekvenciatartományának a megemelését eredményezi átlagosan 6 dB/oktávval. A zöngétlen hangok esetében a $G(\omega)$ gerjesztő függvény lehet súrlódási zörej, mint például a réshangok képzésénél, ahol a súrlódási zörejt a résen kiáramló levegő örvényleszakadása okozza, vagy lehet a hangképző csatornában keletkezett zár felpattanásakor keletkező zárfelpattanási zörej. Mindkét esetben a zörej jellegű $G(\omega)$ színképet a gerjesztett üreg $V(\omega)$ rezonanciája befolyásolja, színképi súlypontokat alakítva ki az eredő színképben. Ezeknél a hangoknál az üregeknek nincs olyan éles rezonanciája, mint a magánhangzók esetében, de a színképi súlypontok a zár vagy a rés képzési helyének függvényében változnak.



3.26. ábra. A beszéd létrehozása fiziológiai és fizikai szinten a zöngés beszédhangokra a gerjesztett szűrő modellje szerint. A beszédjel alapvetően három komponens által képzett átviteli függvény szorzataként áll össze. A formánsok a gerjesztett szűrő rezonanciafrekvenciáinál alakulnak ki. A formáns sávszélességek a frekvencia növekedésével nagyobbodnak. Az ábra a beszédképzés egy kiragadott pillanatát szemlélteti

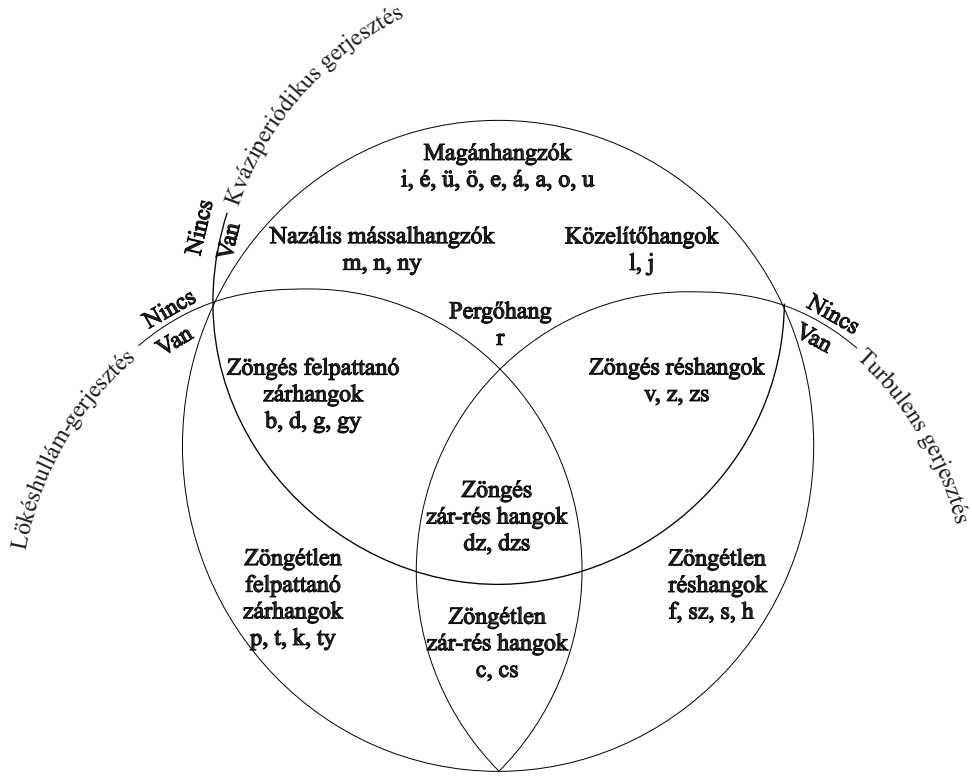
A beszédhangok képzésekor beszélhetünk tiszta (egyfajta gerjesztésű) hangokról, ahol a $G(\omega)$ gerjes tisztán vagy zönge (például magánhangzók), vagy súrlódási zörej (réshangok), vagy pedig lökéshullámszerű zárfelpattanási zörej (zárhangok). A 3.27. ábrán ennek szemléletes rendszerezése látható Gordos–Takács (1983) alapján a magyar beszédhangokra (a részletes hangleírásokra a további vonatkozó fejezetekben kerül sor).

Gyakoriak a vegyes gerjesztésű mássalhangzók is, ahol több egyidejű gerjesztés fordul elő egyetlen hang képzésén belül. A zöngés zár-, illetve réshangok képzésekor a zárfelpattanási-, illetve súrlódási zörej mellett zöngés gerjesztési hang is része lehet a hangképzésnek. A gerjesztés típusait tekintve, a legösszetettebb hangok a zöngés zár-rés hangok, amelyekben mindhárom képzési forma szerepet játszik (zönge, súrlódási zörej és lökéshullámszerű zárfelpattanási zörej).

A gyakorlatban a beszédhez mindig hozzáadódik az akusztikai környezet hatása $N(\omega)$ is, ami időben változóan befolyásolja a beszédjel végleges formáját. A természetes környezetben tehát a beszéddel párhuzamosan minden esetben jelen van valamilyen zaj. A hangstúdióban ez elhanyagolható, viszont egy forgalmas utcán talán a beszédet is elnyomja. A környezeti zajok sokfélék lehetnek: utcazaj, ajtócsapkodás, széknyikorgás, háttérzene, más beszélő hangja stb. Így a hangtérben terjedő beszéd színképéhez a hangtér is mindig hozzájárul. Eszerint a hétköznapi, valós helyzetekben előforduló beszéd eredő színképe minden esetben 4 komponensből tevődik össze, az alábbiak szerint:

$$S(\omega) = G(\omega)V(\omega)R(\omega)+N(\omega). \tag{3.16}$$

A gépi beszédfelismerők modellezésénél ez nagy gondot okoz, hiszen nem tiszta



3.27. ábra. A magyar beszédhangok rendszerezése a gerjesztésük szerint (Gordos–Takács 1983 27. o.)

beszédet kell analizálni, hanem zajjal szennyezett hullámformát. A megoldási kísérletekre a további fejezetekben térünk ki.

A beszédképzés gerjesztett cső modellje. Az egyik legegyszerűbb modell azon a közelítő feltevésen alapszik, hogy a hangképző csatornában – mint egy keskeny csőben – csak tengelyirányban terjednek a hullámok, mivel a keresztmetszeti méretek a hullámhosszhoz képest kicsik. Ilyenkor a cső alakja egy egydimenziós keresztmetszeti függvénnyel ($A(x,t)$) leírható, ahol az x keresztmetszet a t időben folyamatosan változik, és feltételezik, hogy a hanghullám visszaverődése a csőfalról veszteségmentes, veszteség csak a száj- és orrnyíláson keresztül a térbe sugárzásból ered, a toldalékcső csatolásmentes, és a csőfalak merevek. A csőben terjedő hanghullámot a fenti egyszerűsítések mellett az alábbi egyenletek írják le:

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial \frac{u}{A}}{\partial t}, \tag{3.17}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t}. \tag{3.18}$$

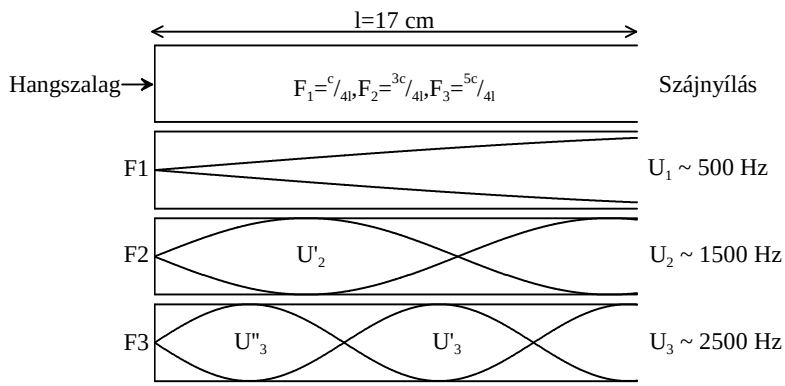
A hangtér e differenciálegyenlet-rendszerének megoldásakor egy állandó $A(x,t) = A_0$ keresztmetszetű cső esetén, szinuszos gerjesztés mellett, az egyik végén zárt (zárt a hangszalagoknál), a másik végén nyitott (nyitott a szájüregnél, orrüregnél) cső kimeneti térfogatsebessége meghatározott, az alábbi átviteli függvény szerint alakul.

$$u(l,t) = \frac{1}{\cos \frac{\omega l}{c}} U_g e^{j\omega t}, \tag{3.19}$$

ahol U_g a gerjesztés komplex amplitúdója. Az ilyen átvitelnek végtelen sok pólusa van, a pólusok frekvenciája csak a cső hosszától függ, és az alábbi képlet szerinti helyeken mutat rezonanciafrekvenciákat.

$$f_n = \frac{c}{4l}(2n-1); n = 1,2,\ldots \tag{3.20}$$

A hang terjedési sebessége c = 340 m/s, 1 atmoszféra nyomáson és 20 C° hőmérsékleten.

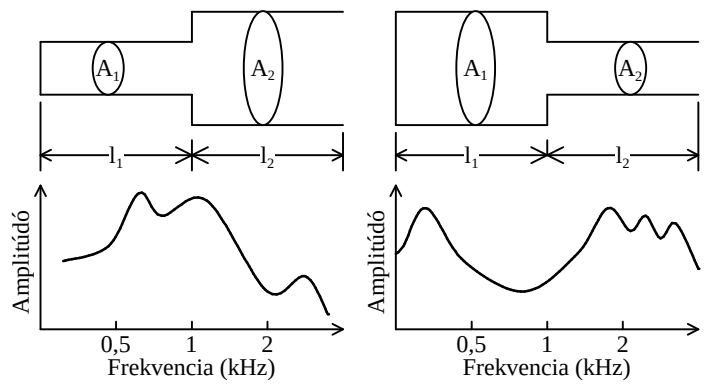


3.28. ábra. Az egyik végén zárt (hangszalagoknál), a másik végén nyitott (szájüreg) cső állóhullámú rezonanciahelyei. U_1, U_2, U_3 jelzi a maximális térfogatsebességeket (Kent–Read 1992)

Ha a cső hossza l = 0,17 m (átlagos férfi artikulációs csatornája), akkor az első, második és harmadik állóhullámú rezonancia F1 = 500 Hz, F2 = 1500 Hz, F3 = 2500 Hz (3.28. ábra). Az egyenletes keresztmetszetű csőhöz hasonlítható a beszédképzésben a semleges magánhangzóhoz (lásd később) tartozó artikulációs csatorna alakja. Ennek a beszédhangnak az első három formánsa rendre az 500, 1500, 2500 Hz körüli értékeknél látható a hangspektrogramokon. Mint ahogy már említettük, a valóság-

ban a hangképző csatorna keresztmetszete folytonosan változik, ami az állóhullámú rezonanciát befolyásolja. Azonban a csatorna felosztható közel állandó keresztmetszetű szakaszokra, melyekre a hullámegyenletek pontosan leírhatók és a rezonanciaértékek kiszámolhatók (Gordos–Takács 1983). A pontos leírást bonyolítja, hogy a keresztmetszet-változásoknál a hullámimpedancia megváltozik, ezért ezeknél a helyeknél visszaverődések lépnek fel.

Magánhangzók esetében már két különböző A1, A2 keresztmetszetű cső együtteséből álló modell átviteli függvénye is egész jó közelítést ad. Az [aː] és [i] hangokra jellemző artikulációs csatorna keresztmetszeteket modellező csőformációkat mutatunk be a 3.29. ábrán. Az átviteli függvények maximum helyei jól közelítik a fenti két hangra jellemző formánsokat.



3.29. ábra. Az A1, A2 keresztmetszetű cső együtteséből álló két modellillusztráció és átviteli függvényeik. A bal oldal megfelel az [aː], a jobb oldal az [i] beszédhang modellezésének (Stevens 2000)

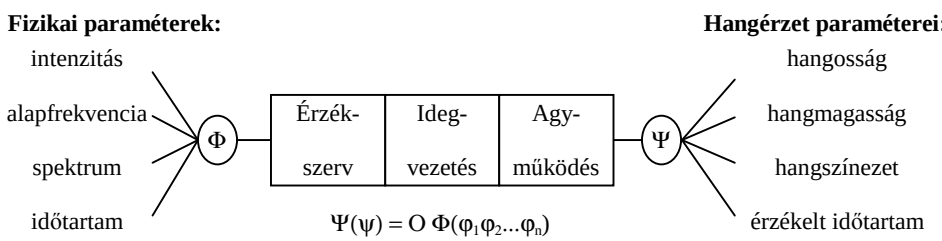
A mássalhangzók artikulációjának ilyen típusú modellezése már jóval bonyolultabb, de Stevens (1972) megmutatta, hogy a cső mentén, a csőben lévő szűkület helyének változtatásával leírható az energiamaximumok kialakulása a képzési hely függvényében.

3.4. Pszichofizikai tényezők

Vicsi Klára

A minket körülvevő világ információi érzékelés útján jutnak el hozzánk: látás, hallás, ízlelés, szaglás, tapintás, hőmérséklet. Mindegyik érzékszervünk csak egy bizonyos ingertípusra és csak korlátozott energiatartományban reagál. Például a szemünk az elektromágneses hullámokból nagyon keskeny frekvenciatartományt érzékel. Az *érzékelés* nemcsak az információ megfelelő érzékszervekkel történő felvételét jelenti,

hanem a kódolást, az átadást, és az információ feldolgozását is, amelyet a központi idegrendszer végez el. A kutatás ezen a területen számos tudományág összehangolt munkáját igényli, mint például a fizika, pszichológia, fiziológia, mérnöki tudományok, matematika stb. Számos érzékelési képesség velünk született, mások pedig tapasztalat és tanulás útján szerezhetők meg, vagy kifejleszthetők. Az ingerek és a szubjektív érzékelés közötti kapcsolat tanulmányozása a pszichofizika alapvető témaköre. A tudományterület elnevezését Gustav Fechnertől származtatják. Ő próbálta meghatározni a mennyiségi kapcsolatot az inger és az érzékelés között (*Fechner-törvény)*. E törvény azt mondja ki, hogy az ingerek *sokszorozódással* nőnek, az érzékelés viszont *hozzáadással*. Például, ahogy a hang intenzitása megduplázódik, az érzeti oldalon a hangerő egy lépéssel nő a skálán. A matematikusok az ilyen viszonyt logaritmikusnak nevezik; Fechner törvénye állítja, hogy az érzékelés az inger logaritmusával nő. Fechner azzal érvelt, hogy ugyanazon viszony alkalmazható bármely ingerre, és az annak megfelelő érzékelésre: például a fényre és a látásra stb. A legutóbbi felfedezések rámutattak arra, hogy ez nem állja meg a helyét, bár pontos matematikai leírást a komplex összefüggések miatt még ma sem fogalmaztak meg. A hangélmény kialakulását a fül mint érzékszerv, valamint a hallási idegvezetés és az agyműködés együttesen határozza meg. Ez valójában egy nemlineáris átviteli rendszer. A hang mérhető fizikai paraméterei: az intenzitás, az alapfrekvencia, a spektrum, az időtartam, az irány stb. A kiváltott hangérzet a hangosság, a hangmagasság, a hangszínezet, a tartósság (érzékelt időtartam) és az irányérzet. A fizikai és az érzetoldal között bonyolult kapcsolat van (3.30. ábra). A hangérzet paramétereinek mindegyike függ egy



3.30. ábra. A hang mérhető fizikai paraméterei és az általuk kiváltott hangérzet közötti kapcsolat

vagy több mérhető fizikai paramétertől. Például a hangosság főként a hangnyomástól függ, de a hang időtartama, spektruma szintén befolyásolják a hangosságérzet kialakulását. A hangmagasság érzete főként az alapfrekvenciától függ, de enyhe függést mutat a hangnyomástól és az időtartamtól is. A hangérzet minőségének a fizikai paraméterektől való függőségét szemlélteti a 3.2. táblázat, ahol a szintek jelölései a következők: + = gyengén függő; ++ = mérsékelten függő; +++ = erősen függő.

A hangérzet és a hang fizikai paramétereinek bonyolult összefüggését már a 3.2. fejezet 3.5. ábrája is szemlélteti, ahol a hallásküszöb, tehát az éppen meghallható

3.2. táblázat. A hangérzet fizikai paraméterektől való függősége

Fizikai paraméterek	Hangérzetek			
	Hangosság	Hangmagasság	Hangszín	Időtartam
Intenzitás	+++	+	+	+
Frekvencia	+	+++	++	+
Színkép	+	+	+++	+
Időtartam	+	+	+	+++

hang intenzitásszintjének erős függése látható a frekvencia függvényében. A fájdalomküszöb, tehát az a hangintenzitásszint, amely már fájdalmat okoz, szintén frekvenciafüggő, de nem olyan nagy mértékben, mint a hallásküszöb. A fizikai paraméterek és a szubjektív hangérzet közötti kapcsolat kifejtése magyar nyelven Tarnóczy (1984), angolul például Hamill–Price (2008) munkájában olvasható.

3.4.1. Hangosságérzékelés

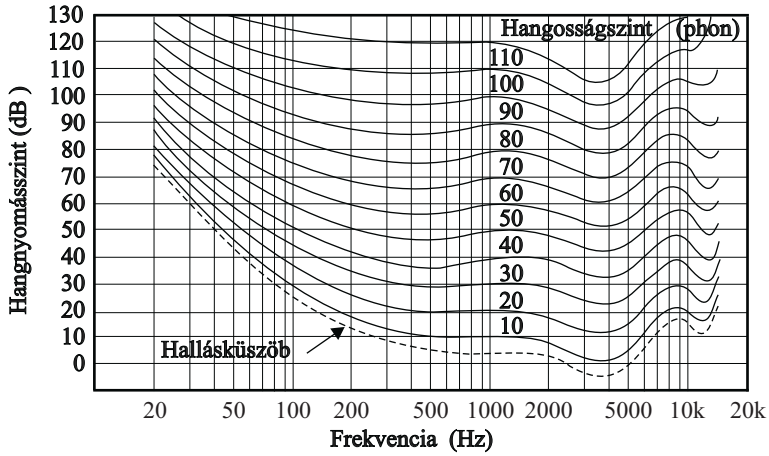
Egy hang bizonyos fizikai hangnyomásszintje bizonyos hangosságérzetet vált ki. A kérdés az, hogy milyen összefüggés van az inger és az érzet között tisztahangok, illetve összetett hangok esetében? Amikor a hangok nagyobb hangintenzitás- vagy hangnyomásszinttel közvetítődnek a fülbe, általában a hangerősség növekedésének érzetét váltják ki, hangosabban szólnak, ez azonban nem minden esetben van így. Ez azt jelenti, hogy a fizikai paraméterek keresztbe is hatnak egymásra a biológiai feldolgozás során.

3.4.1.1. Tisztahangok hangosságérzékelése

A fül érzékenysége, vagyis az agyban keletkezett hangosságérzet nem csak a hangnyomás szintjétől függ, hanem a frekvenciaszerkezettől és a hang minőségétől is. A tárgyalás világosabbá tétele érdekében külön tárgyaljuk a tisztahangok és a komplex hangok érzékelését.

A hangosságszint, [*phon*]. Az alapvető kérdés az volt a 20. század elején, hogy ugyanolyan hangosnak hallunk-e egy rezgést, ha annak frekvenciáját változtatjuk, amplitúdóját viszont nem. Az egyenlő hangosságérzet frekvenciafüggését, sok emberrel elvégzett lehallgatási kisérletekkel (szinuszos hangot alkalmazva) határozták meg (Fletcher–Munson 1933). Azt kérdezték a kísérleti személyektől, hogy mikor hallják egyenlő hangosságúnak a jobb és bal fülükben megszólaló különböző frekvenciájú tisztahangot (miközben az egyik hang amplitúdóját változtatták). Így határozták meg az egyenlő hangosságszinthez tartozó hangnyomásszintadatokat

a frekvencia függvényében. Ezeket a görbéket egyenlő hangosságszintgörbéknek nevezték el. A görbéket tiszta szinuszos hangokra vonatkoztatva a 3.31. ábrán láthatjuk az International Standards Organization alapszabvány (ISO 226 2003) ajánlásának megfelelően, szabad hangtérben mérve. A görbékről elsősorban az



3.31. ábra. Egyenlő hangosságszint [phon] görbék szinuszos hangok esetében, szabad hangtérben mérve, ahol a hangforrás a hallgatóval szemben volt elhelyezve. Az ISO 226: 1987 alapszabvány adatai láthatók

olvasható le, hogy a fül hangosságérzete frekvenciafüggő. A hallási érzékenység gyengébb az alacsony és a magas frekvenciákon, a legjobb a 3500 és 4000 Hz közötti frekvenciákon (ami közel áll a külső fül járatának első rezonanciafrekvenciájához). Hasonló, enyhébb érzékenységi szakasz látszik 13 kHz környékén, ami a második rezonanciafrekvenciával hozható kapcsolatba. Fontos megjegyezni a fül alacsony frekvenciájú hangokra való relatív érzéketlenségét. Leolvasható továbbá a legalsó görbén a hallásküszöb (ami 1000 Hz-en megfelel 20 μPa nyomásértéknek), valamint annak frekvenciafüggése. A hallásküszöb az a hangnyomásszint, amit az ember süketszobában még éppen meghall (nagyjából egy szúnyog repülésének hangja 3 méterről). Az egyenlő hangosságszint kontúrjait mint szubjektív érzeti mértékegységet phonnak nevezett egységenként jelölik, a phonban megadott szint numerikusan egyenlő a decibelekben megadott hangnyomásszinttel az f = 1000 Hz esetében. Egy adott frekvenciájú hang erőssége tehát annyi phon, ahány dB a vele azonos hangosságérzetet keltő 1 kHz-es szinuszos hang hangnyomásszintje. Más frekvenciákon a görbékről olvasható le a phon-dB viszony. A görbékből az is leolvasható, hogy a hangintenzitás növelésével a frekvenciafüggőség szintje csökken. A hangosságérzet erős frekvenciafüggése az egyik oka annak, hogy miért használnak a hangosságszint-mérő eszközöknél különböző súlyozógörbéket. A hangszintmérők egy vagy több súlyozó görbével rendelkeznek, amelyek biztosítják a hallásnak megfelelő frekvenciafüggés figyelembevételét a hangszintméréseknél.

2003-ban nemzetközi együttműködéssel az alapszabvány módosítása jelent meg (ISO 226 2003). Az új és a régi szabványadatok között 1000 Hz alatt 15 dB-es küszöbemelkedés is előfordul, 1000 Hz felett a görbékben csak néhány dB-es eltérés van.

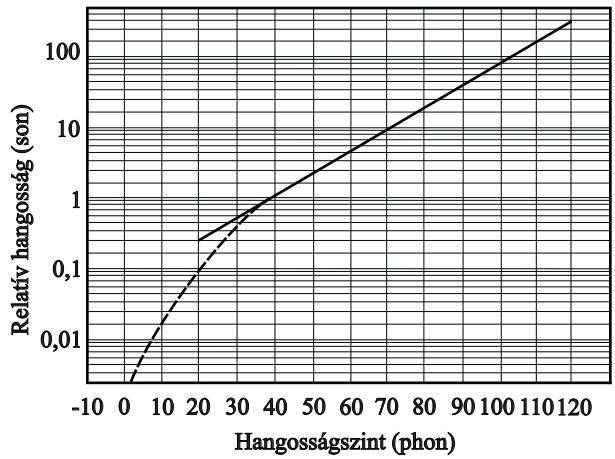
Relatív hangosság [*son*]. A relatív hangosság meghatározásánál arra vagyunk kíváncsiak, hogy mikor hallunk egy adott ingerhangot kétszer, háromszor stb. hangosabbnak (érzeti szinten), mint az eredeti mintahang, frekvenciától függetlenül. A hangosságszint és a hangosság érzete közötti összefüggést mutatja a 3.32. ábra. A relatív hangosságérzet egysége a son. Két son kétszeres, tíz son tízszeres hangosságérzetet jelent. A skálát rögzíteni kellett a frekvenciafüggő hangosságszinthez. Ez a rögzítési pont a következő 1 son = 40 phon és 1000 Hz-en 40 dB. A hangosság és a hangosságszint közötti összefüggést a 3.32. ábra mutatja. Például egy 40 phon hangosságszintű hangot, hogy kétszeres hangosságúnak érezzük 50 phon hangosságszintre kell emelni. Az 50 phon 1000 Hz-en 50 dB hangnyomásszintnek felel meg, viszont egy 100 Hz-es hang esetében a kétszeres hangosságérzet eléréséhez szintén 50 phon hangosságszint szükséges, ami megfelel 60 dB hangnyomásszintnek. A 40 phonos vagy annál nagyobb hangossági szinten a sonban megadott S hangosság és az L_L hangossági szint közötti viszony phonokban az International Standards Organization (ISO 532 1975) által ajánlva a következő:

$$S = 2^{\frac{L_L - 40}{10}} \tag{3.21}$$

A son-skála valóságos érzeti skála. Lineáris a kapcsolat a son-értékek között, tehát 1 son + 1 son = 2 son, *holott* 40 phon + 40 phon = 43 phon volna a logaritmizálási szabályok szerint. Látható, hogy milyen nagy eltérés van a valóságos hangosság alakulása és a hangosságszint alakulása között, ugyanis 2 son nem 43 phonnak, hanem 50 phonnak felel meg.

3.4.1.2. Összetett hangok hangosságérzékelése

A hangosság, mint azt az előző fejezetben említettük, főként a hangnyomástól függ, de az alapfrekvenciával, a spektrummal és az időtartammal is változik. Azt már láttuk, hogyan függ a hangosság az alapfrekvenciától; most pedig összehasonlítjuk a spektrális összetevőktől való függőségét. Az összetett hangok hangosságszámítása a gyakorlatban fontos. Például hány hegedűnek kell játszania egyszerre, hogy kétszer olyan hangosan szóljanak? Milyen hangosságérzetet ad, ha két ember egyszerre ugyanolyan hangerővel beszél? Hogyan függ a forgalom zaja a járművek számától? Amikor az intenzitások két vagy több össze nem függő hangforrásból erednek, együtt adják meg a teljes intenzitást. A hangosságérzet nem a fizikai intenzitás-

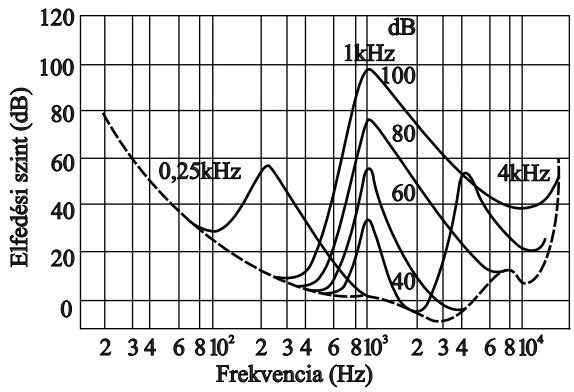


3.32. ábra. A hangosság [son] és a hangosságszint közötti [dB] összefüggés (Tarnóczy 1984)

összegzés eredményét adja. Amikor két vagy több hang összekeveredik, hogy azt milyen hangosnak halljuk, attól függ, hogy az összetevő hangok frekvenciaszerkezete milyen. Ha a hangok frekvenciája egyezik, vagy a kritikus sávszélességen belül esik, akkor a hangosságot a teljes intenzitásból kell kiszámítani, tehát a fizikai törvényszerűség érvényesül: a hangosság az intenzitások összegzéséből adódik. Ha a sávszélesség meghaladja a kritikus sávszélességet, az eredményül kapott hangosság nagyobb, mint amit az intenzitások egyszerű összegzéséből nyertünk. Ahogy a sávszélesség nő, a hangosság megközelíti azt az értéket, ami az egyéni hangosságok összege (de annál alacsonyabb marad). Ha a frekvenciakülönbség nagyon nagy, az összegzés komplikálttá válik. A hallgatók hajlamosak elsődlegesen egy komponensre koncentrálni (például a leghangosabb vagy a legmagasabb csúcsok egyike), és egy teljes hangosságot a komponens hangosságához közel egyenlőnek jelölnek meg (Roederer 1975). A sokkomponensű komplex hangok hangossága sonban meghatározható az következők szerint: mérendő oktáv-, vagy pedig az 1/3 oktávsávokban a hangintenzitás, majd a szabványosított diagramok alapján (ISO 532 1975) a sávokban mért értékek összegzendők. Az oktávsávok olyan frekvenciasávok, amelyek egy oktáv szélesek (azaz a maximumfrekvencia kétszerese a minimumfrekvenciának). Ma már hangosságmérő eszközök, programok léteznek, amelyek az ISO 532:1975 szabvány alapján megadják egy összetett hang hangosságát sonban.

Hangelfedés. Amikor a fül két vagy több különböző hangingernek van kitéve, az egyik elfedheti a másikat. Az elfedés függ attól, hogy a hangok frekvenciában és időben milyen távolságban vannak egymástól, valamint hogy milyen intenzitásúak. A hangelfedést felhasználják a beszédtechnológiában is hangtömörítésre.

Hangelfedés a frekvenciatartományban. A frekvenciaelfedés azt jelenti, hogy az egymáshoz közeli frekvenciakomponensek elfedik egymást. Ezt fiziológiailag úgy is mondhatjuk, hogy az elfedett frekvenciákon megemelkedik a hallásküszöb az elfedő hang hatására. Ez az elfedés függ a frekvenciatávolságtól és függ az egyes frekvenciakomponensek intenzitásától. Zwicker kísérleteket végezett keskeny sávú zörejekkel, amelyek sávszélessége kisebb vagy egyenlő a kritikus sávszélességgel. Azt határozták meg, hogy az ilyen zörejek milyen mértékben fedik el a környező frekvenciákat, azaz milyen módon változtatják meg a hallásküszöböt. A 3.33. ábrán a szaggatott vonal a teljes csöndben mért hallásküszöbgörbét mutatja,



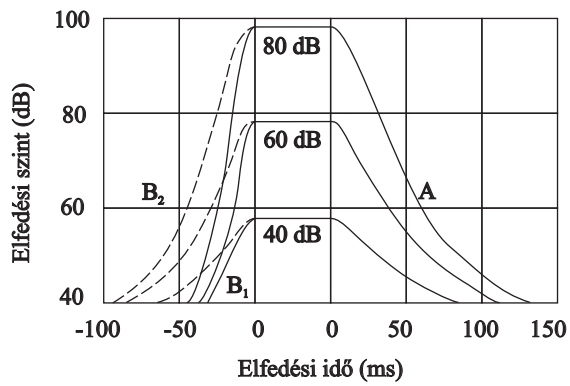
3.33. ábra. Keskeny sávú zörejek elfedő hatása szinuszhangokra 250 Hz, 1 kHz és 4 kHz sávközép-frekvenciájú keskeny sávú zörejek esetében. Az 1kHz-es keskeny sávú zörej növekvő intenzitásával mutatja az elfedési görbe növekedését és kiszélesedését is (Zwicker 1982)

vagyis a frekvencia függvényében azokat az intenzitásszinteket, amiket még éppen meghallunk. A folyamatos vonalak pedig a megváltozott hallásküszöböt mutatják, különböző keskeny sávú zörejek jelenlétében, szinuszos hangra vonatkoztatva. Például az 1 kHz-es, 100 dB intenzitásszintű keskeny sávú zaj nem hat a 200 Hz frekvenciájú szinuszos hang hallásküszöbére. Ugyanakkor 500 Hz frekvenciájú hangra oly módon hat (elfedi), hogy annak hallásküszöbe 25 dB-lel megemelkedik. Az 1 kHz-es keskeny sávú zörejhez képest minél közelebbi frekvenciát vizsgálunk, annál jobban megemelkedik a hallásküszöb. A 990 Hz frekvenciájú színuszhangnál már 97 dB-es hallásküszöb-növekedés olvasható le. Az ábra azt is mutatja, hogy különböző intenzitású 1 kHz-es keskeny sávú zörejek (80, 60, 40 dB-en) hogyan hatnak az elfedési görbe alakulására.

Összefoglalva, azok a tisztahangok vagy keskeny sávú zörejek, amelyeknek a frekvenciája közel van egymáshoz, jobban elfedik egymást, mint azok a hangok, amelyek frekvenciában távol esnek egymástól. Egy tisztahang jobban elfedi a magasabb frekvenciájút, mint az alacsonyabbat, vagyis az elfedés frekvenciában erősen aszimmetrikus. Minél nagyobb egy elfedő hang intenzitása, annál szélesebb

frekvenciatartományban képes az elfedésre. Ha két hang frekvenciában messze esik egymástól, kismértékű lehet az elfedés, vagy egyáltalán semmilyen elfedés nem történik. Ezek az azonos idejű elfedési jelenségek akkor érthetők meg igazán, ha figyelembe vesszük, hogyan ingerlik a tisztahangok az alaphártyát. A magas frekvenciájú hangok az alaphártyát az ovális ablak közelében ingerlik (ott a legkeskenyebb), míg az alacsony frekvenciájú hangok a végén hozzák létre a legnagyobb amplitúdót (ott szélesebb az alaphártya). A tisztahang keltette ingerlés aszimmetrikus, a magas frekvenciájú rész felé nyúlik. Így könnyebb elfedni egy magasabb, mint egy alacsonyabb frekvenciájú hangot. Ahogy az inger intenzitása nő, a nagyobb kimozdulással az elfedés erőssége is nő. A tisztahangok, a komplex hangok, a keskeny és széles sávú zajok mind eltérő módon képesek más hangok elfedésére (Tarnóczy 1984). A széles sávú (fehér) zaj által történő elfedés hozzávetőlegesen lineáris viszonyt mutat az elfedés és a zajszint között (azaz a zajszint 10 dB-lel való megemelése ugyanennyivel növeli a hallásküszöb szintjét). A széles sávú zaj tehát az összes frekvencián elfedi a hangokat.

Hangelfedés az időtartományban. Az egymás után bekövetkező hanghatásoknál akkor következik be időbeli elfedés, ha az időkülönbség kicsi. Az ilyen elfedés mindkét irányba felléphet. A pontos értékeket a 3.34. ábra mutatja 40, 60, és 80 dB intenzitású fehérzaj hatására. Az előre történő elfedés (előelfedés) azt jelenti, hogy



3.34. ábra. Az elfedés időbeli alakulása. Az előre történő elfedést az A görbék mutatják; a visszafelé történő elfedést a B_1 görbesereg azonos fülben, B_2 különböző fülben (Tarnóczy 1984)

az elfedő hang az elhangzása után megjelenő hangot elfedi. Az A jelű görbesereg szerint ez az elfedés az elfedő hang intenzitásától függ (20–100 ms). Az előre történő elfedést az okozza, hogy a stimulált sejtek egy darabig nem olyan érzékenyek, mint a nyugalomban lévők.

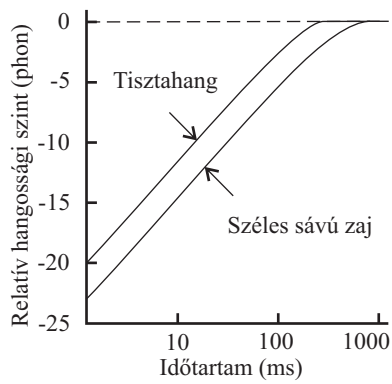
A visszafelé történő elfedés (utóelfedés) azt jelenti, hogy a később megjelenő hang fedi el a korábbi hangot. A B_1B_2 görbesereg egy olyan hang által elfedett hangra

vonatkozik, amely 25–100 ms-mal később kezdődik. Egy hangot elfedhet egy zaj, amely tíz milliszekundummal később kezdődik, bár az elfedés mértéke csökken, ha az időintervallum a két jel között nő (Elliott 1962). A visszafelé történő elfedést az idegi feldolgozási sebességekkel magyarázzák. A nagyobb intenzitású később megjelenő inger megelőzi a korábbi gyengébbet, ezért elfedi azt.

Az egyik fülben az elfedést okozhatja a másik fülben megjelenő zaj, bizonyos körülmények között; ezt központi elfedésnek nevezzük.

3.4.1.3. Hangosság és időtartam

A hangérzet hangossága erősen időfüggő a 200 ms alatti időtartományban. Például egy 25 ms időtartamú rövid idejű hang hangosságérzeti értéke csak a fele a 200 ms-os, vagy annál hosszabb hang hangosságérzetének. A hangosságérzeti szint az inger időtartamának függvényében változik (3.35. ábra).



3.35. ábra. A relatív hangossági szint változása a hang időtartamának függvényében (Zwislocki 1969)

3.4.2. Hangmagasság-érzékelés

A hangmagasság szubjektív érzet. Egy tisztahangnál a hangmagasságot főként a frekvencia határozza meg, bár egy tisztahang hangmagassága a hangosságszint szerint is változhat. A komplex hangok hangmagassága a hang spektrumától (hangszínétől) is függ és annak tartósságától.

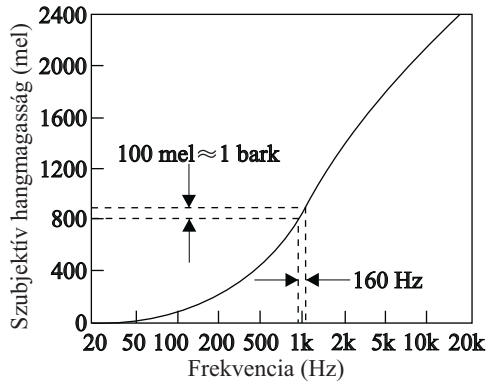
3.4.2.1. Hangmagasságskálák

A hangmagasság a hang egy olyan jellemzője, ami magas vagy mély érzetűvé teszi a hangot, vagy pedig egy skálán a hang elfoglalt pozícióját határozza meg. Két személy, akik ugyanazon hangot hallják, a hangmagasságskálán azt különböző pozícióba jelölhetik.

Zenei hangmagasságskálák. A legtöbb zenei skála alapegysége az oktáv. Az egy oktávot alkotó hangok frekvenciája közel áll (de mint látni fogjuk, nem mindig pontosan) a 2:1 arányhoz. A zenében az oktáv kétféleképpen osztható fel. A nyugati zene általában 12 intervallumra osztja az oktávot, amelyeket *félhang*nak nevezünk; ezek adott hangnevek (A-tól G-ig, fél hanggal felemelve vagy leszállítva), kijelölve a kotta öt vonalán.

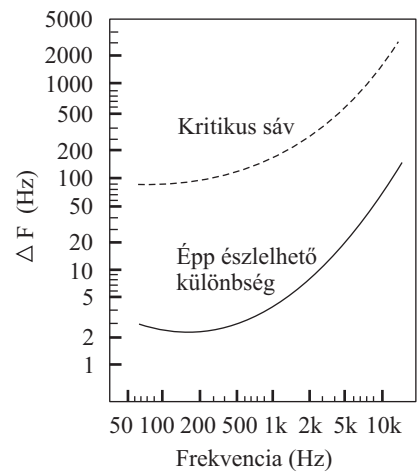
Pszichofizikai hangmagasságskálák [mel] és [bark]. Egy pszichofizikai hangmagasságskála felállítására számos kísérletet tettek már. Ha egy átlagos hallgató meghallgat egy 4000 Hz-es hangot, amelyet egy jóval alacsonyabb frekvenciájú hang követ, azután pedig megkérik, hogy hangoljon be egy oszcillátort a hallott két hang közötti távolság felére, akkor a legvalószínűbb választás 1000 Hz körül lesz. A hangmagasságskálán így az 1000 Hz-et 0 és 4000 Hz között félúton lévőnek ítélték meg. A szubjektív hangmagassághoz használt mértékegység a mel; a skálát úgy állították be, hogy az értékek megduplázása a hangmagasságérzetet is megduplázza, és 131 Hz felejen meg 131 melnek. Ekkor a 0–16 kHz közötti frekvenciatartományt a 0–2400 mel értéksorral jellemzik. A mel és hertz értékek közötti összefüggést a 3.36. ábra szemlélteti. Egy másik pszichofizikai skála a hallás kritikus sávjain alapul és kritikus, sávarányú skálának nevezik (részletezve a 3.2.2 fejezetben). Egy kritikus sávszélességet egy barknak neveznek, és a hangmagasság értékeit a bark-sávok sorszáma fejezi ki. Egy bark megközelítőleg egyenlő száz mellel.

Hangmagasság-megkülönböztető képesség. Annak a képessége, hogy meg tudunk különböztetni két közel egyforma ingert, gyakran kerül jellemzésre a pszichofizikai tanulmányokban. Az éppen észlelhető különbség (just noticeable difference, jnd) kifejezéssel jellemzik azt a percepciós képességet, amikor két közel egyforma ingert meg tudunk különböztetni. Az éppen észlelhető különbség a hangmagasság esetében függ a frekvenciától, a hangerősségszinttől és a frekvenciaváltozás gyorsaságától, továbbá a hallgató zenei affinitásától és bizonyos mértékig a mérési módszerektől is. Az átlagos jnd-t 80 dB szintű tisztahangok esetében a frekvencia függvényében a 3.37. ábra szemlélteti. 1000 és 4000 Hz között a jnd a tiszta színuszos hang frekvenciájának mintegy 0,5 százaléka, ami egy félhangnak körülbelül egytizenkettede. Néha a frekvenciafelbontás kifejezést is használják a



3.36. ábra. Hangmagasságskála a frekvencia függvényében. A hangmagassági skálán 100 mel, megfelel az 1000 Hz középfrekvenciájú kritikus sáv sávszélességének ami 160 Hz (Rossing 1990)

frekvenciák által meghatározott jnd megjelölésére. Összehasonlítva a 3.37. ábra felső és alsó görbéit, látható, hogy a kritikus sávszélesség alakulása tendenciájában hasonló a jnd-vel. Ez a jelenség mutatja, hogy a fülben ugyanez a mechanizmus felelős a kritikus sávokért és a hangmagasság-diszkriminációért. Valószínűleg az alaphártya gerjesztési területeivel van összefüggésben.

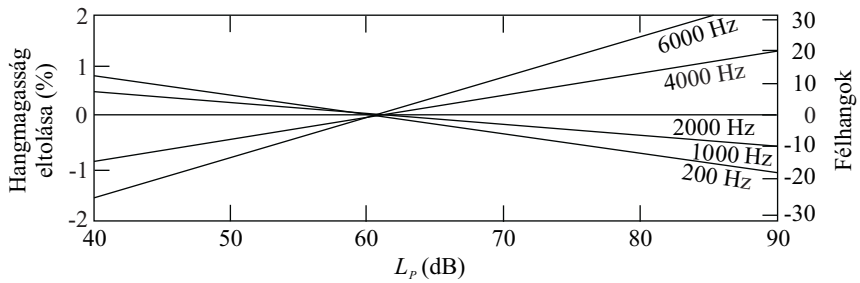


3.37. ábra. Az éppen észlelhető frekvenciakülönbség és a kritikus sávszélesség a frekvencia függvényében (Zwicker et al. 1957)

3.4.2.2. Tisztahangok hangmagasságérzete

Azt már láttuk, hogy a hangmagasság főként az alapfrekvenciától függ. Most pedig vegyük figyelembe a tisztahangok más fizikai mennyiségektől való hangmagasságérzékelési függőségét, ezek például a hangintenzitás, a tartósság és más hangok jelenléte.

Hangmagasság és hangenergiaszint. Korábbi hangmagasság kontra hangenergiaszint-kísérletek a hangmagasság hangszinttől való függőségét mutatták színuszos hangok esetében (3.38. ábra). Az alacsony frekvenciájú hangok hangmagasságérzete a növekvő intenzitással csökkent; a magas frekvenciájú hangoké növekvő intenzitással nőtt, a közepes frekvenciájú hangok (1–2 kHz) pedig csak kis mértékű változást mutatott. A hatás kismértékű, még tisztahangoknál is, és megfigyelőről megfigyelőre váltakozik. Kevesebbet tudunk a komplex hangok hatásairól. A hangszerekkel végzett tanulmányok az intenzitással nagyon kismértékű hangmagasság-változásokat mutattak. Az, hogy egy komplex hang érzeti szinten magasabb vagy alacsonyabb a növekvő intenzitással, attól függ, hogy milyen részösszetevő van túlsúlyban, 1000 Hz alatti vagy feletti (Terhardt 1979). Szerencsés



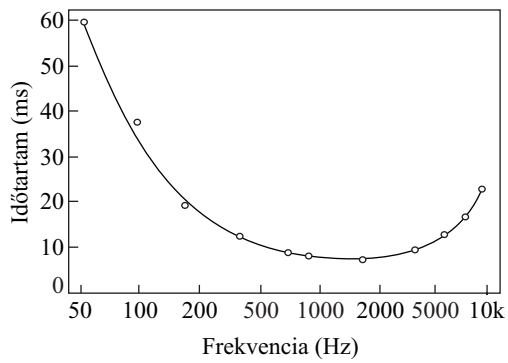
3.38. ábra. Tisztahangok frekvenciaeltolódása a hangnyomásszint változásának függvényében. A görbék 15 személy vizsgálatán alapulnak (Terhardt 1979)

dolog az előadózenészek és a hallgatók számára, hogy a hangenergiaszinttel történő hangmagasság-változás a komplex hangok esetében sokkal kevesebb, mint a korábbi kísérletek során a tisztahangoknál. A zenei előadás nagyon bonyolult lenne, ha lényeges hangmagasság-változások történnének a dinamikai szint változásai során.

Hangmagasság és tartósság. Milyen hosszan kell egy hangnak szólnia, hogy azonosítható hangmagassága legyen? A nagyon rövid hangokat kattanásként érzékeljük, ám ahogy hosszabbodnak a hangok, az érzet szintjén egyre jobban érzékelhető a hangmagasság (3.39. ábra). A hangmagasságnak a tartósságtól való függése egyfajta „akusztikai bizonytalansági alapelvet” követ:

$$\Delta f \Delta t = K, \tag{3.22}$$

ahol Δf a frekvenciabeli bizonytalanságot jelenti, a Δt pedig a hang tartósságát. Optimális körülmények között K kevesebb, mint 0,1 (Majerník–Kaluzny 1979). Tehát az érzeti oldalon is érvényes az a fizikai tétel, mely szerint a frekvencia és az időfelbontás szorzata állandó. Az, hogy a „kattanás” mikor megy át periodikus hang



3.39. ábra. Adott hangmagasság kialakulásához szükséges időtartam

érzetébe, az időtartamon kívül a hangenergiaszinttől is függ; vagy ha a hang nem váratlanul szólal meg, hanem egy enyhe indítással, akkor a hangfelismerési idő 3 ms is lehet, ami rövidebb, mint a legtöbb hangszer megszólalási ideje (Winckel 1967).

Hangmagasság és burkológörbe. Egy rövid, exponenciálisan lecsengő szinuszhang érzékelt hangmagasságát következetesen magasabbnak találták, mint egy egyszerű szinuszhang hangmagasságát ugyanazon frekvenciával és energiával (Hartmann 1978). A hangmagasság-eltolódás éppúgy függ a hangnyomásszinttől, mint a hang amplitúdójának emelkedési, illetve esési sebességétől (Rossing–Houtsma 1986). A hangmagasság burkológörbe-függőségének köze van a már korábban tárgyalt intenzitással való hangmagasság-eltolódáshoz. Ez mindenképpen egy olyan hatás, amelyet a zenészeknek figyelembe kell venniük, amikor ütőhangszerek hangmagasságával foglalkoznak.

Környezeti hangok hatásai. A hangok ritkán hallhatóak elszigetelten. A környezeti hangok jelenléte szintén hatással van a tisztahangok hangmagasságérzetére. Azok a kísérletek, amelyeket zavaró hanggal vagy zajjal elvégeztek, a következőket összegzik. Ha a zavaró hang frekvenciája a teszthang alatt van, akkor mindig felfelé való eltolódás történik. Ha a zavaró hang frekvenciája a teszthang felett van, akkor lefelé való eltolódást figyeltek meg alacsony frekvencián. A zavaró zaj mindig felfelé való hangmagasság-eltolódást okoz, ha alacsonyabb volt a frekvencia, mint

a teszthang (de ha magas volt a frekvencia, akkor az eltolódás mindkét irányba megtörténhet). A hangmagasság-eltolódás annyival nő, amennyivel a zavaró hang vagy zaj amplitúdója meghaladja a teszthangot (Terhardt–Fastl 1971).

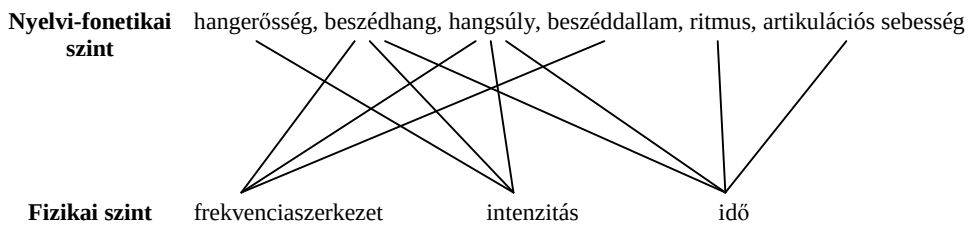
3.4.2.3. Komplex hangok hangmagassága, virtuális hangmagasság

Amikor a fül egy harmonikusokból álló komplex hangot hall, könnyű megjósolni, hogy milyennek ítéljük a hangmagasságot. Egyszerűen a legalacsonyabb frekvenciaösszetevőt fogja az érzet hangmagasságként megadni. Ugyanakkor a fül képes azonosítani a hangmagasságot, még akkor is, ha az alaphang nagyon gyenge, vagy hiányzik. Például, ha a fül egy olyan hangot hall, amelynek 600, 800, 1000 és 1200 Hzes frekvenciájú részösszetevői vannak, a hangmagasságot csaknem mindig 200 Hzes hangként azonosítja. Ilyen esetben beszélünk *virtuális hangmagasság*ról. A fül azon képessége, hogy meg tud határozni egy virtuális hangmagasságot, lehetővé teszi, hogy vezetékes telefonátviteli sávban (300 Hz és 3400 Hz közötti sáv) határozott alaphangmagasságokat halljunk, annak ellenére hogy az alaphangot, ami férfiaknál 80–150 Hz és nőknél 150–300 Hz, az átvitelnél kiszűrjük. Felmerül a kérdés, mely harmonikusok a legfontosabbak a virtuális hangmagasság érzékeléséhez. A kísérletek azt mutatták, hogy egy 200 Hz körüli alapfrekvenciával hangzó komplex hang esetében a hangmagasságot főként a negyedik és az ötödik harmonikus határozza meg. Amikor a komplex hangok részösszetevői nem harmonikusok, a virtuális hangmagasság meghatározása még finomabb. A hangmagasság legutóbbi elméletei szerint a fül kiválaszt egy sorozat csaknem harmonikus részösszetevőt valahonnan a hallási tartomány közepéből, és ez alapján határozza meg a hangmagasságot (Houtsma et al. 1988). Jó zenei példák a csengők és harangok hangjai a hallórendszer virtuális hangmagasságának a keletkezésére nem harmonikus komplex hangok esetén. Az ilyen zengő hang hangmagasságát minden esetben főként három részösszetevő határozza meg, amelyek frekvenciái majdnem 2:3:4 arányban állnak egymással.

3.5. Fizikai-nyelvi megfeleltetések

Olaszy Gábor

A beszédben mint akusztikai jelben a beszélő minden közlési szándéka benne van, a saját jellemző hangszínezetével és fizikai állapotával együtt. Ahhoz, hogy eredményesen tudjuk ezt a fizikai jelet jellemezni, párhuzamot kell vonni a beszédre vonatkozó legfontosabb fizikai és nyelvi-fonetikai jellemzők között (3.40. ábra) Látható, hogy ugyanazon fizikai paraméter több nyelvi szinten is megjelenhet. Nyelvi szinten a két legösszetettebb szerkezetű elem a beszédhang és a hangsúly. Mindkettő



3.40. ábra. Kapcsolatrendszer a fizikai és a nyelvi jellemzők között

kialakításában mindhárom fizikai paraméternek van szerepe. Fizikai szinten pedig a paraméterek mindegyike több nyelvi elem kialakításában is aktív. Az alapfrekvencia értéke például adhat a személy nemére, életkorára vonatkozó információt a beszéddallamba ágyazva, de szerepe van a hangsúlyozásnál is, és utalhat a mondat modalitására is. Az intenzitás szintje a beszédstílust, esetleg az érzelmi töltetet határozhatja meg (hangos, halk, mérges, szomorú), ugyanakkor tartozhat egy hangsúlyhoz is (ilyenkor a szótagra vonatkozik), de fontos tisztában lenni a relatív hangossággal is, ami a beszédhangokat jellemzi. Az időszerkezet is legalább ennyire sokrétű a beszéd esetében. Ha tehát fizikai paramétereket nyerünk ki egy beszédjelből, akkor tudnunk kell, hogy az adott paramétert mely nyelvi jellemzőhöz akarjuk kötni, vagyis hogy mit vizsgálunk.

Az időszerkezet képezi az alapot, ugyanis a két másik akusztikai komponens változása (hangintenzitás, spektrális tartalom) az időszerkezetbe ágyazódva valósul meg. Ha a beszéd időszerkezetét nyelvi szempontból szegmentáljuk, akkor a legkisebb szerkezeti elem (5–10 ms) a zöngés hangperiódus. Idesorolhatók még bizonyos gyorslefolyású hangrészek (zárfelpattanások) is mint egyes beszédhangok fontos és jellemző részei. A hosszabb időintervallumok felé haladva számos olyan paraméter lehetséges, amit az időszerkezettel jellemezhetünk, például a hangok hossza, a hangok belső időszerkezeti részei (hangátmenetek, zárszakaszok, zárfeloldódások). Tovább menve jutunk el a hangkapcsolódások, majd a szótagok, szavak hosszához és végül a jellemző dallamformák kezdeti és végpontjai közötti időintervallumokhoz. Vizsgálhatjuk a hangsúlyok között eltelt időtartamot (fonetikai frázis), a szünetek hosszát, a közöttük eltelt időszakaszt, a zöngés-zöngétlen hangszakaszok hosszát, a nem szorosan a beszédhez tartozó elemeket (hangos szünet). Amennyiben mondatszintű elemzést végzünk, érdekes lehet a mondatok hosszának vizsgálata is. Jelfeldolgozás szempontjából szokásos a beszéd időfüggvényén mérhető nullátmenetek közötti időtartamokat, azok eloszlását is vizsgálni. Ezek képezik talán a legrövidebb időszegmenst, akár 1 ms-nál is rövidebb időszakaszokat.

A hangintenzitás tekintetében is több szinten kell gondolkodni. A legalacsonyabb szint a hangokat érinti. Minden beszédhangnak saját, jellemző intenzitásszintje van, ami azt jelenti, hogy a környező hangokhoz képest milyen intenzitással van jelen a beszédhangsorban. Ezek a jellemző intenzitásviszonyok érvényesülnek a folyamatos

beszédjel bármely pontján. A másik szint a beszéd folyamatának pillanatnyi intenzitásszintje (mikor halkabb, mikor hangosabb). Ezt lehet értelmezni mondaton belül és mondatok között is. Az intenzitással függ össze a hangsúlyozás is. A hangsúlyos szótagok intenzitása általában nagyobb, mint a környezetüké.

A beszédben az alapfrekvenciánál magasabb frekvenciaszerkezeti képet az adott nyelv határozza meg az artikulációs bázisán keresztül, tehát a pillanatnyi spektrumot úgy tekinthetjük, mint az artikuláció pillanatnyi akusztikai vetületét (lásd később részletesen). Nyelvi szempontból a zöngés hangoknál a spektrumot két részre bontjuk, az alapfrekvenciára és a felharmonikusokat tartalmazó frekvenciasávra, mert eltérő a szerepük. Zöngétlen hangoknál ilyen kettéválasztásra nincs szükség. Az alapfrekvencia változását külön szokták mérni és ábrázolni, továbbá annak mesterséges változtatására is külön algoritmusokat fejlesztettek ki.

4. fejezet
A beszéd és az írás kapcsolata

A nyelvi kommunikáció két legfontosabb formája a beszéd és az írás. Míg a beszéd folyamatos levegőrezgés, melynek akusztikai szerkezete állandóan változik, az írás alapvetően a beszéd tartalmi rögzítésére jött létre, ennek megfelelően diszkrét, véges számú egymástól elkülöníthető elemből (graféma, írásjel) álló jelkészletet használ (Nádasdy 2006). A beszéd és az írás kapcsolatrendszerét a beszédtechnológiában hangsúlyozottan veszik figyelembe. Sok esetben írott szöveget kell a gép segítségével automatikusan beszéddé alakítani, tehát a diszkrét szimbólumsorozatból kell folyamatosan változó hullámformát készíteni (beszédszintézis). Ennek fordítottja pedig, amikor az elhangzott beszédből kell a gépnek meghatároznia, hogy mi volt a hangsor nyelvi tartalmát reprezentáló diszkrét elemek sorozata (beszéd-írás konverzió).

4.1. Írásrendszerek

Kiss Géza

Az írás célja kialakulásától fogva az emberi beszéd minél pontosabb rögzítése volt, grafikus formában (Fischer 1999, 86. o.). Az idők folyamán elődeink három lényegileg különböző *írásrendszer*t hoztak létre: a logografikust, a szótagírást és a betűalapút (Fischer 2004, 86–111. o.).

A *logografikus* rendszerekben az írásjelek általában egy-egy szót jelölnek, ezeket nevezzük logogramoknak. A logogramok gyakran úgynevezett piktogramokból, a jelzett tárgy képszerű ábrázolásából fejlődtek ki, azok absztraktabbá válásával. Logografikus írás volt például az ókori egyiptomi, ma pedig például a kínai és japán írás ilyen. Egyik sem teljesen az, mivel mindegyik tartalmaz a hangzást leíró elemeket is. Például a kínai nyelvben a szavak legnagyobb részét két jellel írják le: az egyik a kiejtését adja meg, a másik a szemantikai kategóriáját, ami azonos hangzású, de eltérő jelentésű szavak esetén egyértelművé teszi, hogy melyikről van szó. Néhány

logogramot, illetve ideogramot (egy adott fogalmat jelölő szimbólum) szinte minden nyelv használ, például az €, $ és @ jelek ilyennek vehetők. Sőt, az arab számok is ebbe a kategóriába tartoznak, hiszen szinte minden írásrendszer ezeket használja a számok jelölésére, de a beszélők a saját anyanyelvüknek megfelelő betűsort rendelik a számokhoz (például a 2 szimbólumhoz: *kettő, two, zwei, dwa, dos* stb.).

Szótagírás esetén a leírt szimbólumok szótagokat jelölnek; a hasonló hangzású szótagok írásjelei között rendszerint nincs alaki hasonlóság. Az ókorban először a sumérek használtak szótagírást, később számos közel-keleti nép átvette, és a kereskedelemben is fontos szerepe volt. Ma például egyes afrikai és észak-amerikai indián népek használnak ilyent, valamint a szótagok jelölése megtalálható a kínai nyelvben és a japánban (kanaírás) is.

A betűírás fokozatosan alakult ki. A főníciai és más sémita (arab, héber) írások írásjelei a szavak mássalhangzóit jelölik, ami e nyelvek reprezentálására elégséges. Ezt átvették az ókori görögök, és nyelvük szükségleteinek megfelelően kiegészítették a magánhangzók jelölésével, valamint néhány további mássalhangzóval. Ismereteink szerint így jött létre a nyelv legkisebb szegmentális egységeihez (fonémák) írásjegyet rendelő betűírás. A görög ábécé lett az alapja a latin és a cirill betűknek, amelyek ma sok nyelv írásának szimbólumrendszerét adják. Azokhoz a nyelvekhez, amelyek a legutóbbi időkig nem rendelkeztek írásbeliséggel, szinte kivétel nélkül latin betűs ortográfiát alakítanak ki.

A nyelvek struktúrája nem határozza meg, hogy milyen jellegű írásrendszerrel kell írni: ugyanolyan típusú nyelveket nagyon eltérő írásrendszerekkel rögzítettek már (Saenger 2000, 2. o.). Az, hogy egy nyelv melyiket használja, részben történelmi okokra vezethető vissza, részben arra, hogy az adott nyelvet jól lehet ábrázolni a használt írásmódban. A betűírást kétségtelenül sokkal egyszerűbb megtanulni, mint a logogramírást vagy akár egy szótagírást, az utóbbiak szimbólumainak nagy száma miatt. Viszont kutatások azt igazolták, hogy a szótagalapú írások közelebb állnak az ember természetes beszédfeldolgozási mechanizmusához, mint akár a magyar fonémikus betűírás. Ezt igazolja, hogy az írástudatlan gyermekek és felnőttek számára nehézséget okoz a beszédhangok elkülönítése, ami arra utal, hogy ez a betűírással együtt tanult képességünk; továbbá, hogy olyan feladatokban, amikor a nyelvük szavainak hangalakját kell megállapítani, kevesebbet hibáznak azok, akik szótagírást tanultak, mint akik betűírást (Field 2003).

Léteznek speciális célú írásmódok is, például a matematikai nyelvek, a programnyelvek, a zene nyelvének rögzítésére készült kották stb. Íráson a továbbiakban olyan rendszert fogunk érteni, amelynek közvetlen célja a beszélt nyelv rögzítése.

Az írás univerzáléi. Amint a beszélt nyelvek esetében léteznek mindre érvényes jellemzők, ez még nyilvánvalóbban igaz az írásrendszerekre (Coulmas 1999, 531. o.). Ilyen közös jellemzője mindegyik írásrendszernek:

- a nyelvreprezentáció: a (beszélt) nyelvet jeleníti meg látható formában;
- a diszkrét ábrázolás: a folytonos beszédjelet véges számú különálló szegmenssel ábrázolja; ezért szelektív és nem teljes reprezentációja a beszédnek;
- a véges számú jelkészlet: szimbólumok véges halmazát használja a végtelenül változatos beszédforma ábrázolására; ezért olyan modellje a beszédnek, amit nagyon sokféleképpen lehet kiolvasni;
- a konvenciók használata: a közösség által elfogadott szabályok és konvenciók szabályozzák a használatát, amelyek bizonyos önkényes módon meghatározzák az elemek jelentését, sorrendjét és egymáshoz való viszonyát;
- az időtállóság: egy többé-kevésbé tartós médiumon (papír, kőtábla) rögzíthetjük a beszédet, hogy később is értelmezni lehessen.

Az írás és a beszéd viszonya. Az írás és a beszéd egymással szoros kapcsolatban vannak, ezt a kapcsolatot a fonémákhoz való kötésük adja. Ugyanakkor szerkezetileg élesen különböznek egymástól. A kölcsönös viszonyukban vannak olyan összetevők, amelyek egymásnak pontosan megfeleltethetők, és vannak olyanok, amelyek csak az egyikre vagy csak a másikra jellemzőek. Az alábbiakban megnézünk néhány eltérést, mégpedig a beszéd különböző szintjeinek megjelenését az írásban, a beszédhez képest jelen lévő többletinformációt, az azonos alakú és azonos hangzású szavak kérdését és a két kifejezésmód közötti pragmatikai eltérést.

Mi hiányzik az írásból? A beszéd szupraszegmentális szintjének ábrázolása nagyrészt hiányzik az írásból (lásd a 6 fejezet). A beszéd paralingvisztikus összetevői, például a közlemény megfogalmazójának kilétére utaló jellemzők jelentős része sem érzékelhetők. A paralingvisztikus jegyeknek nincs nyelvi funkciója, de sokat elárulnak a beszélő fizikai és érzelmi alkatáról és állapotáról. Más módon természetesen az írás is tartalmaz információt az azt leíró személyről, például a kézírást több ezer éve használják a szerző személyének igazolására. (Pál apostol is erre az ismertetőjegyre hivatkozik az i. sz. 1. században annak igazolására, hogy nem más írt az ő nevében.)

Mi hiányzik a beszédből? A helyesírásnak megfelelő írott szöveg elárulhat olyan információt, ami beszédben nem kifejezhető. Például egy nem mondatkezdő szó nagy kezdőbetűje utalhat arra, hogy az tulajdonnév (a német nyelvben arra, hogy főnév), a csupa nagybetű arra, hogy rövidítéssel vagy betűszóval van dolgunk; speciális formátumuk kiemeli a szövegből a dátumokat, időpontokat. Az idegen szavakat, külföldi neveket eredeti kiejtésüket imitálva, de anyanyelvünk beszédhangjaival szoktuk kiejteni, ezért a beszéd elrejtheti nyelvi eredetüket, de írásban ez gyakran láthatóvá válik (például Kennedy elnök). Ez nehézséget is jelenthet a szöveget olvasó személy (vagy akár gép) számára, mivel neki kell megbirkóznia az idegen írásmód feloldásával.

Az írás és a beszéd kölcsönös megfeleltetésében további nehézséget jelentenek a homofónok és a homográfok. Homofónoknak nevezünk két vagy több eltérő jelentésű és írásmódú szót, amelyeknek azonos a kiejtése (*hej, hely; estéje, estélye;*

belefojt, belefolyt; csukja, csuklya). Homográfoknak azokat hívjuk, amelyeknek a leírása egyezik, a kiejtése eltér. Például az *egyek* szónak más az ejtése az *Egyek valamit?*, illetve az *Egyek vagyunk?* mondatokban. A magyar nyelvben a kiejtést nagyban követő írásmód miatt kevés példát találunk a homofónokra és a homográfokra. Kapcsolódó fogalom a homonima is, amely azonos kiejtésű és leírású szavak egyikét jelöli, amelyek jelentése között nincs kapcsolat; ilyenek például magyarul a *nyúl, ég, dob, csap, áll, fog, vár* szavak, amelyeknek van teljesen független főnévi és igei jelentése is. Ha a különböző jelentésű változatok között van etimológiai kapcsolat, azaz közös az eredetük, akkor többjelentésű szónak szoktuk nevezni (poliszémia). Ahogy látjuk, ezek a szavak egyszerre homográfok és homofónok.

Az írás grafémái és a beszéd hangjai közötti megfeleltetés nyelvenként eltérő lehet. Ez egyes nyelvekben történhet úgy, hogy egy szó írásképéhez megtanuljuk a kiejtését (ebben segíthet a hasonló írásmódú szavak közötti analógia felismerése), illetve úgy, hogy egyes betűcsoportokra vonatkozó szabályokat tanulunk meg (Field 2003, 28. o.). A magyar nyelv többnyire fonémikus írása miatt általában elégséges az utóbbi, méghozzá viszonylag egyszerű kiejtési szabályok használatával.

A graféma-beszédhang átírás fontos eleme a beszédtechnológiának is. Az írott szöveg kiejtés szerinti átírása az automatikus szövegfelolvasó rendszerekben (10.3.2. fejezet) jut fontos szerephez (Olaszy 1984, Olaszy et al. 2000a). Ugyanez fordítva is szükséges, a kiejtés írott szöveggé való átalakítására a gépi beszédfelismeréshez (9.5.1. fejezet) alakítanak ki modelleket, eszközöket, például automatikus beszédlejegyző rendszerekhez (Mihajlik et al. 2002).

Pragmatikai szempontból is teljesen más egy olyan szituáció, amelyben beszédre van szükség, mint amikor írásra (Tátrai 2009). A szöveg e két formájának alapvető prototípusa a társalgás és az irodalmi szöveg; közöttük számos átmeneti forma létezik: telefonbeszélgetés, személyes levél, újságcikk, műszaki leírás és a többi. Társalgás esetén a beszélők ugyanazt a környezetet érzékelik, és sok esetben egymást is szemtől szemben látják. A dialógus résztvevői többé-kevésbé tudatában vannak (vagy menet közben tudatába kerülnek), hogy a másik milyen ismeretekkel rendelkezik, így erre a tudásra építhetnek. Továbbá a verbális kommunikációval párhuzamosan nonverbális és paralingvisztikus kommunikációt is folytat(hat)nak. Így más jellegű verbális közlésre van szükségük. Ha látják egymást, nincs szükség a környezet leírására, sőt dolgok megnevezése helyett mutatással vagy pillantással is utalhatnak rájuk. Viszont a spontán szituáció miatt nincs idő előre megtervezni a mondandót, ezért megjelennek a gondolkodási folyamatot jelző néma és kitöltött szünetek (közbeszédben az ö-zés) és a diskurzusjelölők (*nos, értem*). Időnként hibázunk is a megfogalmazásban, vagy megbotlik a nyelvünk, ami miatt megjelenhetnek ismétlések, újrakezdések és ezek jelzésére közbeiktatott szavak (*úgy értem, akarom mondani*). Ezzel szemben az irodalmi szöveg szerzője nem ismeri azt, akihez beszél, más időben, más helyen van, mint aki olvassa az írását. Ezért le kell írnia a helyszíneket, körülményeket, és minden olyan ismeretet, amiről feltételezhető, hogy

a megcélzott olvasóközönség nem minden tagja ismeri. Ha érzelmeket, hangulatot, attitűdöket szeretne kifejezni, ezt is csak azok körülírásával teheti meg.

A szupraszegmentális szint egyik funkciója a tagolás, melynek jelölése sokáig hiányzott az írásból. Példa erre, hogy az indoeurópai nyelvek írásában egészen a 9. századig nem jelölték a szavak határát (Fischer 2004, 91. o.). Több nyelvben a mai napig sem jelölik a szóközöket (kínai, japán). Ma már számos nyelvben segédjelekkel tagolják a szöveget, jelölik a mondat- és tagmondathatárokat, és szóközök választják el a szavakat. A jelölésük könnyebbé teszi az olvasást, illetve a beszédhez képest csökkenti a kétértelműségek számát. A folyamatos beszédben nem tartunk szünetet a szavak között, ezért egy elhangzott hangsort sok esetben csak a környezete segítségével tudunk értelmezni. Például a *kicsikém*, illetve *kicsi kém* betűsorozatok felolvasva ugyanúgy hangzanak, lehet szó egy kedves kisbabáról, esetleg valakinek a kedveséről, de egy apró termetű titkos ügynökről is. Természetesen az ilyen kétértelműségek eldöntése általában nem okoz nehézséget számunkra, sőt észre sem vesszük a többértelműség lehetőségét, mivel a kontextusból egyértelművé válik számunkra a kimondott hangsor értelme. A gépi döntésnél viszont problémát okozhat.

Az írás és beszéd fent leírt kapcsolatrendszerét eltérő szinten, de valamennyien ismerjük, és többé-kevésbé tudatosan alkalmazzuk is, amikor írott szöveget felolvasunk, vagy amikor fogalmazunk és írunk. A nyelv- és beszédtechnológiában ezeket számítástechnikai eszközökkel kell megvalósítani. A gépi döntések hatásfoka még jócskán elmarad az emberéitől.

Összefoglalva tehát azt mondhatjuk, hogy az írott szöveg és annak kiejtett változata között az adott nyelvre vonatkozó szabályrendszer teremt kapcsolatot. Tulajdonképpen ennek alapvető szabályait tanulják a kisgyerekek is az olvasástanulás időszakában. A fő különbségek a tagolásban vannak. A magyar írás több szinten diszkrét: betűk, szóközök a szavak között, írásjelek a mondaton belül és a mondatok között. Egy szöveg felolvasásakor nem mindig igazodunk az írás diszkrét tagolási szintjeihez, más szempontok szerint szervezzük a beszédet.

4.2. Hangjelölések

Olaszy Gábor

Egy elmondott közlés kiejtést jelölő úgynevezett fonetikai átiratának elkészítéséhez hangszimbólumokat kell definiálni. A hangszimbólum a kiejtett beszédhangot jelöli valamilyen írott (karakteres) formában. Ha szövegből készítünk fonetikai átiratot, akkor a szöveg felolvasásakor valószínűsíthető hangsorozatot adjuk meg a fonetikai átiratban. A fonetikai átirat részletessége több szintű lehet, ez függ az átírás céljától és az átíró személytől (esetleg szoftvertől) is. Hangjelölésre azért van szükség, hogy írásos formában is érzékeltetni tudjuk (nyelvtől függetlenül), hogy milyen hangról

beszélünk. Az első hangjelölő rendszert 1889-ben a Nemzetközi Fonetikai Társaság (International Phonetic Association) hozta létre. Ez az IPA-szimbólumkészlet, ami mind fonémák, mind fonémavariánsok jelölésére alkalmas. Később alakították ki a számítógépeken is használható, karakterekkel kifejezett hangszimbólumokat. Ez a SAMPA-jelrendszer. A magyar beszédhangok SAMPA jelölési rendszerét is kidolgozták (Vicsi 1996).

Ebben a könyvben többféle hangjelölést (szimbólumot) alkalmazunk (4.1. táblázat), a beszédhang magyar betűkkel jelölt formáit, a hang nemzetközi jelét szögletes zárójelek között (IPA-szimbólum), a hang nemzetközi számítógépes jelét (SAMPA-jel). Használunk továbbá két egyedi hangjelölési formát is (E1, E2), amiket a magyar beszéd- és nyelvtechnológiai gyakorlati alkalmazásokhoz alakítottak ki. Ezek szűkebb skálában tartalmazzák a beszédhangokat, mint a szabványosított rendszerek, de a felhasználási célnak megfelelnek. A hangjelölés fontos a tudományos vizsgálatok pontos leírásánál, meghatározó szerepe van a beszédtechnológiában is (beszédszintézis, gépi beszédfelismerés, automatikus szókeresés beszédanyagokban, kiejtésikivétel-szótárakban stb.).

4.1. táblázat. A magyar beszédhangok jelölései. A variánsokat a hangszimbólum mellé tett csillaggal jelöltük

Betű	IPA-jel	SAMPA-jel	E1	E2	Betű	IPA-jel	SAMPA-jel	E1	E2
á	[aː]	A:	A:	a1	m,mm	[m], [mː]	m, m:	m, m:	m, m:
a	[ɔ]	O	a	a	m*	[ɱ]	M	-	-
o	[o]	o	o	o	n, nn	[n], [nː]	n, n:	n, n:	n, n:
ó	[oː]	o:	o:	o1	n*	[ŋ]	N	-	-
u	[u]	u	u	u	ny, nny	[ɲ], [ɲː]	J, J:	N, N:	ny, ny:
ú	[uː]	u:	u:	u1	j, ly, jj, lly	[j], [jː]	j, j:	j, j:	j, j:
ü	[y]	y	U	u2	j*	[ç]	X)	-	-
ű	[yː]	y:	U:	u3	h, hh	[h], [hː]	h, h:	h, h:	h
i	[i]	i	i	i	h*	[ɦ]	h	-	-
í	[iː]	i:	i:	i1	h*, ch*	[x]	x	-	-
é	[eː]	e:	E:	e1	h*	[ɣ]		-	-
ö	[ø]	2	O	o2	v, vv	[v], [vː]	v, v:	v, v:	v, v:
ő	[øː]	2:	O:	o3	f, ff	[f], [fː]	f, f:	f, f:	f, f:
e	[ɛ]	E	e	e	z, zz	[z], [zː]	z, z:	z, z:	z, z:
- *(svá)*	[ə]				sz, ssz	[s], [sː]	s, s:	s, s:	s, s:
b, bb	[b], [bː]	b, b:	b, b:	b	zs, zzs	[ʒ], [ʒː]	Z, Z:	Z, Z:	Z, Z:
p, pp	[p], [pː]	p, p:	p, p:	p	s, ss	[ʃ], [ʃː]	S, S:	S, S:	S, S:
d, dd	[d], [dː]	d, d:	d, d:	d	dz	[d͡z], [d͡zː]	dz, dz:	dz, dz	dz
t, tt	[t], [tː]	t, t:	t, t:	t, t:	dzs	[d͡ʒ], [d͡ʒː]	dZ, dZ:	dZ, dZ	dZ
gy, ggy	[ɟ], [ɟː]	d', d':	G, G:	gy, gy:	c, cc	[t͡s], [t͡sː]	ts, ts:	c, c:	c, c:
ty, tty	[c], [cː]	t', t':	T, T:	ty, ty:	cs, ccs	[t͡ʃ], [t͡ʃː]	tS, tS	C, C:	C, C:
g, gg	[g], [gː]	g, g:	g, g:	g, g:	l, ll	[l], [lː]	l, l:	l, l:	l, l:
k,kk	[k], [kː]	k, k:	k, k:	k, k:	r, rr	[r], [rː]	r, r:	r, r:	r, r:

A hangjelölés fontos része a fonológiai hanghosszúság érzékeltetése is (*tör, tőr; sok, sokk*). A hosszú beszédhangok jelölésére az IPA-rendszerben speciális jelölést alkalmaznak, a kettősponthoz hasonló, két egymás felett elhelyezett, csúcsukkal

szembe néző egyenlő oldalú háromszöget (*hajókkal* = [hɔjoːkːɔl]). A SAMPA-, az E1- és E2-rendszerekben kettősponttal jelölik a hosszú beszédhangokat. Az E1- és E2-jelű hangjelöléseket egyedi kutatásokhoz alakították ki, főleg a számítógépes feldolgozást tartva szem előtt. Az E1-rendszer (Olaszy et al. 2000b) hangjelei és a tényleges betűképek között kicsi az eltérés, a kapcsolat egyértelműnek tekinthető, tehát az ilyen jelekkel leírt hangsor viszonylag jól olvasható. Az E2-hangjelek (Prószéky 1985) kissé eltérnek a betűképektől, főleg a magánhangzók esetében, ugyanis betű- és számkombinációt is felhasználnak bizonyos hangok jelölésére. A betűjel utáni sorszám határozza meg, hogy valójában milyen hangról van szó (*hálókkal* = h; a1; l; o1; k:; a; l). Ezt a jelölési rendszert belső gépi reprezentációkban alkalmazzák. A szakirodalomban használják még a magánhangzók általános jelölésére a V jelet (Vowel), a mássalhangzók jelölésére pedig a C jelet (Consonant). A V, C jelek kombinációi általános hangkapcsolatokat jelentenek. Például a CV jel mássalhangzó-magánhangzó kapcsolatot jelöl.

A hangszimbólumok alkalmazási rendje ebben a könyvben a következő: A szövegben általában az IPA-szimbólumokkal fejezzük ki a kérdéses beszédhangot, hangsort (szögletes zárójelek között). A betűjeleket ábrákon és táblázatokban is használjuk, valamint a függelékekben. A számítógéppel generált ábrákban, táblázatokban E1- és E2-jelű hangjelöléseket is használjuk.

4.3. Tagolási különbözőségek

Olaszy Gábor

A beszédtechnológiában fontos szerepet tölt be a szöveges forma, valamint a hozzá tartozó beszédjel (a beszédet a gép valamilyen írott formából állítja elő, illetve elhangzott beszédből készít írott formát). Ezért fontos a tagolás vizsgálata mindkét szinten. A tagolás más képet mutat a beszédhullám tekintetében és mást az annak megfelelő ortografikus szövegnél. A beszédben sokkal gazdagabb eszközrendszer áll a beszélő rendelkezésére, mint az írás szintjén. A beszélő tudatos cselekvéssel szinte bármilyen hozzáadott információt is be tud építeni a beszédjébe a tényleges szövegtartalmon felül. Ilyenek például az érzelem (öröm, gúny, szomorúság, türelmetlenség, harag), a hangerő, a beszédritmus változtatása, a hanglejtés, a hangsúlyozás, a magyarázó beszéd stb.

Egyelőre maradjunk csak az alapvető kérdéskörnél, vagyis az írott szövegnél és annak felolvasott változatánál (például hírolvasás, időjárás-jelentés, regény felolvasása, írott üzenetek hangos közlése). A kérdés az, hogy mennyiben lehet egyforma a tagolás a két szinten, és mely pontokon térhet el? A magyar írás diszkrét elemek sorozatára épül, ezek a betűk. A helyesírás szabályai szigorúan megszabják, hogy mit hogyan írjunk. A szavak szóközzel különülnek el egymástól. Tovább haladva

felfelé a hierarchiában, a szavak sorozatát csoportokba szedhetjük a központozási jelek használatával (vessző, kettőspont, gondolatjel, zárójel stb.), azonban ezek helyét is többnyire szigorú szabályok határozzák meg. A mondatokat a mondatvégi írásjelek választják el egymástól. A több mondatból álló, nagyobb szövegrészek pedig bekezdéseket alkotnak. A felolvasott szöveg hangzó alakjának megformálásakor más szabályok érvényesülnek, és ezek sokkal lazábbak, mint amilyeneket az írásnál láttunk. A beszélőt csak a nyelvi norma köti a szövegfelolvasásban, azon belül szabadon dönthet. Az íráshoz kötődő szigorú szabályok csak arra vonatkoznak, hogy az artikulációt alapvetően az írás vezérli. Ugyanakkor a felolvasott szöveg (pillanatnyi) értelmezése határozza meg például a hangsúlyozást, a dallammenetet, az esetleges megállásokat a kiejtésben. A korábban feltett kérdésre tehát a válasz az, hogy a beszédben másfajta tagolási elveket kell követni, mint az írásnál. Olvasáskor (és spontán beszédben is) a kiejtett hangok és a szavak folyamatosan kapcsolódnak egymáshoz. Rövid megszakításra csak akkor kerül sor, ha az értelmezés megkívánja. Ezt általában a központozási jelek jelölik az írásban, de sok esetben nincsenek ilyen jelölések, mégis szünetet tartunk. Ez utóbbi eseteket a szövegösszefüggések határozzák meg. Egy példát mutatunk be arra, hogy ugyanazt a szöveget kétféle tagolással is fel lehet olvasni. A kétfajta beszédet az alsó sorokkal imitáljuk, az egybefüggően kimondott részeket számokkal jelöljük.

A szöveg felolvasásakor más szabályok érvényesülnek, és ezek a következők.

(1)Aszövegfelolvasásakor (2)másszabályokérvényesülnek (3)ésezekakövetkezők.

(1)Aszövegfelolvasásakormásszabályokérvényesülnek (2)ésezekakövetkezők.

Előfordulhatnak olyan esetek is, amikor az írásban központozási jel van, a beszédben mégsem tart szünetet a beszélő. Ilyenre is mutatunk egy példát.

A közelítés ugyanazt az eredményt kell, hogy adja.

(1)Aközelítés (2)ugyanaztazeredménytkellhogyadja.

A fentiek alapján a tagolással kapcsolatos fő tudnivalók a következők.

Az első és talán legfontosabb, hogy amit az írásban tagolásként találunk, ugyanazt, ugyanott általában nem találjuk meg a beszédjelben. A beszédjelben tehát más helyen találhatunk szüneteket, mint ahogy a szövegből következne. A szüneteket a szöveg értelmezéséből alakítja ki a beszélő, csak néha egyeznek meg a szünetek az ortografikus szöveg központozási jeleivel.

A második, hogy a beszédjelben a beszédhangok nem különíthetők el egymástól, tehát a hanghatárok megjelölése a hullámformában sok esetben nem egyértelmű. Ugyanez vonatkozik a szavak kapcsolódására is.

A harmadik, hogy a beszédben olyan akusztikai elemek is előfordulnak, amelyek nem tartoznak a beszédhangsorozathoz (áttételesen a szöveghez). Ezek, ha az íráshoz akarjuk hozzáigazítani a beszédet, megzavarhatják a feldolgozást. Ilyen elemek a hangos szünetek, a tévesztések és javításaik, a zárhangok végének túlartikulálása, megnyomása a hangsor végén (svá), glottalizált hangindítások, nyelvcsettintések, a

szó újraindítása stb. Ezek elkülönítése a beszédjeltől bizonyos feldolgozásoknál elengedhetetlenül szükséges, de nagyon nehéz.

Összefoglalva azt mondhatjuk, hogy egyértelmű tagolást az írott szövegből nem lehet levezetni a beszéd szintjére, és a beszédből sem az írott szövegére.

4.4. Az írott szöveg és a hangalak kapcsolata

Olaszy Gábor

A magyar nyelvre még nincs egyértelmű és mindenre kiterjedő nyilvános számítógépes fonetikai hangátíró szoftver egyik irányban sem (szövegből hangalak, illetve beszédből hangalak, majd írott szöveg). A beszédkutató műhelyekben születtek már eredmények, de azok minden esetben igazodtak a saját feldolgozási követelményeikhez, ezért nem tekinthetők általánosnak (nem is nyilvánosak). Részletesen a későbbi fejezetekben mutatunk be ilyen rendszereket.

A hangalak, azaz a fonetikai átirat meghatározása két irányból közelíthető, az ortografikus szövegből kiindulva, illetve az elhangzó beszédből visszafejtve. A gondolati tartalom e kétfajta reprezentációját tehát a fonetikai átirat kapcsolja össze. Elméletileg a két irányból való közelítés ugyanazt az eredményt kell, hogy adja.

Szövegből kiinduló példaként említhetők a nyomtatott kiejtési szótárak, amelyek azt adják meg, hogy egy szónak mi a kiejtett alakja (Fekete 1992). Fekete szótárában 10 800 szó kiejtési formája található meg. Fontos az is a kiejtés szempontjából, hogy az idegen írásmódú szavaknak, kifejezéseknek mi a kiejtése (Tóthfalusi 2006). Ez utóbbi szótárban 40 000 elem van. Elektronikus formában a teljes magyar nyelvet reprezentáló kiejtési adattár is elkészült 2010-re (8.4.3. fejezet). Ez 1,5 millió szóalak kiejtési formáját adja meg hangszimbólumokkal (a szóalak terminus itt azt fejez ki, hogy a tőszó toldalékokkal ellátott alakja is különálló lexikai elem a szótárban). Egy hangos közlés kiejtés szerinti automatikus lejegyzése sokkal komplexebb feladat, mint egyes szavak kiejtésének megadása.

Amennyiben folyamatos beszédet akarunk lejegyezni, akkor a teljes közlési folyamatot kell megfejteni, figyelemmel kell lenni a szavak összekapcsolásakor létrejövő koartikulációs jelenségekre, továbbá a szünettartásra is. A beszédből történő visszafejtésnél további hangjelenségek is előfordulhatnak (hangos szünet, dupla hangindítás, krákogás, hümmögés, hangos levegővétel stb.), amelyek nehezíthetik a feldolgozást.

A fonetikai lejegyzés nyelvfüggő. A magyar a fonematikus helyesírású nyelvek közé tartozik, ami azt jelenti, hogy a leírt ortografikus szöveg közel áll a hangalakhoz, nem kell túl bonyolult szabályrendszer a kiejtés szerinti átíráshoz. A német például kissé bonyolultabb, az angol viszont az egyik legbonyolultabb ilyen szempontból. A hangalak megadása, vagyis a fonetikai átirat elkészítése fontos mind a gépi

beszédfelismerés, mind pedig a gépi beszéd-előállítás szempontjából, de fontos lehet nyelvészeti-fonetikai kutatásokban is. A következőkben a beszédtechnológia vonatkozásában vizsgáljuk a témakört. Használjuk a fonemikus szöveg fogalmát mint a hangátírás köztes lépcsőjét. Mit tekintünk fonemikus szövegformának? Ez olyan szövegformát jelent, amelyben a fonémákat a betűjelükkel (vagy szabványos fonetikai hangszimbólumokkal) jelöljük, a szóhatárok megmaradnak, csak minden olyan elem szövegszintű feloldásra kerül, amelyik az ortografikus betűképe alapján nem felel meg a kiejtés hangalakjának (például számok, mértékegységek, nevek, rövidítések stb.). Ezt a műveletet kanonikus átírásnak is nevezik (lásd a 8. fejezetet). A fonemikus átírás folyamatát szövegnormalizálásnak is szokták hívni, mivel a helyesírás szerinti ortografikus formát egy egységesebb, csak írott szavakat tartalmazó formára hozzák. A fonemikus szövegformát automatikusan állítják elő az ortografikus szövegből. A fonemikus szövegforma információt ad például arról, hogy közelítőleg hány beszédhangot tartalmaz az adott közlés. A fonemikus szöveg általában fizikailag hosszabb, mint az ortografikus forma (több betű lesz benne a feloldások miatt). A koartikulációból adódó hangváltozásokat a fonemikus szöveg még nem tartalmazza. Az átiratban még tényleges szavak szerepelnek, amelyek között szóközök vannak. Ezért ezt az átiratot jól lehet alkalmazni a szóhatárok közelítő jelzésére is a beszédjelben.

Tekintsük át, hogy milyen módon jutunk el a hangalakhoz például az ortografikus szövegből. Egy lehetséges megoldás, hogy először az ortografikus formából készítünk fonemikus szintű szöveget, majd ebből készítjük el a fonetikai átiratot, a hangalakot reprezentáló karaktersorozatot, ami a hullámforma szintjére vetíthetően jelöli a tényleges beszédhangokat hangszimbólumok formájában. A fonetikai átirat már tartalmazza a koartikulációból eredő hangváltozásokat (hasonulások, hangkiesések, hangbetoldások stb.), és szerepelnek benne a beszélő személy által tartott szünetek is. A fonetikai átirat tehát már pontosan megmutatja, hogy hány és milyen beszédhangot tartalmaz az adott közlés, azaz pontos leképezését adja a hullámformának (a hangidőtartamok és egyéb időadatok megadása nélkül). Ezt nevezik fonotipikus átírásnak is (lásd a 8. fejezetet).

Az alábbi példában mindhárom szintet megmutatjuk egy példamondaton. A (sil) jelzés a szünettartást jelenti.

Ortografikus alak: *A hőmérséklet –2 fok, a többit SMS-ben írom meg.*

Fonemikus alak: A hőmérséklet mínusz kettő fok, a többit esemesben írom meg.

Fonetikai átirat az E2-hangjelekkel: A h o3 m e1 r s e1 k l e t m i1 n u sz k e t: o3 f o k (sil) a t o2 b: i t e s e m e zs b e n i1 r o m: e g

Miért van szükség a köztes, fonemikus szövegformára? Elsősorban azért, mert segíti a beszédjelen elvégzendő automatizált hanghatárjelölés elvégzését. Másodsorban pedig azért, mert ez a forma jól tükrözi a hullámforma hangjainak számát, ugyanakkor még ember számára olvasható, tehát manuálisan is lehet ellenőrizni a megfeleltetést például a hullámformával. A felolvasásból készített nagy méretű beszédadat-

bázisok első feldolgozási lépése minden esetben az, hogy az elhangzott beszédet egy erre betanított személy összeveti azzal az írott anyaggal, amit a bemondó felolvasott. Az egyezést a fonemikus átiraton ellenőrzi. A bemondó ugyanis téveszthet (más szót is mondhat, hozzátehet egy toldalékot egy szóhoz), ami nem változtatja meg a mondat értelmét, viszont a hangzás és a neki megfelelő szövegrész nem lesz szinkronban. Ilyen esetekben mindig a hanghoz igazítják a szöveget, egyszerűen kijavítják a fonemikus átiratot és az ortografikus szöveget is. A feldolgozás eme szakaszát jelenleg csak ember tudja elvégezni. A rövidítések feloldását már segítik szoftverek, ami azt jelenti, hogy nem kell az embernek begépelni a rövidítésnek megfelelő betűsort, hanem azt automatikusan elvégzi a gép. Az ilyen szigorú megfeleltetés alkalmazása mind a beszédszintézisben, mind pedig a gépi beszédfelismerésben fontos. A gyakorlatban a beszédfelismerők betanításához elengedhetetlen az ilyen hármas ábrázolási forma, főleg a tanítási szakaszhoz, hiszen a tanító algoritmus a fonemikus átiratból tudja meg, hogy mit kell kinyernie a beszédjelből és megtanulnia. A beszédhangok határait kijelölő, úgynevezett kényszerített felismeréssel dolgozó algoritmusok csak a fonemikus átiratra támaszkodva tudnak eligazodni a beszédhullámban (9.5.3.2. fejezet). Az ilyen módszerrel készített, felcímkézett beszédadatbázisok sokféle egyéb kutatásnál, fejlesztésnél is jól használhatók.

4.5. Hang- és szóhatárok kijelölése a beszéd hullámformáján

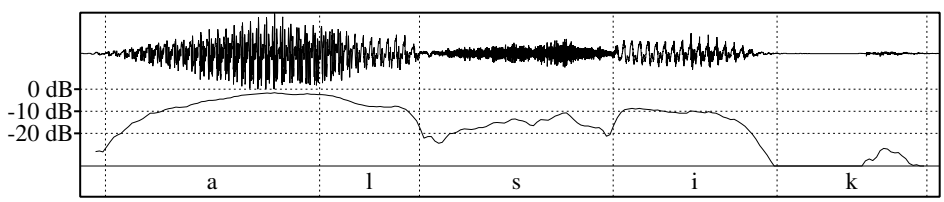
Olaszy Gábor

A cél, hogy a beszéd hullámformájában tudjuk követni a hangok egymásutánját, hogy mettől meddig tartanak a beszédhangok. A hanghatár kijelölése kézi és gépi módszerekkel lehetséges. Ma már a kézi jelölést is támogatják célszoftverek. A manuális kijelölés előnye, hogy a szakértő nagy pontossággal meg tudja ítélni a hanghatár helyét. Az ilyen munkát általában audiovizuális támogatással végzik. Ekkor a jelölő személy látja a hullámformát, a spektrumot, az intenzitás alakulását, és mindezekhez még meg is tudja hallgatni a kijelölt hangrészt, akár zöngeperiódusonként szűkítve, bővítve a kijelölést. A meghallgatáshoz célszerű biztosítani egy úgynevezett lépésenként tágítható-zsugorítható (periódusszintű időablakos) kapuzásos megszólaltatási lehetőséget is. Ezzel akár balról jobbra, vagy fordítva fokozatosan hallgathatjuk a hullámforma hangját, és percepciósan is meghatározhatjuk a hangok csatlakozási pontját. Ilyen meghallgatási lehetőséget biztosít a ProfiVox beszédszintetizáló rendszer vizsgálati szoftvere, a Profidev (Ferenczy et al. 1997, Olaszy 2001a). A hanghatárok megállapításában segíthet a hangspektrogram képének tanulmányozása is. Régen (a számítógépes korszak előtti időkben) ilyen módszerrel végeztek kézi méréseket a magyar beszédhangok időtartamainak a meghatározásához (Mag-

dics 1966). A hangspektrogram regisztrátumán vonalzóval jelölték be a hanghatárt, majd az időtengely osztásai alapján kiszámították a hangidőtartamokat.

A kézi hanghatárjelölés hátránya (még a számítógépes támogatás ellenére is), hogy lassú és szubjektív. Ez utóbbi azt jelenti, hogy a megjelölés helye szubjektív döntés eredménye (nem biztos, hogy más személy ugyanazt a döntést hozná). A gépi hanghatárjelölés pontossága kisebb, mint a manuális jelölésé, viszont gyors, ezért ideális nagy tömegű adat feldolgozásához. A gépi hanghatárjelölés pontosságát javítani lehet utófeldolgozással, amikor félautomatikus módszerekkel megkeressük és kijavítjuk a gépi eljárás hibáit (Olaszy–Bartalis 2008). Számos akusztikai tényező befolyásolja, hogy a hangok határát milyen pontossággal tudjuk kijelölni. A vizsgálat eredménye szempontjából három hangkapcsolódási helyzetet kell megkülönböztetnünk: azokat, amelyeknél a hanghatárt a rezgésképből szinte egyértelműen ki lehet jelölni (pontos lesz az eredmény); azokat, amelyeknél ez nehezebb, itt bonyolultabb vizsgálatokat kell végezni; és végül azokat, amelyeknél nincs egyértelmű hanghatár a két hang között, a határpont helye minden esetben jórészt a mérést végző személy (vagy szoftver) döntésétől függ.

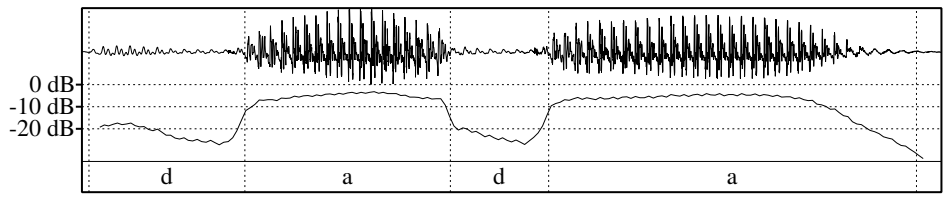
Hanghatár-kijelölés a gerjesztés alapján. A hanghatár viszonylag pontosan kijelölhető a hangsor azon pontjain, ahol az egyes hangokat alapvetően a gerjesztési jel megváltozása különíti el. Ilyenkor a zöngés hangperiódus indulási pontjára kell tenni a jelzést. A zöngés-zöngétlen váltás mellett ebben a helyzetben jellemzően intenzitásminimum is keletkezik, tehát két információ is segíti a hanghatár megtalálását. Ilyenkor akár gépileg, akár vizuális ítélet alapján meghatározható a két hang határpontja (***alszik, átnéz***).



4.1. ábra. Az *alszik* (753 ms) szó hanghatárainak kijelölése a gerjesztésváltás helyének detektálásával az időfüggvényen (fent). Jól látható, hogy a zöngés hangok egymás utáni periódusokból állnak, a zörej pedig aperiodikus

Hanghatár-kijelölés az intenzitásváltozás alapján. Azoknál a hangkapcsolatoknál, amelyeknél nincs gerjesztésváltás a két hang kapcsolódási pontján, a hanghatár kijelölése már nehezebb. A meghatározott feltételek között lezajló intenzitásváltozás jelzésként szolgálhat a hanghatár megállapításához. Ha jelentős intenzitásváltozás van 20–30 ms-on belül, akkor ott valószínűleg hanghatár van. A hanghatárt itt is a kiválasztott hangperiódus indulási pontjára kell elhelyezni. Meghatározott hang-

kapcsolatok rendelhetők ehhez a vizsgálati csoporthoz, nevezetesen amikor nagy energiájú zöngés hangok (magánhangzók, a [j] és [l]) kapcsolódnak zöngés zár-, zár-rés hangokhoz (4.2. ábra). A hanghatár megállapítását a hangspektrogramon mérhető jellegzetes változások segíthetik (formáns gyengülése, eltűnése).



4.2. ábra. A *dada* (621 ms) szó hanghatárainak kijelölése az időfüggvény (fent) és az intenzitásváltozás (lent) együttes vizsgálatával. A hangintenzitás időfüggvénye egyértelműen jelzi a hanghatárhoz kapcsolható időintervallumot. A magánhangzó periódusainak amplitúdója lényegesen nagyobb, mint a zöngés zárhangé

Nincs egyértelmű hanghatár. Azoknál a hangkapcsolatoknál, ahol nincs gerjesztésváltás, továbbá az artikulációs mozgások időben csak viszonylag lassan változnak, és a hangintenzitásban sincs lényeges módosulás, a hanghatár megállapítása csak közelítőleg lehetséges még hangspektrogram támogatásával is. Ilyenek például azok a helyzetek, amikor két azonos magánhangzó találkozik (*ráálltak*), különböző magánhangzók kapcsolódnak egymáshoz (*ráadás*), a [j]+V (*kijárat*) és [l]+V (*elad*), egyes CC-kapcsolatok (*romnak*). Ezekben az esetekben a kutató (gépi algoritmus) döntésétől függ, hogy melyik pontot (hangperiódus kezdetét) nevezi ki hanghatárnak. A feldolgozó szoftverek megtervezésénél külön kell figyelni ezekre a tényezőkre. Ebből következik, hogy az ilyen hangkapcsolatoknál megállapított hangidőtartam-adatok nagyobb szórást mutathathatnak, mint a korábbi kategóriákéi.

Szóhatárok kijelölése a hullámformán. A szóhatárok kijelölésénél is felmerülhetnek egyértelműsítendő kérdések. Ezek visszavezethetők a hanghatároknál felsoroltakra. Ha a két szó határán ugyanaz a hang van, akkor a hullámformában nem egyértelmű a két szó összekapcsolódási pontja (*nagyobb betűvel, fa alatt, Kis Sándor* stb.). Vegyük a *nagyobb betűvel* szókapcsolatot példaként. A zöngés zárhangok találkozásából egy hosszú zárhang [b:] keletkezik a szóhatáron. Ez a zárhang mindkét szóhoz tatozik. Szétválasztani akusztikai szempontból nem lehet. Ha ennek a hangnak a jobb oldali határát az első szó végének jelöljük, akkor az első szó szempontjából helyesen járunk el, a második szónak viszont nem lesz jelölt kezdő zárhangja. Egy kezelhető megoldás a következő: alkalmazzunk külön jelzést a szó kezdetére és befejezésére. Jelölje a szókezdetet a <, a befejezést a > jel (Fék et al. 2006). Ekkor a fonetikus átiratban a következő jelölések lesznek (az E2 hangjeleket használva): <n a gy o <b:> e t u3 v e l>. A jelölés mutatja, hogy a [b:] befejező hang is és kezdő is egyaránt. Az ilyen jelzéskörnyezettel

rendelkező hangokat a későbbi feldolgozó algoritmus külön tudja majd kezelni, amikor beszédből kell szöveget előállítania.

4.6. Magyar hang-, betű- és szóstatisztika

Zainkó Csaba

A nyelv írásos formája (betűkép) és a hangalak (kiejtés) szoros összefüggésben van egymással. A számítógépes nyelv- és beszédfeldolgozás felszínre hozta azt az igényt, hogy a statisztikai elemzéseknél vegyük figyelembe a két szint egymásra hatását is, hiszen egymásból következnek. Ez újfajta megközelítést igényel, olyat, amely alapjaiban kapcsolódik a szóstatisztikai adatokhoz, továbbá a szavakat felépítő betűk statisztikai feldolgozásához, valamint ahhoz, hogy a betűkép milyen hangszintű információkat tartalmaz. Teljes képet a nyelv statisztikai jellemzőiről csak akkor kaphatunk, ha mind a szövegszintű elemek, mind az elhangzó hangok szintjén ugyanarra a nagyméretű nyelvi anyagra végzünk méréseket. Az ilyen összefüggések megállapítására végzett kutatások eredményeit ismertetjük a következőkben.

4.6.1. Betűstatisztika a hangalak figyelembevételével

Az első átfogó fonémastatisztikát Szende (1976) kézi méréseiből ismerjük. A szerző 80 000 fonémát felölelő beszédanyagot dolgozott fel. Magyar betűstatisztikáról nem találtunk szakirodalmi adatot. A betű és hang mint két önálló szint egymásra hatását is feldolgozó mérést Zainkó (2009) végezte el. Az osztályozáskor figyelembe vette a betűsorozathoz rendelhető hangalaki reprezentációt is. Kutatásában a vizsgálat nyelvi anyaga a Magyar Nemzeti Szövegtár 2005-ös verziójának teljes szöveganyaga volt. Ez a szövegtár 187,6 millió szövegszót tartalmaz, 5 nagyobb szövegtípust dolgoz fel: sajtószövegeket, szépirodalmi műveket, tudományos, hivatalos és személyes szövegeket.

A módszer ismertetése. A mérésekhez gépi gyűjtő és szortírozó algoritmusok készültek, kifejezetten ehhez a kutatáshoz. A betű fogalmát kiterjesztettük és a célkitűzéshez alakítottuk. A karakter- és betűstatisztikához a vizsgált leghosszabb betűsorozat a szó volt, a nem betű típusú karaktereket (számok, relációs jelek stb.) figyelmen kívül hagytuk. A hangstatisztikát mondatokra vonatkoztatott egységekből készítettük, hogy figyelemmel legyünk a szóhatárokon a folyamatos ejtésből eredő hangváltozásokra is. A betű fogalmának kiterjesztése azt jelentette, hogy a beszédhang oldaláról is visszavetítettünk elemeket az írás szintjére. Például a *pech* szó klasszikus értelemben vett *ch* betűkapcsolatát az *sz* betűhöz hasonlóan kezeljük, két karakterből álló betűnek tekintjük, ugyanakkor eltérően kezeljük például a *lánchíd ch* betűkapcsolatá-

tól, ahol külön *c* és *h* betű szerepel. Az új betűszintű osztályozás miatt megmaradnak olyan információk is, amelyek a fonetikus átírás közben elvesznek. Például különkülön rendelkezésre áll az új típusú betűstatisztikában a [j] hangként kimondott *j* és *ly* betű, vagy az [i] hangként kimondott *i* és történelmi nevek végén gyakran szereplő *y* betű.

A betűstatisztika készítésekor betűként a következő tulajdonságú karaktereket vagy karaktersorozatokat értjük:

- a 44 betűs magyar ábécé tagjai: *a, á, b, c, cs, . . . , s, sz,. . . , q, w, x, y, z, zs*
- régi magyar családnevekben, idegen szavakban gyakran előforduló betűkombinációk: *cz, ch*, amelyeknek többféle ejtése is lehet (*technika, charter*).

Néhány ilyen betű jelölését ki kellett bővíteni, hogy érzékeltessük a belőle keletkezett hangot. Ezt a jelölést betű-hangjelnek nevezzük.

A *ch* betűkapcsolat esetében háromféle hangot vizsgáltunk ([x], [t͡ʃ] és [k]). Ennek megfelelően ennyi különböző jelölést alkalmazunk: ch_x, ch_cs, ch_k (ezeknél a jelöléseknél az aláhúzás utáni betű jelöli a kiejtési formát)

A *h* betű hangalakja is többféle lehet. Alapesetben zöngétlen glottális [x] hang (*hó*). Néma lehet szó végén (*cseh* [t͡ʃ ɛ]). Jelölése: $h_{néma}$.

A zöngésen ejtett forma a [ɦ] hang, ami esetenként intervokális helyzetben fordul elő. Jelölése: $h_{zöngés}$. Szó végén és mássalhangzókapcsolat első tagjaként lehet veláris zöngétlen réshang [x] (*sah, sahnak*).

A *j* betűt egyes esetekben zöngétlen [ç] réshangként ejtjük. Jelölése: $j_{zöngétlen}$.

Az *sch* német eredetű betű is gyakran előfordul, a 3 karakteres hossza miatt fontos a külön kezelése. Jelölése: sch.

Az *y* betű többnyire régi nevekben és idegen eredetű szavakban fordul elő általában [i] vagy [j] hangként valósul meg ejtéskor. Jelölésük: y_i, illetve y_j. Nem vizsgáltuk azokat az eseteket, amikor az *y* betű [ji] formában valósul meg, mint például a *Fáy* szóban.

Fontos megjegyezni, hogy ezek az osztályozások elsősorban beszédtechnológiai szempontok figyelembevételével történnek, nyelvészeti vonatkozásban bizonyos döntések hiányosnak tűnhetnek. A hangjelölések megállapítására és osztályozására az elválasztási szabályokra épített algoritmust használtuk (*Ri-chárd*, *Mün-chen*, *Ben-czúr)*, miszerint ezek a betűk nem elválaszthatóak. A döntéseket a magyar elválasztásiminta-gyűjtemény szószedete (Nagy 2008) alapján hoztuk meg. Az algoritmusunk figyelembe veszi a két karakterből álló betűk mellett az ábécé (*gy, ty, ny, sz, zs, cs* betűit is, azok hosszú változatával egyetemben. A hosszú változatokat két betűnek tekintettük (*zzs* = *zs* + *zs*) a statisztikai feldolgozás során.

A szövegekből előállítottuk a hangalakot (hangszimbólumok írott sorozatát) beszédtechnológiai gépi módszerek alkalmazásával. Három eszközt használtunk, egy szabályalapú algoritmust: a ProfiVox szövegfelolvasó rendszer fonetikai átíróját és szabálygyűjteményét (Olaszy et al. 2000a), a magyar elektronikus kiejtési szótárt

(Abari–Olaszy 2006) és a Névmondó tulajdonnév-kiejtési gyűjteményt (Németh et al. 2003). A kiejtési forma meghatározásához kialakított szabályok a magyar nyelvi normát képviselik. A hangstatisztika teljesen gépi módszerrel készült, nincsenek megkülönböztetve a fonémarealizációktól a variánsok. Manuális ellenőrzés nem történt a fonetikus átiratokon. A hangstatisztikát a szöveg fonetikus alakjából állítottuk elő.

A módszer előnyei. Nagyméretű szövegkorpuszon alapul, ezért statisztikailag megbízható eredményeket ad. Megismételhető, mivel számítógépes támogatással készült, az algoritmusok többször is futtathatók. Újszerű tudományos vizsgálatok is végezhetők. Összekapcsolható a beszédhang- és a karakterreprezentáció.

A módszer hátrányai. A gépileg gyűjtött és ellenőrzött szöveg tartalmaz(hat) hibákat. A feldolgozott szövegkorpusz nagy mérete miatt manuális ellenőrzés nem jöhet szóba. A felhasznált kiejtésikivétel-szótárak szintén részben gépi módszereken alapulnak, ezért tartalmazhatnak hibákat, vagy hiányosak is lehetnek. A vizsgált betűk meghatározása önkényes, a gépi beszédfeldolgozás egyes szempontjait tartotta szem előtt, más felhasználás esetén a vizsgált betűk kiválasztása korlátozást jelenthet. Például a régi írásmódú betűk vizsgálata nem teljes körű, amely a beszéd szempontjából megengedhető, de névelemzés esetén már esetleg nem.

Eredmények. A vizsgált szöveg és hangalak statisztikai elemzése 3 formában készült el. Az eredmények a 4.2. táblázat oszlopaiban láthatók. Az első két oszlop a karakterstatisztika, a második kettő a betűstatisztika, az utolsó két oszlop a hangstatisztika eredményeit mutatja. A táblázatban szereplő számértékek megadják, hogy átlagosan 1000 elemből hány adott elem fordul elő. Az üres mezők azt jelentik, hogy az adott típusú statisztikában olyan elem nem szerepelt.

A speciális betűk statisztikáját a 4.3 táblázatban adjuk meg, ebben a számértékek 1 millió elemre vonatkoznak. Az összes vizsgált betűre vonatkozó gyakorisági sorrendet a 4.4. táblázat mutatja.

A különböző statisztikák elkészítése nagyságrendileg eltérő erőforrást igényelt. A karakterstatisztika másodpercek alatt elkészült, a betűstatisztika több tíz perc, míg a hangstatisztika elkészítése 3–4 órát vett igénybe. A karakterstatisztika használata tehát akkor előnyös, ha gyors működés elengedhetetlen.

A karakterstatisztika csak 36 karakterre tartalmaz információkat. Ahol ugyanazon karakter több betűreprezentációban is előfordulhat, ezt figyelembe kell venni a számadatok értelmezésénél.

Például a 4.2. táblázat *s* karakteréhez és betűjéhez tartozó gyakoriságokat összevetve látható, hogy az *s* karakter jóval gyakrabban fordul elő, mint az *s* betű. Ennek oka a kettős betűk szétdarabolása. Betűstatisztika helyett ezért csak fenntartásokkal használható. Ennek ellenére az egyszerű programozhatóság miatt sok helyen így használják.

A megadott hangstatisztika is tartalmaz egyszerűsítéseket, csak 39 beszédhang szerepel benne (nem kezeli külön a hosszú-rövid mássalhangókat, mivel közöttük

4.2. táblázat. Magyar karakter-, betű- és hangstatisztika

Karakter	1000-ből	Betű	1000-ből	Hang	1000-ből	Karakter	1000-ből	Betű	1000-ből	Hang	1000-ből
a	89,37	a	92,85	[ɔ]	90,21	o	40,93	o	40,21	[o]	42,26
á	35,95	á	37,58	[aː]	37,99	ó	10,03	ó	10,49	[oː]	9,95
b	19,66	b	20,56	[b]	18,28	ö	10,90	ö	11,39	[ø]	11,75
c	7,64	c	3,97	[t͡s]	6,10	ő	8,94	ő	9,35	[øː]	9,68
		cs	3,91	[t͡ʃ]	3,85	p	11,14	p	11,65	[p]	12,42
d	19,74	d	20,42	[d]	19,49	q	0,04	q	0,04		
		dz	0,03	[d͡z]	-,-	r	42,47	r	44,41	[r]	44,02
		dzs	0,02	[d͡ʒ]	-,-	s	60,35	s	39,08	[ʃ]	35,89
e	98,70	e	101,31	[ɛ]	106,59			sz	19,27	[s]	24,55
é	33,46	é	35,02	[eː]	35,69	t	79,42	t	82,72	[t]	81,58
f	9,18	f	9,59	[f]	9,04			ty	0,27	[c]	4,09
g	33,80	g	22,69	[g]	19,82	u	10,18	u	10,73	[u]	11,19
		gy	12,70	[ɟ]	11,45	ú	3,01	ú	3,06	[uː]	2,69
h	15,32	h	13,07	[h]	17,56	ü	5,51	ü	5,85	[y]	5,54
i	44,06	i	46,39	[i]	47,28	ű	1,86	ű	1,86	[yː]	1,74
í	5,82	í	5,60	[iː]	5,51	v	19,89	v	20,80	[v]	21,52
j	11,19	j	11,98	[j]	14,27	w	0,28	w	0,29		
k	49,22	k	51,46	[k]	53,63	x	0,36	x	0,38		
l	62,27	l	60,78	[l]	58,46	y	22,71	y	0,21		
		ly	3,77			z	43,48	z	26,48	[z]	24,51
m	35,00	m	36,56	[m]	36,61			zs	0,73	[ʒ]	2,21
n	58,12	n	53,78	[n]	54,37						
		ny	7,02	[ɲ]	8,21						

csak időtartam-különbség van, spektrális nincs). Ennek ellenére a karakterstatisztikához képest jobban tükrözi a nyelv tulajdonságait, mert az egyszerűsítések fonetikailag megengedhető helyeken történtek. A betűstatisztikával összehasonlítva az elemek hasonló gyakorisággal szerepelnek

Szende adataival összehasonlítva, a magánhangzók közül az [yː] és a [y] estében 50% eltérést tapasztaltunk. A legkisebb eltérés az [i] esetében volt megfigyelhető. Ha a teljes statisztikára vetítve vizsgáljuk az eltéréseket, akkor 1 százalékpont volt a legnagyobb, amely elhanyagolható. A leggyakoribb mássalhangzókat vizsgálva az adatok eltérőek, Szende 68,8 [n] hangot számolt meg 1000-ből, míg itt csak 54,4 hang volt. A [t] hang esetében fordított a helyzet, Szende 65 hangjával szemben itt 81,6 hang szerepel. Gósy (2004b) spontán beszédre készített hangstatisztikát, amelyben a magánhangzó-mássalhangzó arány 43% és 57% volt. Itt ez az arány 42% és 58%. A leggyakoribb hangot összehasonlítva szintén hasonló számokat kaptunk, az [ɛ] hang Gósy statisztikájában 11,4%-os gyakoriságú, itt 10,7%. A 4.2. táblázat betűstatisztika részében mind a 44 betű gyakoriságát megtalálhatjuk. Ez a 44 betű a vizsgált szövegkorpusz 99%-át alkotja. A speciális betűk csak 1%-ot tesztnek ki. Az ábécé betűi közül a *dz, dzs, q* szerepel nagyon ritkán, a 1 millió szóban átlagosan 20–40 db található.

Az *y* betű [j] hangként való realizációja gyakoribb, mint az [i] hangként való megjelenése. Ez abból adódik, hogy idegen nevek többször szerepelnek (például *Toyota*), mint a történelmi nevek (például *Dessewffy*).

A *ch* betű leggyakrabban [x] hangként jelenik meg, majd [t͡ʃ] hang a második leggyakoribb formája, [k] hangként ritkán ejtjük.

Az új típusú betűstatisztika használható kutatási feladatokra, a magyar szöveges állományok statisztikai tulajdonságainak vizsgálatára (Milyen gyakran jelöl a *ch* betűkapcsolat [x] hangot?). Az algoritmus felhasználható beszédadatbázisok készítésekor a felolvasandó szövegállományok elemzésére, válogatására. Például megbecsülhető, hogy egy adott szöveg felolvasása esetén a felolvasott szöveg egy kiválasztott hangból elegendő számút fog-e tartalmazni.

4.3. táblázat. Betűstatisztika speciális betűkre

Betű-hangjel	1 000 000-ból
ch_cs	28,12
ch_h	129,04
ch_k	2,33
ck_k	4,96
cz_c	25,24
$h_{\text{néma}}$	60,27
$h_{\text{zöngés}}$	2470,19
sch	85,09
ts_cs	34,25
tz_c	11,84
y_i	65,27
y_j	111,27
$j_{\text{zöngétlen}}$	3,59

4.6.2. A magyar szavak eloszlásai

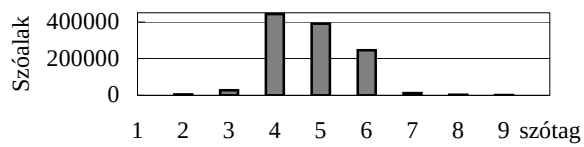
Kétfajta statisztikát adunk meg, szóalakokra és szavakra vonatkozót. A szóalakok fontossága abból adódik, hogy a magyar ragozó nyelv, egy-egy szótőhöz számos rag, jel, toldalék kapcsolható. Ennek következtében az egy szótagú szavaktól egészen az igen hosszú szóalakkokig terjed a hosszúsági skála. Szóalaknak tekintünk minden szövegszót, ami szóköz között szerepel az ortografikus szövegben.

Az alábbi szóalak-statisztika azt mutatja meg, hogy milyen a magyar szóalakok gyakorisága a szótagszám függvényében. A mérés nyelvi anyaga egy 80 millió szóból álló szövegkorpusz volt (Németh–Zainkó 2002), amiből kiválogattuk a szóalakokat (betűkép szerinti válogatással). Különbözőnek tekintettünk két szóalakot, ha két szóköz közötti betűsor egyetlen karakterben eltért. Eredményként kaptunk egy 1,5 millió szóalakból álló korpuszt, amelyben minden szóalak különbözött, és mind-

4.4. táblázat. Betűstatisztika gyakorisági sorrendben

Betű	db/1000	Betű	db/1000	Betű	db/1000	Betű	db/1000
e	101,31	d	20,42	ú	3,06	y_i	0,065
a	92,85	sz	19,27	$h_{zöngés}$	2,47	$h_{néma}$	0,060
t	82,72	h	13,07	ű	1,86	q	0,040
l	60,78	gy	12,70	zs	0,73	ts_cs	0,034
n	53,78	j	11,98	x	0,38	ch_cs	0,028
k	51,46	p	11,65	w	0,29	dz	0,026
i	46,39	ö	11,39	ty	0,27	cz_c	0,025
r	44,41	u	10,73	y	0,21	dzs	0,017
o	40,21	ó	10,49	ch_h	0,13	tz_c	0,012
s	39,08	f	9,59	y_j	0,11	ck_k	0,005
á	37,58	ő	9,35	sch	0,09	$j_{zöngétlen}$	0,004
m	36,56	ny	7,02			ch_k	0,002
é	35,02	ü	5,85				
z	26,48	í	5,60				
g	22,69	c	3,97				
v	20,80	cs	3,91				
b	20,56	ly	3,77				

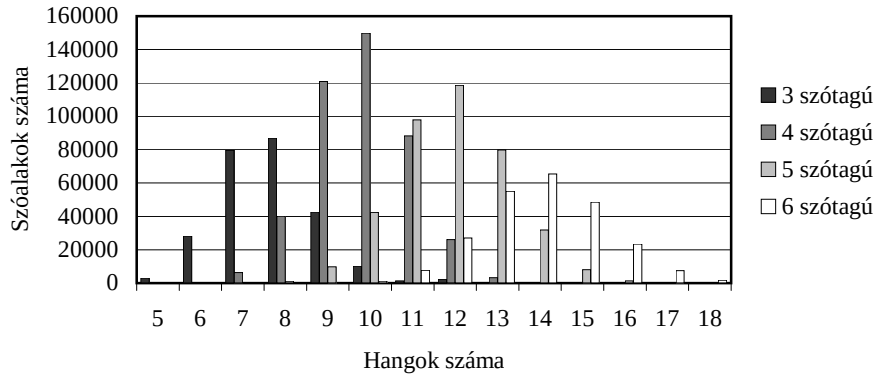
egyik csak egyszer fordult elő. Minden szóalaknak meghatároztuk a szótagszámát. Ebből készült a magyar szóalakok statisztikája a szótagszám függvényében (4.3. ábra). A szótagok száma szerinti eloszlás képe azt mutatja, hogy nyelvünkben a 4 és 5



4.3. ábra. Magyar szóalakok szótagszám szerinti eloszlása az összes szóalak függvényében

szótagú szóalakokból van a legtöbb, majd a 6 és 3 szótagúak következnek. Legkevesebb az 1 és 2 szótagú, valamint a 8 és 9 szótagú szavakból van. A szótagszám szerinti szóalakcsoportok szóalakjai hosszban önmagukban is eltérnek egymástól attól függően, hogy hány hangból épülnek fel. A hangszám meghatározására felhasználtuk a ProfiVox hangátíró algoritmust (Olaszy et al. 2000a), valamint a magyar szavak elektronikus kiejtési szótárát (Abari–Olaszy 2006). A szóalakcsoportok hangszám szerinti eloszlását a leggyakoribb csoportokra a 4.4. ábra mutatja.

A magyar szavak szótagszám szerinti gyakoriságát a Magyar Nemzeti Szövegtár anyagán és ehhez kapcsolt egyéb magyar szövegkorpuszok összességén mértük meg (150 millió szó). A gyakoriságot a 4.5. ábra mutatja. A mérésbe nem számoltuk bele az *a*, *az* névelőket. A gyakorisági adatok szerint a leggyakrabban két szótagú szavakat használunk a szövegekben, mármint akkor, ha a névelőktől eltekintünk.



4.4. ábra. A 3, 4, 5 és 6 szótagú szóalakcsoportok szavainak hangszám szerinti eloszlása

4.5. ábra. Magyar szavak előfordulásának gyakorisága szövegekben a szótagszám függvényében, ha a határozott névelőket nem vesszük bele a mérésbe

A BESZÉD SZERKEZETI ELEMZÉSE

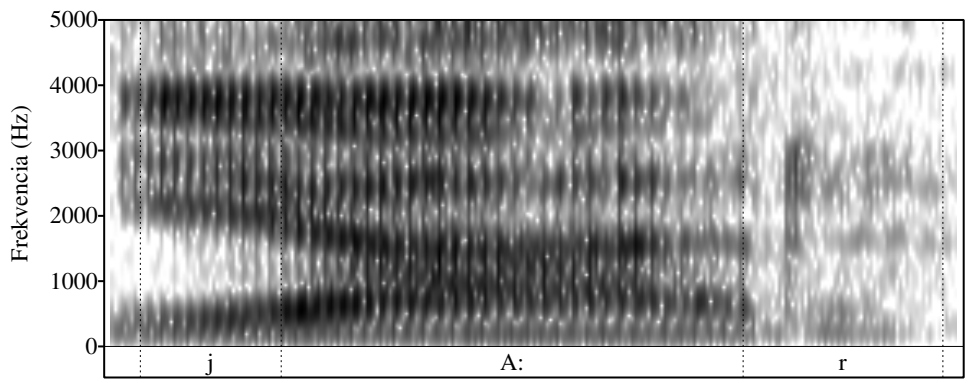
5. fejezet
A beszéd szegmentális szerkezete

Olaszy Gábor

A beszédet – mint korábban már említettük – elméletileg két szerkezeti részegységre lehet bontani, szegmentálisra és szupraszegmentálisra. A szegmentális szerkezethez tartoznak azok a komponensek, amelyek alapvetően szükségesek a beszéd létrehozásához, ezek létrejötte többnyire akaratunktól független. Ilyenek a beszédhangok és hangkapcsolódások szerkezeti elemei, a hozzájuk tartozó nyelvi időtartamok és arányok (specifikus időtartamok), valamint a hangok egymáshoz viszonyított hangintenzitás-különbségei (specifikus intenzitások). Ezek alkotják a beszédképzés (a hangtest) alapvető elemeit. A szegmentális szerkezet elemeit felhasználva már érthető beszéd hozható létre (főleg géppel). A szupraszegmentális szerkezethez tartoznak a beszéddallam, a hangsúlyozás, a ritmikai megvalósítások és a hangszínezet. Ezeket összefoglalva szokták prozódiának nevezni. A prozódia elemeinek megvalósítása inkább akaratfüggő, és ez főleg a felolvasott beszédben érvényesül. A szegmentális és szupraszegmentális szerkezet egyszerre van jelen a beszédprodukcióban, azonban tárgyalásukkal külön fejezetben foglalkozunk.

Az artikuláció akusztikai vetülete. Az artikuláció által folyamatosan változtatott mechanikai üregrendszer egyfajta rezonátorteret képvisel. Ha ezt az üregrendszert a zöngehanggal gerjesztjük, akkor a keletkezett hang formánsstruktúrája jellemezni fogja a pillanatnyi artikulációs állapotot. Ezt a formánsképet nevezzük **akusztikai vetület**nek, és adatszerűen az F1, F2, F3 formánsfrekvenciákkal fejezzük ki (a magasabb formánsok mozgása elhanyagolható). Az akusztikai vetület tehát egyfajta kapcsolatot teremt az artikulációs csatorna pillanatnyi térbeli formációja és a keletkezett hang között. Ha változnak az artikulációs csatorna fizikai méretei, változik az akusztikai vetület is. A definíciónk szerint **minden pillanatnyi artikulációs konfigurációnak megfelel a saját akusztikai vetülete**. Az akusztikai vetület független a gerjesztéstől. A zöngétlen hangokra vonatkozó vetületeket a zöngés párjuké képviseli, a vegyes gerjesztésű hangoknál pedig a zöngés elemre vonatkozó három formánsérték. Az akusztikai vetület az artikulációs pozíciót fejezi ki, függetlenül

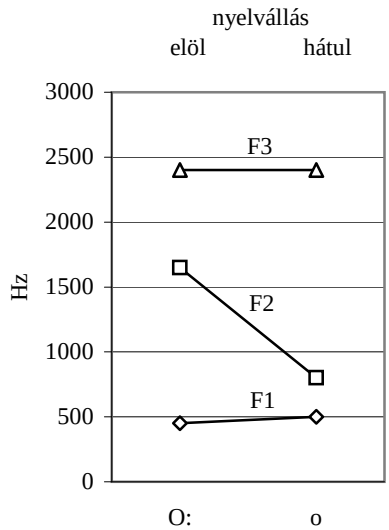
attól, hogy van-e hangkibocsátás, vagy nincs az adott beszédhangban. Az akusztikai vetület minden hangra és hangátmeneti szakaszra értelmezhető. Pontosabban fogalmazva azt mondhatjuk, hogy a zöngés hangok esetében minden koartikulációs konfigurációnak az adott periódus időpontjában létezik a saját akusztikai vetülete (ez olvasható le a széles sávú dinamikus hangspektrogramok függőleges bordázatú időtengelyén minden borda formánsadataiban). A beszéd akusztikai rezgése tehát úgy is jellemezhető, mint az egyes beszédhangokra jellemző akusztikai vetületek sorozata és azok összekapcsolása (5.1. ábra). Gépi beszéd-előállításnál ilyen alapon működnek például a formánsszintetizátorok. A gépi beszédfelismerésnél is lényeges szerepe van az akusztikai vetületek ismeretének, útmutatást adhatnak, hogy milyen hang következik a hangsorban.



5.1. ábra. A beszédhullám minden zöngés periódusa (függőleges vonalkák) saját akusztikai vetületet képvisel (*jár* szó 526 ms)

A magánhangzókra egyszerű meghatározni az akusztikai vetületet, hiszen a formánsok folyamatosan mérhetők. Vegyünk egy példát az [øː]-[o] magánhangzóátmenetre a *kőomlás* szóban. Ebben az esetben a két hang képzésében csupán egyetlen artikulációs paraméter változik, a nyelv az elülső helyzetéből hátrafelé mozog. Hogyan fejeződik ki ez a mozgás az akusztikai vetületekben? Úgy, hogy az F2 formáns magasabb értékről [øː] egy alacsonyabbra [o] változik (5.2. ábra). A kapcsolódás két végpontján létrejövő akusztikai vetületek adják a két magánhangzóra jellemző hangzást, a köztes értékek pedig az átmenetét. Az [øː] artikulációjakor a jellemző akusztikai vetület: F1 = 450, F2 = 1600, F3 = 2400 Hz, az [o] jellemző vetülete: F1 = 500, F2 = 700, F3 = 2400 Hz. Az átmeneti fázisban (amikor a nyelv elölről hátrafelé mozog) pedig annyi akusztikai vetületi pont határozható meg, ahány periódus van a két végpont közötti köztes hangszakaszban, vagyis a hangátmenetben. Minden egyes periódusra vonatkoztathatunk egy-egy akusztikai vetületet (5.1. táblázat).

A formánsalapú beszédszintetizátorok vezérlésénél ugyanezt hajtjuk végre. Minden hangperiódusra megadjuk az arra jellemző formánsokat.



5.2. ábra. Az artikulációs mozgás akusztikai vetületének folyamatos változása sematikus ábrázolásban az [ø:o] hangkapcsolatban

A táblázatból látható, hogy az átmenet során olyan formásértékek is megvalósulnak, amelyekkel nem hozható kapcsolatba egyetlen magyar beszédhang sem.

5.1. táblázat. Az [ø:]-[o] hangkapcsolat átmeneti fázisának periódusaiban mérhető akusztikai vetületek F1, F2, F3 értékei Hz-ben, amelyek a nyelv elölről-hátra mozgásának az eredményei. Feltételezzük, hogy a hangátmenet 7 periódusnyi idő alatt zajlik le

Átmeneti hangperiódus	Hang	F1	F2	F3	Megjegyzés
	[ø:]	450	1600	2400	A hang vége
1		450	1490	2400	Az átmenet kezdete
2		460	1380	2400	Átmeneti fázis
3		470	1265	2400	Átmeneti fázis
4		480	1150	2400	Átmeneti fázis
5		490	1040	2400	Átmeneti fázis
6		495	930	2400	Átmeneti fázis
7		500	815	2400	Az átmenet vége
	[o]	500	700	2400	A hang eleje

Ha a beszélő a *kőomlás* szó kiejtésekor, mondjuk az átmenet 4. periódusában megállítaná az artikulációját, a zöngeképzést azonban folytatná, akkor egy F1 = 480, F2 = 1150, F3 = 2400 Hz-es hangzót hallanánk (ilyen szerkezetű hang nincs a magyarban). Minden hangátmeneti periódus tehát más-más hangzást képvisel az átmenetben. A percepciós rendszer folyamatában dolgozza fel az ilyen fizikai jelenségeket, és a hangátmenetek belső részhangzásait nem tekinti számára ismeretlennek. Így van kódolva a percepciós bázisunk. A fenti példában azt láttuk, hogy az artikulációs mozgások egyetlen paraméterének változása milyen akusztikai vetületsorozatot

hoz létre. A beszéd során a beszélő az esetek többségében egyidejűleg több artikulációs mozgást is végez (nyelv, ajkak, állkapocs, uvula), és a pillanatnyi akusztikai vetületekben ezeknek mozgásoknak az összegzett végeredménye tükröződik. Mivel az artikuláció folyamatos, az akusztikai vetületek változása is az. Ezzel a folyamatos változással lehet akusztikailag jellemezni a hangátmeneteket, mivel bármely két hang találkozásakor azok akusztikai vetületei kapcsolódnak. Egy hangsor artikulációs folyamatát úgy transzformálhatjuk át jó közelítéssel a hangzást jellemző spektrális térbe, hogy leírjuk az akusztikai vetületek sorozatával. Megjegyezzük, hogy a gondolat meg is fordítható. Az akusztikai szerkezetből visszakövetkeztethetünk az artikulációs mozzanatok sorozatára. Ezt jelfeldolgozási eljárásokkal meg is lehet tenni (7.2. fejezet).

Az artikuláció akusztikai vetülete a mássalhangzókra is hasonlóan levezethető. Ezzel jellemezzük a továbbiakban a CC kapcsolódási pontok akusztikai képét. A későbbi adatok jobb megértése érdekében kitérünk itt még egy fontos jellegzetességre az artikulációs vetületekkel kapcsolatban. Az egyes hangokra megállapított artikulációs vetület formánsadatai (ugyanazon személynél) sem állandók, hanem sávokat alkotnak (hasonlóan, mint a magánhangzók formánsa). Bizonyos határok között az F1, F2, F3 értékek mozoghatnak a szomszédos hangok függvényében (vö. Olaszy 1985), vagyis a szomszédos hangok artikulációs vetületei befolyásolják egymást. Így van ez a CV, VC és a CC kapcsolatoknál is. Vegyük példának az [l] hang adatait. Más az akusztikai vetülete a [lo] hangkapcsolatban és más a [li]-ben (mélyebb hangzású érzetet ad az elsőben, magasabbat a másodikban). A későbbiekben egy adott hang akusztikai vetületének megadásakor ezeket az elmozdulási lehetőségeket tól-ig adatokkal érzékeltetjük, vagyis megadjuk a mozgási sáv határértékeit. Az [l] általános akusztikai vetületét tehát a következőképpen adjuk meg: F1 = 400–500 Hz, F2 = 1300–1500 Hz, F3 = 2400–2800 Hz. Azt, hogy ezeken a sávokon belül konkréten milyen érték valósul meg, az [l]-hez csatlakozó hang akusztikai vetülete határozza meg. Az itt elmondottakból következik, hogy az ugyanazon képzési hellyel rendelkező hangok akusztikai vetülete jó közelítéssel meg fog egyezni. Tehát az akusztikai jellemzésnél sokkal kevesebb akusztikai vetülettel kell dolgoznunk, mint ahány beszédhang van. Megjegyezzük, hogy egyes hangkapcsolatokban nemcsak a szomszédos hangok, hanem az azokat megelőző, illetve követő hangok is befolyásolhatják az adott hang belső akusztikai tartalmának a változását.

A következő alfejezetekben részletesen végigvesszük a beszéd szegmentális elemeit.

5.1. A magyar beszédhangok

A magyar nyelv rendszere 65 fonémát használ. Ezen fonémák megszólaltatott formáit tekintjük beszédhangoknak. A beszédhangok akusztikai szempontból jól megkülönböztethetők egymástól, azonban ejtésük mindig a beszélő személytől, annak mentális és fiziológiai jellemzőitől is függ. A beszédhangok két fontos tulajdonsága, hogy egyrészről önmagukban is többféle hangzásban jelenhetnek meg bizonyos határok között (más a *király k*-ja és a *katona k*-ja), másrészről, hogy hangátmenettel kapcsolódnak egymáshoz. Ezek a koartikuláció következményei. Ha a fonéma beszédhang-realizációja függ a hangkörnyezettől (bizonyos koartikulációs helyzetekben), akkor a fonéma variánsát (allofón) találjuk meg a beszédben (más a *támad m*-je, mint a *kámfor m*-je). A beszédhangok száma tehát nagyobb, mint a fonémáké. A beszélés tehát egy folyamatosan változó akusztikai rezgést eredményez, amelyben beszédhangokra jellemző elemek és az őket összekapcsoló hangátmeneti részek találhatók. A beszéd folyamatosságát a beszédtechnológiában kitüntetett figyelemmel kezelik.

5.1.1. A beszédhangok osztályozása

A beszéd építőelemeit sokféle szempont szerint lehet osztályozni, attól függően, hogy milyen céllal tesszük ezt. A legfontosabb osztályozásokból adunk példákat.

Magánhangzó-mássalhangzó. Ez a klasszikus osztályozás. Nyelvi szempontból indokolt, mivel a magánhangzó szótagképző, a mássalhangzó a legtöbb esetben nem. A magánhangzóknál nincs akadály az artikulációs csatornában, a mássalhangzóknál van. A mássalhangzó elnevezés tradíció, mivel vannak olyan mássalhangzók is amelyek ejtéséhez nem szükséges másik hang, önmagukban is lehet kitartva ejteni őket, hasonlóan, mint a magánhangzókat ([m] [n] [j] [ʒ]).

Rövid-hosszú hang. Fonológiailag a magyarban létezik a rövid-hosszú oppozíció, azaz a hosszú hangok fonémaértékűek (*varr, var*). Nyelvi szerkezet szempontjából tehát indokolt ez az osztályozás. Fizikailag azonban ezek a kategóriák egymásba is csúszhatnak, a nyelvileg hosszú hang a hangsorban rövidebb is lehet, mint a rövid párja. Ez a percepciót többnyire nem zavarja, de a jelfeldolgozásban nehezíti a modellezést.

Orális-nazális. Az artikulációs csatorna szerinti osztályozáskor bizonyos akusztikai szerkezeti különbségeket tudunk szétválasztani. Például a nazális palatális zárhang zárszakaszára jellemző hangrész lényegesen különbözik az orális ugyanilyen képzési

helyű zöngés zárhangétól (*nyanya, baba*). Az előbbiben az amplitúdó lényegesen magasabb, mint az utóbbiban. Ez a nazális képzésből ered, hiszen a [ɲ] hang esetében az orális zár alatt is van hangkiáramlás az orron keresztül, míg a tisztán orális zöngés zárhangnál nincs.

Gerjesztési hely. A gége szintjén jön létre a zönge (mint kváziperiodikus rezgés), amely a hangszalagok működésének az eredménye. A zönge képződési helye tehát állandó. A beszédben előfordulnak zörejelemek is. A turbulens áramlásból keletkező zörej (mint gerjesztés) az artikulációs csatorna különböző pontjain létrehozott szűkületekben keletkezik. Ez utóbbi képződési helye tehát változó.

Gerjesztési forma. A kétféle gerjesztés helyének különbözősége lehetővé teszi, hogy a beszédhangot tisztán zöngés vagy tisztán zörejes, illetve zöngés-zörejes gerjesztéssel hozzuk létre. A legösszetettebb beszédhang tehát a zöngés-zörejes gerjesztésű, mivel a zöngés komponensre zörejelemek is szuperponálódnak. Ezzel az osztályozással tehát három csoportba lehet osztani a beszédhangokat: zöngés, nem zöngés és kevert gerjesztésűek.

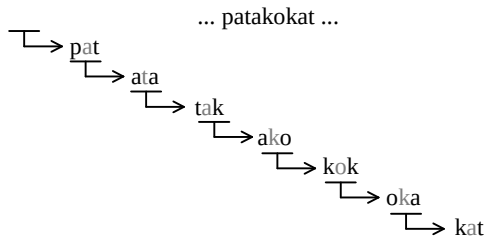
Egyszerű-összetett szerkezetű hang. Ez az osztályozás bizonyos képzési módokból következik. Az egyszerű szerkezetű hangokban az akusztikai szerkezetben nincs lényeges változás a hangon belül. Az ilyen hangok folyamatosan, kitartva is ejthetők (magánhangzók, réshangok). Az összetett szerkezetű hangok belső időszerkezete megosztott, két részből tevődik össze. Ez azt jelenti, hogy vagy a frekvenciakomponensekben vagy a hangintenzitásban (általában mindkettőben) jelentős változás áll be a hang valamely belső pontján. Úgy is szokták jellemezni ezeket a hangokat, hogy folyamatosan nem ejthetők. Ilyen hangok azok, amelyeknél az artikuláció során zárat képezünk rövid időre. Ilyenek az orális zár- és zár-rés hangok függetlenül attól, hogy zöngések, vagy zöngétlenek (például a *cica* szó mássalhangzói zárszakaszból és réselemekből állnak). A zár- és zár-rés hangok zárszakaszának és az utána következő réshangszakasz időszerkezeti aránya hangonként változó (Kovács 2002). A leghosszabb a zárszakasz a bilabiális orális zárhangokban, a legrövidebb az alveoláris zár-rés hangokban. Az összetett szerkezetű zöngétlen hangoknál százalékosan szokták kifejezni a két szerkezeti elem arányát. Az ilyen osztályozás beszédtechnológiai szempontból lehet fontos. Például a rövid zár- és zár-rés hangok hosszú változatát a zárszakasz jelének megnyújtásával lehet létrehozni.

5.1.1.1. A beszédhangok specifikus időtartamai

A specifikus időtartam elméleti fogalom. Olyan alapidőtartamot fejez ki, amelyik a beszédképzés alap-, azaz szegmentális szintjén jellemző a hangra. Ilyenkor csak az artikulációból eredő hatások befolyásolják a hang időtartamát. A specifikus időtartamok számszerű meghatározása nehéz, mivel a fenti kritérium szerinti beszéd nem fordul elő a természetes beszédképzés során.

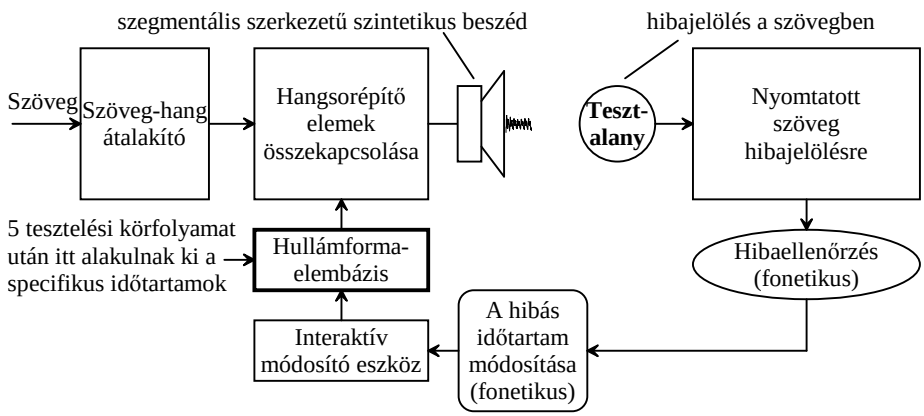
Korábbi nemzetközi kutatásokban az *„intrinsic duration"* terminussal, illetve az *„inherent duration"* meghatározással találkozhatunk, ezek közelítik a fenti definíciót. Lehiste (1970, 18. o.) szerint *„egy beszédhang időtartamát sok fonetikai faktor befolyásolhatja. Egy szegmens időtartamát bizonyos mértékig meghatározza a saját természete, azaz az artikuláció módja és helye. Az intrinsic duration terminussal tehát hivatkozhatunk a szegmens időtartamára, amit a fonetikai minőség határoz meg"*. Ez a meghatározás a szegmens saját fonetikai minőségét veszi figyelembe. A specifikus időtartam-terminus ezt a gondolatot folytatja azzal, hogy figyelembe kívánja venni a folyamatos artikulációból adódó befolyást is (hogy milyen hang előzi meg és milyen hang követi a vizsgált hangot). Az angolra elkészített beszédszintézismodellben (Allen et al. 1987) az időtartamokat az *„inherent duration"*-ból mint alapból építették fel. A szerzők ezt így határozták meg: *„Az inherent duration nem jelent speciális státuszt, szerepe csak annyi, hogy kiindulási pontot adjon a szabályoknak. Ez az időtartam közelítőleg megfelel a hangsorba helyezett értelmetlen CVC kapcsolatok ejtésekor keletkező időtartamnak. Például: Kimondom, hogy BAB és BAB. Kimondom, hogy VAV és VAV"*(i. m. 94). Ebben a megfogalmazásban az *„inherent duration"* közel áll a mi meghatározásunkhoz. A jelen modellünkben annyival léptünk túl Allen meghatározásán, hogy nem egy szótagú hangsorokból határoztuk meg a jellemző időtartamot, hanem magából a folyamatos beszédből. Így közelítettük elgondolásunkat a természetes beszédben lezajló folyamathoz. A természetes hangsorban (ha például szöveget olvasunk fel) a hangok egymás után kapcsolódnak, tehát automatikusan rendelkezésünkre állnak olyan hármas hangkapcsolatok, amilyenekről Allen beszélt. Erre mutat példát az 5.3. ábra. Ha a hármas hangkapcsolat középső hangjára sikerül meghatározni a specifikus időtartamot, akkor egyrészről figyelembe vettük a környezeti artikulációs hatást, másrészről nem szakadtunk el nagyon a természetes beszédfolyamatra jellemző állapottól, a folyamatosságtól. Ha minden egymás után következő hármas hangkapcsolatot sikerül így feldolgozni, akkor minden hangra minden hangkörnyezetben megkapjuk a rá jellemző egyedi specifikus időtartamot. Ezek átlagolásából származtatható az adott hármas hangkapcsolat középső hangjának a specifikus időtartama.

A specifikus időtartamok meghatározására inverz eljárást dolgoztunk ki, amelynek lényege, hogy nem közvetlen méréssel, hanem szintetikus beszéd percepciós tesztjével jutottunk az adatokhoz (Olaszy et al. 2000a). Az eljárásban két újdonság szerepelt. Egyrészről a szintetikus beszéd használata, másrészről, hogy nem fizikai



5.3. ábra. A hangsor hármas hangkapcsolatainak feldolgozása balról jobbra a *patakokat* szó esetében. A specifikus időtartamot a középső hangban mérjük

időtartammméréssel állapítjuk meg a mért beszédhang időtartamát, hanem közvetetten, percepciós ítélet alapján. A méréshez speciális szerkezetű, beszédtechnológiai eljárással kialakított beszédet használtunk, olyant, amelyben nincsen prozódia, ezért közel áll ahhoz a kritériumhoz, hogy csak a szegmentális szintű artikulációs hatások alakítják a hangidőtartamokat. A tesztben résztvevő személyek ilyen szerkezetű folyamatos beszédet (mondatokat) hallgattak, és meg kellett jelölniük a túl hosszúnak, illetve túl rövidnek hallott hangokat. Ezek után időkorrekciót végeztünk a kritizált hangkapcsolatokban, majd újból meghallgatták a beszédet. Így több forduló után kialakult az a helyzet, hogy nem volt több kritika a hangidőtartamokkal kapcsolatosan. Az így kialakított hangidőtartamokat tekintettük specifikus értékeknek. Az alkalmazott módszer tipikusan az analízis szintézissel eljárás (5.4. ábra). Az eljárás részletes leírása Olaszy (2006a) munkájában található.



5.4. ábra. A specifikus hangidőtartamok meghatározására kialakított eljárás folyamatábrája

A specifikus időtartamok hangkapcsolatfüggőek. Értéküket az adott nyelv és azon belül pedig az artikuláció és a koartikuláció befolyásolja. Amennyiben a hangátmenet létrehozásához nem szükséges bonyolult artikulációs mozgásokat végezni, a hang specifikus időtartama rövidebb lesz, ellenkező esetben hosszabb. Példaként be-

mutatjuk az [o] hangra kapott adatok egy részét az 5.2. táblázatban. A számadatokból kitűnik, hogy a mért magánhangzó időtartama széles határok között mozog, és attól függ, hogy milyen hang előzi meg, illetve milyen követi. A leghosszabb az [o] hang, amikor palatális mássalhangzók veszik közre (ToT hangkapcsolat esetében 115 ms), a legrövidebb, amikor [k] előzi meg és [m n] követi (a kom, illetve kon hangkapcsolatoknál 74 ms). A két szélső érték között 40 ms van, ami beszédhangszinten jelentős értéknek számít. A magyar beszédhangokra jellemző specifikus időtartamok átlagait az 5.3 és az 5.4. táblázat mutatja.

5.2. táblázat. Az [o] hang (bal felső sarok) specifikus időtartamai C1VC2 kapcsolatokban, ha C1=zárhang (első oszlop), C2 pedig tetszőleges mássalhangzó (első sor). A számadatok a magánhangzó időtartamát jelentik ms-ban. A hangokat az E1-jelű hangszimbólumokkal jelöltük. A mérésben használt szegmentális szerkezetű gépi beszéd artikulációs sebessége 10,5 hang/s volt

o	**b**	**p**	**d**	**t**	**g**	**k**	**G**	**T**	**m**	**n**	**N**	**j**	**h**	**v**	**f**	**z**	**s**	**c**	**Z**	**S**	**C**	**l**	**r**
b	88	93	84	95	93	90	93	103	83	84	94	95	95	85	94	93	90	94	94	85	83	84	94
p	88	93	83	95	92	90	92	103	82	83	93	95	95	84	93	93	90	94	94	84	83	83	93
d	86	91	82	93	91	88	91	101	81	81	92	93	93	83	91	91	88	92	92	83	81	82	92
t	84	90	80	92	89	86	89	100	79	80	90	92	92	81	90	90	87	91	91	81	80	80	90
g	87	92	83	94	92	89	92	102	82	82	93	94	94	84	92	92	89	93	93	84	82	83	93
k	79	84	75	86	84	81	84	94	74	74	85	86	86	76	84	84	81	85	85	76	74	75	85
G	90	95	85	97	94	92	94	105	84	85	95	97	97	86	95	95	92	96	96	86	85	85	95
T	99	104	95	106	104	101	104	115	94	95	105	106	106	96	105	105	101	106	106	96	95	95	105

5.3. táblázat. A magyar CVC helyzetű magánhangzók specifikus időtartamai folyamatos beszédre (ms-ban) Olaszy (2006a) alapján, az E1-hangjelöléseket alkalmazva

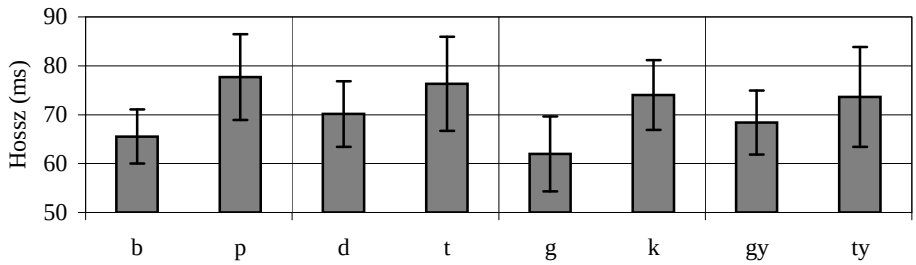
	i	u	U	o	a	e	O	E:	A:
Átlag	80	86	86	90	91	91	92	146	164
Szórás	7,07	9,54	9,06	7,64	7,06	8,88	6,8	10,44	14

5.4. táblázat. A magyar VCV helyzetű mássalhangzók átlagos specifikus hangidőtartamai ms-ban Olaszy (2006a) alapján, az E1-hangjelöléseket alkalmazva

C	b	p	d	t	g	k	G	T	m	n	N	j	h
Időtartam	65	77	70	76	62	74	68	76	67	48	66	59	62
Szórás	5,55	8,75	6,85	9,63	7,66	7,13	6,54	10,22	6,68	7,07	10,51	15,5	7,91
C	v	f	z	s	Z	S	dz	dZ	c	C	l	r	
Időtartam	61	85	68	82	67	83	128*	132*	92	98	52	37	
Szórás	9,57	6,58	4,62	10	11,57	6,15			6,76	8,04	8,12	6,35	

Mivel magyar beszédhangok specifikus időtartamait korábban még nem határozták meg, ellenőrző méréseket végeztünk, melyek során a korábbi kutatásokból megállapított tendenciákat hasonlítottuk össze a specifikus időtartamokkal. Ezekből mutatunk be egy példát. Ismert Magdics (1966) méréseiből, hogy a zöngés zárhangok átlagosan rövidebbek, mint a zöngétlenek. Ez az eltérés a specifikus időtartamoknál

is kiolvasható az 5.4. táblázatból. Megvizsgáltuk, hogy szignifikáns-e ez az eltérés. Minden párra elvégeztük a Student-féle párosított t-próbát egyenlő elemszámokra. Az eredmény 0,05 konfidenciaszintre számolva a következő: a [b p]-re a zöngétlen hang hosszabb időtartama szignifikáns ([t(160)= 10,55, p<0,0001]), [d t]-re is ([t(160)= 4,72, p<0,0004]), [g k]-ra is ([t(160)= 10,34, p<0,0001]) és [J c]-re is ([t(160)= 3,88, p<0,0001]) (5.5. ábra). A magyar beszédre jellemző, a hangkörnyezettől függő specifikus időtartamok minden hangkapcsolatra megtalálhatók Olaszy (2006a) munkájának függelékében. Ezeket az időtartamokat használták fel az 1990-es évek végén egy magyar időmodell alapjaként is. Az időmodell lényege, hogy szabályokat ad arra, hogy az itt közölt specifikus hangidőtartamokból hogyan lehet előállítani a hangsorban ténylegesen mérhető hangidőtartamokat. A modell helyes működését beszédszintézissel igazolták (Olaszy et al. 2000a).

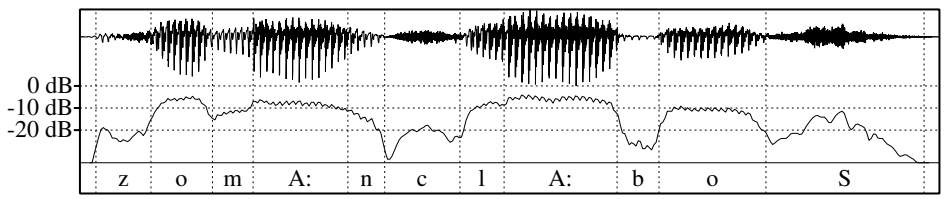


5.5. ábra. A zöngés/zöngétlen párok specifikus időtartamának különbsége a zárhangokra

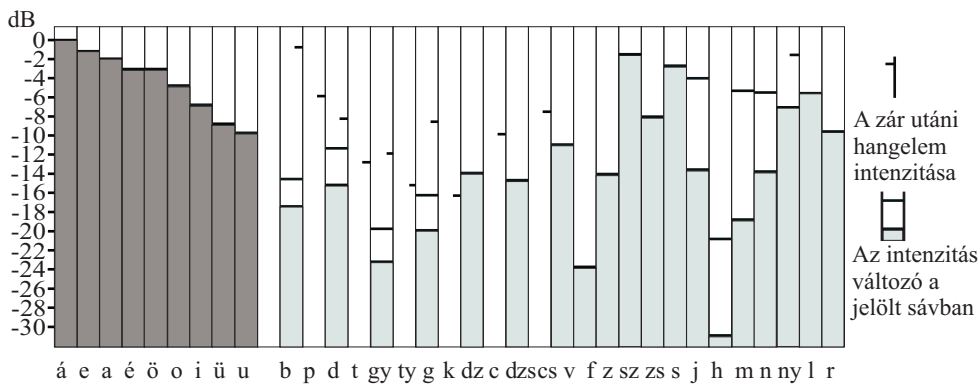
5.1.1.2. A beszédhangok specifikus intenzitásai, hangzósság

A fonetikában a hangzósséggal fejezik ki a beszédhangok egymáshoz viszonyított erősségét (Laziczius 1944, Olaszy 1989a, Gósy 2004b). Ez azt jelenti, hogy a beszédhangokat különböző hangerősségűnek halljuk akkor is, ha teoretikusan a képzésük során ugyanolyan gégeszintű levegőenergiával képezzük és ugyanolyan távolságról hallgatjuk őket. Ilyen szempontból a leghangzósabb az [aː], a legkevésbé hangzós a [h] hang. A beszédben a beszédhangok közötti relatív intenzitásszint függ a hangképzési energiától és az artikulációtól. A gégeszintű hangképzés adja az alaphangerőt (például a zönge jelének amplitúdója), az artikuláció pedig ezt befolyásolhatja. Azoknál a hangoknál, amelyeknél a zönge amplitúdója nagy, az artikulációs csatornában nincs akadály, a levegőáramlás útjában és a szájnyílás is nagy, a kisugárzott hang intenzitása is nagy ([aː]). Amennyiben a gége szintjén kis amplitúdójú a jel, és az artikulációs csatorna akadálymentes, kicsi lesz a kisugárzott hang intenzitása ([h]). Fizikai szempontból a specifikus hangintenzitás értékét a hangra jellemző spektrális komponensek összegzett intenzitása határozza meg (lásd részletesen az egyes hangoknál). Minden hangnak megvan a hozzávetőleges specifikus intenzitásszintje a

többi hanghoz képest. Ez hangsorba szerveződéskor is jellemzi a hangot (5.6. ábra). Az átfogási sáv mintegy 30 dB. Ha ezek az intenzitásszintek túllépik a hangra jellemző intervallumot (például gépi beszéd-előállításnál rossz erősítési adatot adunk meg), akkor a beszéd hangzása darabos, huppogó, lüktető lesz, ami egyrészről negatív benyomást tesz a hallgatóra, másrészről a megértést is erősen leronthatja. Ezért fontos a megfelelő intenzitásszintek kialakítása beszédszintézisnél is. A beszédfolyamatra jellemző specifikus intenzitásokat a 5.7. ábrán mutatjuk be. A hangzósság érzeti fogalom, a specifikus intenzitás fizikai.



5.6. ábra. A hangintenziás alakulása a *zománclábos* (1,23 s) szóban (alsó görbe)



5.7. ábra. A magyar beszédhangok jellemző specifikus intenzitásszintjei a mért szegmentális szintű beszédben. A hangokat a betűjelükkel jelöltük

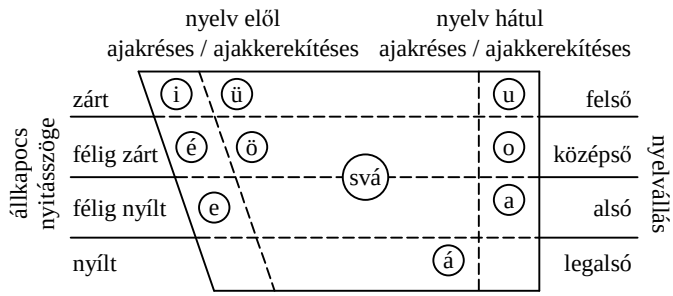
A beszéd intenzitásviszonyainak más, magasabb szintjei is vannak. A hangsúlyos részeknél a beszélő nagyobb hangerőt használhat, mint a nem hangsúlyosaknál. A beszélő saját kifejezésmódja, akaratlagos ejtése is befolyásolja a beszéd komplex intenzitásszerkezetének a kialakulását. Végül a beszéd intenzitásviszonyait meghatározza az is, hogy egy-egy közléshez meghatározott levegőmennyiségre van szüksége a beszélőnek. Ha a levegő fogyóban van, az intenzitás csökken (például a mondat végén). A beszédhangok intenzitásának alakulását különböző hangsorokban jól

tanulmányozhatjuk a következő honlapokon http://fonetika.nytud.hu/cccc és a http://fonetika.nytud.hu/cvvc.

5.1.2. A magyar magánhangzók

Ebben a fejezetben a magyar magánhangzók artikulációs és akusztikai sajátságait mutatjuk be (a nyelvjárási változatokat nem érintjük). A magánhangzók a hangsor lényeges információhordozó elemei, intenzitásuk általában lényegesen nagyobb, mint a mássalhangzóké. A magyar hangrendszerben 9-féle magánhangzó létezik: [aː ɔ o u i y eː ø ɛ]. Ebből ötnek van nyelvileg hosszú párja [oː uː yː iː øː], az [aː]-nak pedig rövid, ez az [a]. Tehát fonémaszinten 15 magánhangzóval kell számolni. A rövid-hosszú oppozíció magánhangzóinak artikulációja és spektrális szerkezete eltérhet egymástól. Egyes hosszú magánhangzók zártabb ejtésűek, mint a rövidek, ami abban nyilvánul meg, hogy a nyelv állása kissé magasabbra kerül, az ajakartikuláció is változik. A rövid-hosszú magánhangzók között tehát nem csak időtartamkülönbség van, hanem hallható akusztikai is lehet (*honlap, hónap*). A hangzásbeli különbség mértéke hangonként változik, a legkisebb eltérés az [i] [iː] és az [aː] [a] között mérhető. A rövid [a] főleg idegen szavakban fordul elő, de a hosszúsági jegye jelentést megkülönböztető. A *bájt* szó jelentheti a számítástechnikai kategóriát, vagy a kedvességet is. A rövid [a] a helyesírásunkban egyaránt szerepelhet *á*, illetve *a* betűvel (*Fiat, sztrájk*).

A magánhangzókat három fő artikulációs alapparaméterrel jellemezhetjük: a nyelv, az ajkak és az állkapocs állása. Mindhárom paraméter belső tere több szintre osztható: a nyelv függőleges állásának két szélső pontja az alsó és felső állás; a vízszintes mozgás két végpontja az elülső és a hátsó nyelvállás. A végpontok között lehetnek köztes fokozatok (Gósy 2004b). Az ajkakra két állapot jellemző: ajakréses, illetve ajakkerekítéses hang. Az állkapocs nyitási fokának két sarokpontja a zárt és a nyílt állapot, közötte még két fokozatot különböztetünk meg. A magyarban csak orális magánhangzók képeznek fonémaértékű beszédhangot. A téma részletesebb tárgyalása Gósy (2004) munkájában található. A magánhangzók egyszerű szerkezetű hangok, kitartva is ejthetők, zöngés gerjesztésűek. A zönge kváziperiodikus jelként gerjeszti a gége feletti artikulációs csatornát (szupraglottális üregek). Az artikuláció során a nyelv, az ajkak és az állkapocs helyzete határozza meg az adott magánhangzó hangzását (5.8. ábra), vagyis, hogy a kiejtett hang melyik beszédhangnak felel meg a nyelvi rendszerben. Az ábra sarokpontjait alkotó [aː u i] hangokat egyes esetekben kiemelik a magánhangzórendszerből és külön kezelik (lásd az 5.12 ábrát). Vannak olyan magánhangzók, amelyek csak egy képzési paraméterben különböznek egymástól [i y], [y u], [eː ø], [ø o]. Ezekkel a hangokkal személyes ejtési kísérletek is végezhetők. Például az [i]-ből [y]-be az ajkakat kell csak összecsücsöríteni és a

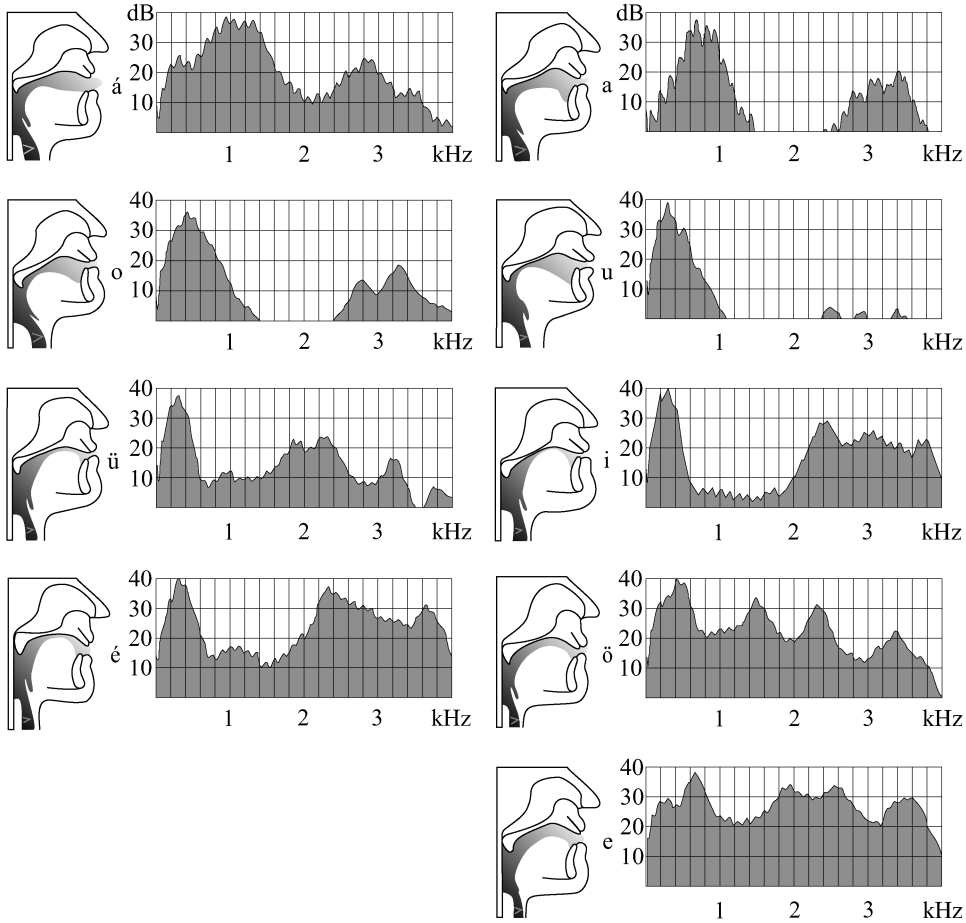


5.8. ábra. A magyar magánhangzók artikulációs konfigurációi. A hangokat a betűjelükkel jelöltük

hang meg fog változni. Az [y]-ből [u]-t képezhetünk, ha a nyelvet hátra mozdítjuk. Ha a képzési különbségeket összevetjük a formánsképekkel akkor az összefüggések leegyszerűsítve a következők. Az F1 elmozdulása a nyelvállás magassági fokával fordított arányosságban van ([aː]-magas F1; [u]-alacsony F1). Az F2 a nyelv vízszintes mozgásával hozható kapcsolatba ([u]-alacsony F2; [y]-magas F2). Az ajakréses magánhangzók F2-je magasabb frekvenciájú, mint az ajakkerekítéseseké ([eː]- F2 magas; [ø]- F2 alacsonyabb). Az összetett artikulációs mozgásoknál ezek a hatások együttesen érvényesülnek.

Említést kell tennünk az úgynevezett semleges magánhangzóról is (svá, fonetikai hangjele a [ə]), amely az 5.8. ábra középpontjára jellemző artikulációs paraméterekkel jellemezhető. A svá képzését úgy kell elképzelnünk, mint ha az 5.8. ábra [ø] hangját jobbra, középre tolnánk el (mediális és illabializáltabb ejtés). Emlékeztetünk arra, hogy a beszédképzés fizikai csőmodellje (3.3.5.1. fejezet) is hivatkozik a svá hangra, mint amelyik a legjobban reprezentálja az egyenlő keresztmetszetű artikulációs csatornát. A semleges magánhangzó a magyarban leginkább hangos szünetként (hezitálás) jelenik meg, ilyenkor időtartama hosszú (Gósy 1997). Előfordulhat CC hangkapcsolódásoknál is (*vadra*), bizonyos mássalhangzók kapcsolódásánál (Olaszy 2007b), ilyenkor csupán 25–35 ms-os (lásd 5.2.3. fejezet). Svá hangot ejthetünk zárhangok végén is hangsorzáró helyzetben, ha nyomatékosan ejtjük az utolsó mássalhangzót (*vám*= [vaːmə], *volt*= [voltə]). A svá hang a magáhangzókhoz hasonló frekvenciaszerkezettel rendelkezik, formánsainak jellemző értéke F1 = 500 Hz, F2 = 1500 Hz, F3 = 2500 Hz. A svára is vonatkozik a koartikulációs hatás, tehát az előbbi formánsokból az F2 magasabb frekvenciára is tolódhat, amikor CC kapcsolatban a svá-t tartalmazó C1 mássalhangzót palatális C2 mássalhangzó követ. Például a *fo**gny**úlvány* szó ejtésekor a zöngés zárhangban keletkező svá F2-je felfelé mozog a 2000 Hz-es érték felé. A beszédtechnológiai adatbázisokban a hangátírásoknál svá-t célszerű jelölni a hangsor fonetikai átiratában, amennyiben hangos szünetként van jelen. A svá hang jelenlétével főleg az automatikus beszédfelismerésnél kell hangsúlyozottan számolni, mivel magánhangzóra emlékeztető tulajdonságokkal bír, és ezért megtévesztheti a felismerő algoritmust.

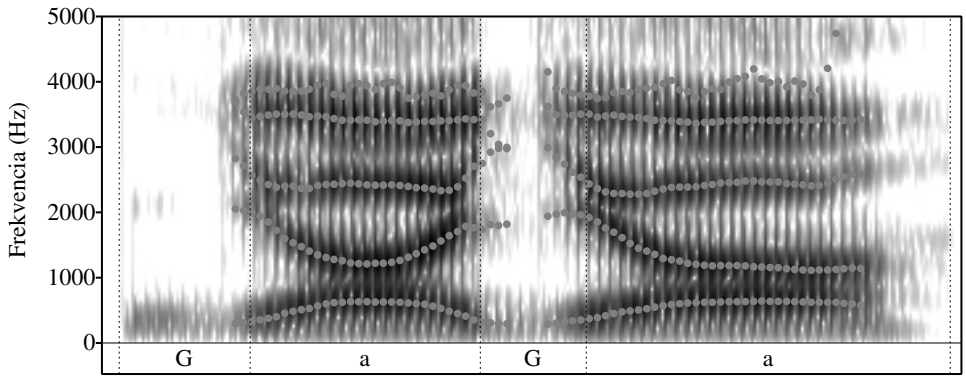
A magyar magánhangzók jellemző nyelv- és állkapocsállásait, azok jellemző spektrumképeit az 5.9. ábrán mutatjuk be.



5.9. ábra. A magyar magánhangzók artikulációs nyelvállásai és a hozzájuk tartozó spektrális szerkezet

A magánhangzók formánsai egy adott sávon belül mozognak. Az, hogy a sávban hol van az adott formáns értéke, egyrészről attól függ, hogy milyen hangokhoz kapcsolódik a magánhangzó, másrészről pedig attól, hogy milyenek a beszélő fiziológiai sajátosságai. Az artikulációs csatorna hossza egyértelműen befolyásolja a formánsok elhelyezkedését. A hosszabb artikulációs csatorna (férfiak) lefelé, a rövidebb (nők) felfelé tolja a formánsokat (vö. a 3.3.5.1. fejezettel). A hangkapcsolódásokban a koartikulációs hatás következménye, hogy mozoghatnak a magánhangzók formánsai (a két hang artikulációs vetülete hat egymásra). Például C1-V-C2 kapcsolatokban a magánhangzók formánsai a C1-V kapcsolódási pontjára

jellemző akusztikai vetületből indulva mozognak a V-re jellemző formánsértékek felé, majd onnan tovább a V-C2 kapcsolódási pontra jellemző akusztikai vetület irányába. Amennyiben a C1 és C2 akusztikai vetülete között nagy a frekvenciakülönbség, akkor a V formánsaiban is jelentős mértékű elmozdulások lehetnek (5.10. ábra). A magyar magánhangzókra jellemző formánsértékeket már sok kutató mérte,



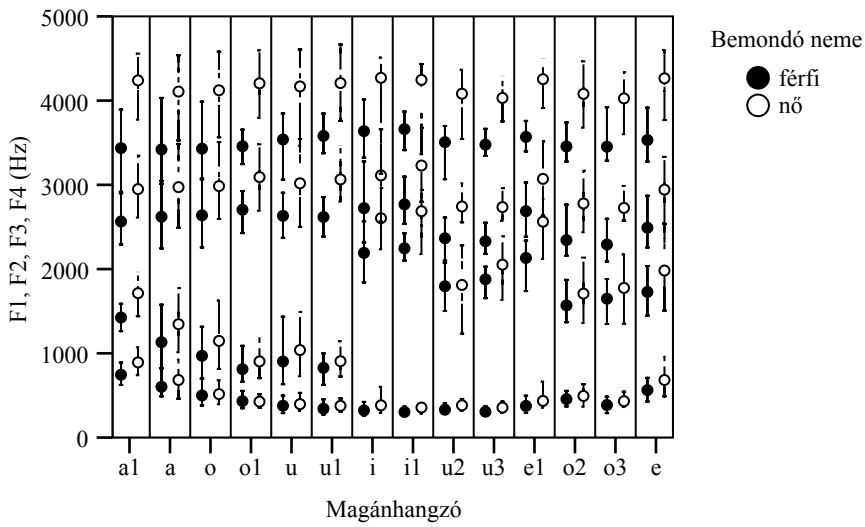
5.10. ábra. A formánsok intenzíven mozognak a *gyagya* (653 ms) szó ejtésekor

adatokkal tehát bőven el vagyunk látva (Magdics 1966, Olaszy 1985, 1989a, Gósy 2004b). A korábbi mérések adatai mind kézi, egyedi és egyszeri feldolgozásból származtak, a méréseket nemigen lehet megismételni. Ebben a könyvben olyan adatokat közlünk, amelyek beszédadatbázison végzett szoftveres, automatikus mérésekből származnak. Az adatbázis egy férfi és egy női beszélő szólista felolvasásos hangfelvételéből áll. A beszédadatbázis biztosítja egyrészről a megismételhetőséget, másrészről azt, hogy más típusú mérések is bármikor elvégezhetők. Az első magyar formánsadattárból származó formánsadatokat mutatjuk be a magyar magánhangzókra (Olaszy et al. 2009), amely referenciaként is használható. A benne szereplő adatok ellenőrizhetők, ugyanis a mérés hanganyaga nyilvánosan hozzáférhető, bárki elvégezhet bármilyen formánsmérést akár újból is. Az adatbázist részletesen a 8.4.4. fejezetben ismertetjük. A formánsadatbázisból mutatunk be néhány lekérdezést a magyar magánhangzók jellemzésére. A lekérdezések adatainak feldolgozásából származnak az ábrák, az eloszlások és a táblázatok adatai.

A magyar artikuláció súlyozása. Az artikuláció súlyozása azt mutatja meg, hogy milyen formánsértékek vannak túlsúlyban a nyelvet reprezentáló beszédben. Ilyen kérdésre csak adatbázis felhasználásával tudunk választ adni. Erre tettünk kísérletet a formánmsadatbázis felhasználásával. Lekérdeztük az összes magánhangzó első három formánsának átlagértékét. A következő eredményt kaptuk. A férfi hangra: F1 =505 Hz, F2 =1488 Hz, F3 2592 Hz ; a női hangra: F1 =594 Hz, F2 =1737 Hz,

F3 2949 Hz. Ezek az értékek a semleges magánhangzóra emlékeztetnek. Tehát a magyar artikuláció kiegyensúlyozottnak tekinthető, szavak felolvasásából kapott formánsdatok alapján.

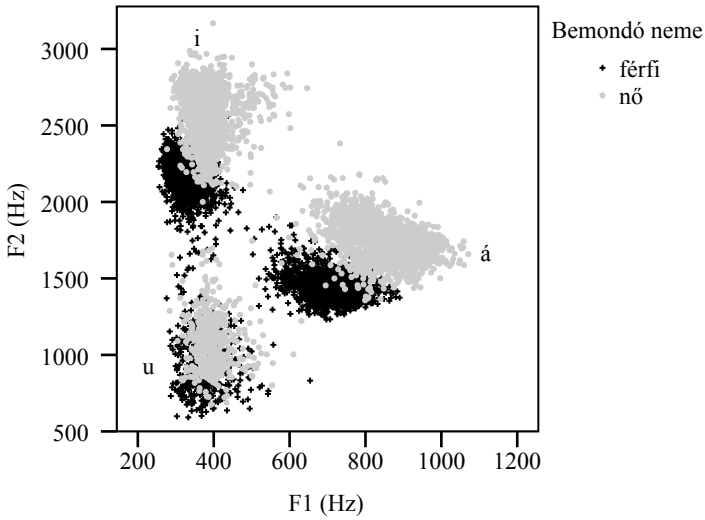
Formánssávok. Közismert tény szakmai körökben, hogy a férfiak formánsai alacsonyabb sávban helyezkednek el, mint a nőké. Az ok, hogy a nők artikulációs csatornája átlagosan rövidebb, mint a férfiaké. Ennek igazolását mutatja be a formánsadatbázis adatai alapján készített diagram (5.11. ábra).



5.11. ábra. Az F1, F2, F3, F4 formánsok frekvenciasávjai egy férfi és egy női bemondó ejtésből mérve a referencia-formánsadatbázis szóanyagán. A diagramok a magánhangzó 50%-os pontjában mért adatokat mutatják

Formánssíkok. A formássíkok informatív képeket adnak a magánhangzókról. A négy formánssal kapcsolatosan bármelyik kettőből képezhetünk síkot. Például a magánhangzók hangzását egyértelműen meghatározza az első két formánsuk. Ezért szokásos az F1-F2 síkban is ábrázolni a magánhangzók formánsainak eloszlását, ezen keresztül a magánhangzók realizációs halmazait (Tarnóczy 1974, Gósy 2004b). Az 5.12. ábra ilyen eloszlást mutat be az [a:u i] magánhangzókra. Az eloszlások elhelyezkedése egyértelműen utal a férfi és női beszélőre.

Hangátmenetek. A kialakított formánsadatbázis lehetőséget ad arra is, hogy CVC kapcsolatokban vizsgáljuk a C hatását a V-re, mivel minden magánhangzóra három adat szerepel a hangon belül, nevezetesen a 25%-os, az 50%-os és a 75%-os ponton. A három formánsadattal jellemezni lehet a formánsok mozgását a magánhangzón belül. Az ilyen típusú vizsgálatokhoz a kapcsolódási variáci-



5.12. ábra. Az [a: u i] magánhangzók F1, F2 formánsai által meghatározott halmazok egy férfi és egy női bemondó ejtéséből

ók száma meglehetősen nagy, mivel három hang határozza meg a mérési teret (25x25x9 = 5625-féle mérés). Egyszerűsíthető a helyzet, ha a mássalhangzókat a képzési helyükkel helyettesítjük, vagyis az akusztikai vetületre adunk meg hangcsoportokat. Egy képzési helyhez több mássalhangzó tartozik, ugyanakkor a képzési helyre jellemző akusztikai vetület azonos, ebből arra következtetünk, hogy az adott képzési helyhez köthető mássalhangzók hasonló hatást fognak gyakorolni a hozzájuk csatlakozó magánhangzók formánsaira. Hét képzési hellyel számolva (BL-bilabiális; LD-labiodentális; DA-dentialveoláris; AL-alveoláris; PA-palatális; VE-veláris, FA-faringális), már csak 49-féle kapcsolódási lehetőség van. Ha minden magánhangzóra külön akarjuk összegyűjteni az adatokat, akkor 49x9 = 441-féle kapcsolódási kombinációval számolhatunk. Ha csupán az [a: u i] magánhangzókra szűkítjük a mérés terét, akkor mindössze 108-féle kapcsolódási forma marad, amiből a jellemző tendenciákat már meg lehet határozni. Az ilyen mérésből kimutathatók, hogy léteznek-e ugyanazokat a formánsmozgási tendenciákat eredményező hatások a képzési helyek között (előfordulnak-e egyforma formánsmozgások a V-ben függetlenül a hangkapcsolódási környezettől).

Hangátmenet és hangkörnyezet. Hogyan hat a magánhangzó formánsaira a hangkörnyezet? Példaként bemutatjuk annak a lekérdezésnek az eredményét, amelyik arra keres választ, hogy a palatális mássalhangzók milyen hatást gyakorolnak az [o] magánhangzó formánsaira (változik-e az átlaghoz képest a formáns értéke). A méréshez kétféle lekérdezést végzünk. Megadjuk a palatális mássalhangzókra

szűkített hangkörnyezetet és így kérjük le az adatokat, majd azok kizárásával is lekérdezzük őket. Ebből a két adathalmazból végezzük el a statisztikai elemzést.

5.5. táblázat. Az [o] hang első két formánsának átlaga, ha palatális hang előzi meg, illetve ha nem

Hang	Hangkörnyezet	Bemondó	F1 50%	F2 50%	Mért esetek száma
[o]	Palatális előzi meg	Férfi	512	1050	18
[o]	Nem palatális előzi meg	Férfi	498	968	526
[o]	Palatális előzi meg	Nő	494	1258	17
[o]	Nem palatális előzi meg	Nő	517	1145	527

5.6. táblázat. Az [o] hang első két formánsának átlaga, ha szomszédai palatális hangok, illetve ha nem

Hang	Hangkörnyezet	Bemondó	F1 50%	F2 50%	Mért esetek száma
[o]	Palatálisok a szomszédai	Férfi	323	1153	3
[o]	Nem palatálisok	Férfi	378	890	162
[o]	Palatálisok a szomszédai	Nő	381	1256	2
[o]	Nem palatálisok	Nő	394	1029	158

A palatális környezet megemeli az F2 értékét az [o] hangban. A hatás gyengébb, ha csak az egyik oldalon van a magánhangzónak palatális szomszédja (5.5. táblázat), mint akkor, ha mindkettőn (5.6. táblázat).

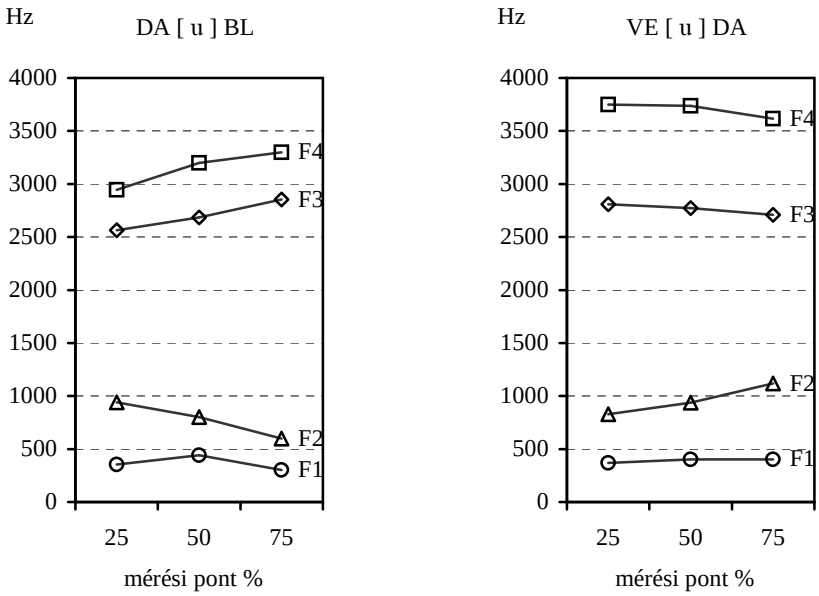
Formánstávolságok. Az adatbázisból lekérdezhetők olyan adatok is, hogy mennyi a jellemző formánstávolság valamely két formáns között ugyanabban a magánhangzóban a megelőző és követő mássalhangzó függvényében. Hasonlítsuk össze például a BL[aː]BL és a DA[aː]PA hanghármasok magánhangzójára jellemző formánstávolságokat férfi ejtésből. A mérésnél a fenti két mintacsoport összes reprezentánsát (azaz a megjelölt képzési helyekhez tartozó összes mássalhangzót) vettük figyelembe az adatbázisból. A kétfajta ejtésből az 50%-os mérési pontokra jellemző formánstávolságokat az 5.7. táblázat mutatja. A példában látható, hogy az

5.7. táblázat. A jellemző formánstávolságok az [aː]magánhangzó közepén két hangkörnyezeti konfigurációra

Hang	F1 és F2 átlagos távolsága Hz	F2 és F3 átlagos távolsága Hz	F3 és F4 átlagos távolsága Hz
BL[aː]BL	585	1335	813
DA[aː]PA	784	1098	863

F1-F2, illetve az F2-F3 távolságában lényeges az eltérés, az F3-F4 közel azonos. Az így meghatározott formánstávolságok jellemzőek lehetnek a hangkörnyezetre. Az ilyen jellemzők felhasználhatók a formánskinyerő algoritmusok finomítására is.

Formánsmozgások. Kirajzoltathatók a jellemző formánsmozgások (lineáris közelítéssel), és azok rendszerezhetők. Példaként bemutatjuk, hogy a férfi ejtésű [u] hang formánsmozgásai hogyan alakulnak a DA[u]BL, illetve az VE[u]DA hangkörnyezeti konfigurációban (5.13. ábra).

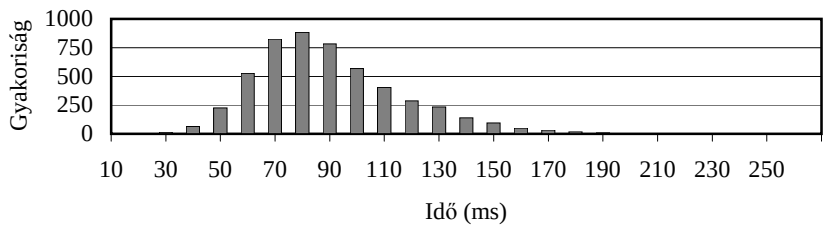


5.13. ábra. Jellemző formánsmozgások az [u] hangban a két hangkörnyezeti konfigurációban, férfi hangból mérve

A magánhangzók formánsai spontán beszédben. A spontán beszéd akusztikai sajátosságainak kutatása Magyarországon is elkezdődött, főleg a BEA spontánbeszéd-adatbázis (8.5. fejezet) hangfelvételeire alapozva. A spontán beszédben a beszédtervezés és a kivitelezés egyidőben zajlik (Gósy 1998), és ez kihat az artikulációra, ezen keresztül pedig a spektrális szerkezetre is. A beszélő személy az adott pillanatban válogatja ki a közlésre szánt gondolatokat és tervezi meg azok nyelvi formáját. Ennek következménye, hogy a magasabb szintű tervezési folyamatok befolyásol(hat)ják az artikulációs tervezéshez szükséges agyi aktivitást. Ez Beke (2009) mérései szerint azt eredményezi hogy lazul az artikuláció, azaz a magánhangzók képzése redukálódik. A laza ejtésből adódhat, hogy a magánhangzókra jellemző formánssávok is szélesedhetnek (Horváth–Gráczi 2010), ami együtt jár azzal, hogy az egyes magánhangzók közötti átfedések is növekednek a spontán beszédben. Beke–Gráczi (2010) tanulmányában az állapították meg, hogy az elemzett magánhangzók 47%-ában a beszédhang az alaprealizációtól eltérően valósult meg, egyszerűsödött. Azt is megállapították, hogy a magánhangzó gyakoriságának növekedésével nő a semleges realizációk előfordulása a spontán beszédben. A fentiekből látható, hogy a spontán beszéd akusztikai szerkezetének vizsgálata nehezebb, mint a tervezett beszédé. A modellezéséhez figyelembe kell venni többféle redukciós tényezőt is.

5.1.2.1. A magyar magánhangzók időtartamadatai

A magyar magánhangzók specifikus időtartamait már bemutattuk az 5.1.1.1. fejezetben. Itt most a magánhangzók felszíni időtartamait vizsgáljuk felolvasásos beszédben végzett mérésekkel. A vizsgálat kézzel címkézett beszédadatbázisra épült. A magyar magánhangzókra felolvasásos anyagból kapott átlagos hossz: 85,7 ms, (egyetlen férfi beszédéből mérve, 5178 hang számítógépes összegzése alapján). A hanghosszak eloszlását az 5.14. ábra mutatja. A beszélő személy artikulációs sebessége 13 hang/s volt. A mérés beszédadatbázisa a könyv honlapján megtalálható (`http://magyarbeszed.tmit.bme.hu`).



5.14. ábra. A magyar magánhangzók időtartamainak eloszlása. A vizsgált anyag: felolvasott beszéd férfi ejtésben, kézi hanghatárcímkézéssel

Részletezettebb képet mutat az 5.8. táblázat, amelyben a 9 magyar magánhangzóra lebontott átlagokat adjuk meg az előbbi korpuszból. A hangokat az időtartamátlagaik alapján rendeztük sorba, növekvő sorrendben. A magánhangzók hossza változik

5.8. táblázat. Magyar magánhangzók időtartamátlagai felolvasott beszédben

Hang	i	u	U	o	a	e	O	E:	A:
Átlag ms	69	75	76	81	79	82	78	108	120
Szórás	20,55	24,7	24	20,92	21,74	21,59	20,61	29,29	27,38
Min.	23	35	26	34	30	17	37	41	40
Max.	192	145	182	160	183	179	156	250	242
A mért hangok száma	624	163	88	579	1098	1406	151	389	387

annak függvényében, hogy a szóban hol fordulnak elő. A CVC helyzetű magánhangzók időtartamátlagát mutatjuk meg a szó elején, belsejében és végén (5.9. táblázat). A mérés eredménye azt mutatja, hogy a rövid magánhangzók a szó első szótagjában a leghosszabbak, a szó belsejében pedig a legrövidebbek. A legkisebb eltérés a három hanghelyzet között az [i] esetében van, ez a hang stabilan a legrövidebb. A legnagyobb időtartambeli különbség az [y]-nél látható, szókezdő helyzetben 16 ms-mal hosszabb, mint a szó belsejében. Az [aː] és [eː]-re mért adatok szerint ezek a hangok konzekvensen a leghosszabbak a szó elején, majd rövidebbek a szó belsejében és a legrövidebbek a szó végén. A mérésnél a szavak pozícióit a mondaton belül nem vettük figyelembe.

5.9. táblázat. A magyar magánhangzók átlagos hossza a szó eleji, szó belseji és a szó végi helyzetben CVC hangkörnyezetben, 13 hang/s-os artikulációs sebesség esetén

Hang	Helyzet	a	o	u	U	i	O	e	E:	A:
Átlag ms	Eleje	79	86	74	80	66	84	84	114	129
A mért hangok száma		489	227	91	40	233	68	561	191	131
Átlag ms	Belseje	73	75	64	64	63	71	77	99	115
A mért hangok száma		189	153	19	19	95	60	329	88	125
Átlag ms	Vége	79	79	71	77	64	71	82	105	114
A mért hangok száma		207	186	22	26	157	19	385	85	112

Hogyan függ a hangidőtartam a szó hosszától? Ez klasszikus kérdés, már a 20. század elején is feltették a kutatók. Gombocz (1909) híres és azóta többször megismételt kísérletére utalunk (*tát, tátog, tátogat, tátogatók, tátogatóknak*). A kísérlet azt bizonyította, hogy az első magánhangzó fokozatosan rövidebb lett a szótagszám növelésével. Itt most a leggyakoribb magyar magánhangzóra, az [ɛ]-re vonatkozó adatokat mutatjuk meg (5.10. táblázat) 1–5 szótagú szavakban a felolvasásos beszédadatbázisból mérve (más hangokra nem kaptunk statisztikailag feldolgozható adatot). A magánhangzót a szókezdő, a szó belseji és a szó végi szótagban mértük. Az eredmények szerint a szó eleji helyzetben a hang időtartama határozottan csökken a szótagszám növekedésével (vö. Gombocz 1909). Ugyanilyen csökkenést mutatnak a szó belseji adatok is, bár a csökkenés kisebb mértékű. A szó végi helyzetben a hang időtartama viszont nem mutat számottevő változást a szó belseji értékekhez képest. Ezek szerint a felolvasásos beszédben is érvényesül valamelyest az a hangrövidülési tendencia, amit szófelolvasáson mutattak ki. Az artikulációs energiával való gazdálkodás feltevése tehát itt is igazolódik.

5.10. táblázat. Az [ɛ] átlagos hossza szó eleji, szó belseji és a szó végi szótagban CVC hangkörnyezetben, a szó hosszának függvényében folyamatos felolvasásban egyetlen személy esetében, 13 hang/s-os artikulációs sebesség esetén

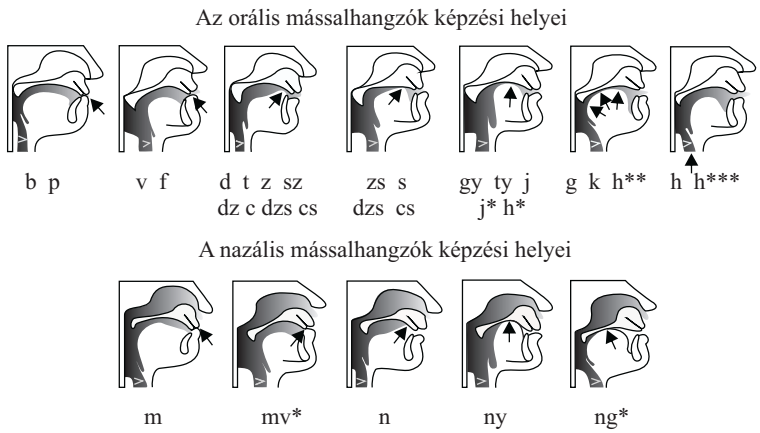
[ɛ]	Helyzet	1 szótagú szóban	2 szótagú	3 szótagú	4 szótagú	5 szótagú
Átlag ms	Eleje	88	86	85	81	78
A mért hangok száma		167	147	153	57	30
Átlag ms	Belseje	-	-	79	75	72
A mért hangok száma				148	94	55
Átlag ms	Vége	-	82	81	78	80
A mért hangok száma		-	199	119	45	18

Hangidőtartamok a spontán beszédben. Spontán beszédben végzett időtartammérésekről ritkábbak a publikációk, mint azok, amiket felolvasott beszédre alapoztak. Egy legutóbbi tanulmány (Gósy–Beke 2010) szerint mindössze néhány ilyen mérésről számoltak be a kutatók az utóbbi 100 évben. A kutatások a spontánbeszéd-adatbázisok egyre növekvő számával meg fognak szaporodni. A mérések nehézsége abban is rejlik, hogy az artikulációs sebesség jobban kontrollálható a felolvasott

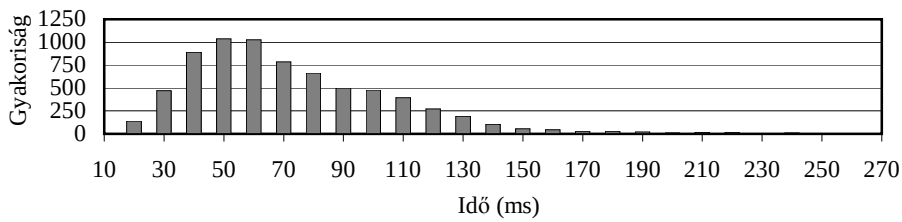
beszédben, mint a spontánban. Itt csak a fenti tanulmány adatait vizsgáljuk meg, mivel ez az első olyan mérés, amelyik nagyszámú beszédhangot vizsgált számítógépes támogatással. Az összehasonlítást Olaszy (2006) felolvasott beszédre vonatkozó munkájával tesszük, mivel ez a mérés is hasonló eszközrendszerrel dolgozott. A magánganzók átlagos hossza spontán beszédben 78 ms, felolvasottban 85,7 ms. A spontán beszéd rövid és hosszú magánhangzói (65 ms és 90 ms) is rövidebbek, mint a felolvasásból mértek (78,8 ms és 111 ms). Ugyanez a tendencia vonatkozik az egyes magánhangzókra. Mindezek az adatok azt mutatják, hogy az átlag szintjén a spontán beszédben a magánhangzók rövidebbek, mint felolvasásból származóban. Ugyanakkor a spontán beszédben a hangidőtartamok átfogási sávja nagyobb, mint a felolvasásnál, hiszen a spontán produkcióban nincs olyan vezérfonal, mint a felolvasásnál a szöveg.

5.1.3. A magyar mássalhangzók

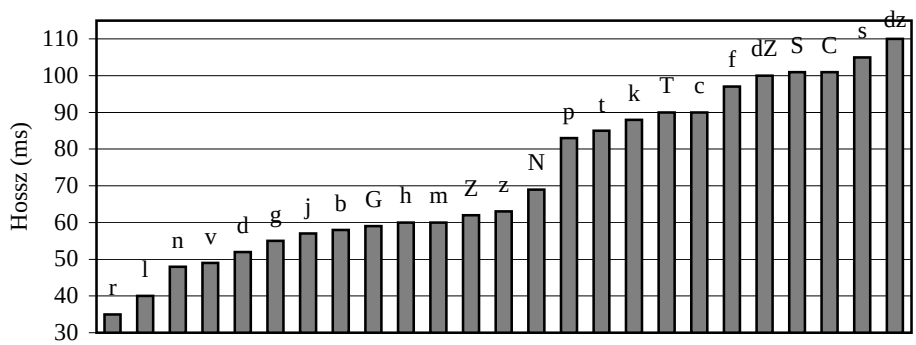
Ebben a könyvben 55 magyar mássalhangzóval foglalkozunk (ez a szám a magyar köznyelvi mássalhangzókat fedi le, nem tartalmazza a nyelvjárási ejtésváltozatokat). Tárgyalásunk a beszédtechnológiai kutatásokhoz alkalmazkodik. A mássalhangzókból 25-25 rövid-hosszú párt alkot, a fennmaradók pedig rövidek. A 25-25 rövid-hosszú pár jelentést megkülönböztető szerepű a nyelvben, tehát ezek megfeleltethetők fonémának, az ezen kívüliek variánsok (jelentést megkülönböztető szerepük nincs, helyesírásunk nem jelöli őket). A mássalhangzókat – az akusztikai szerkezetüket tekintve – három paraméterük alapján lehet egyértelműen meg határozni. Ezek a képzési hely, a mód és a gerjesztés formája (5.11. táblázat). A rövid-hosszú oppozíciót ebben az osztályozásban nem tekintjük paraméternek, ugyanis a rövid-hosszú mássalhangzók csak időszerkezeti szempontból térnek el egymástól, spektrális különbség nincs közöttük. Ezt a tényt ki is használják a beszédtechnológiában adatredukcióra. A hosszú mássalhangzók időtartama átlagosan a rövidek 1,5-szöröse. A mássalhangzók artikulációs konfigurációit az 5.15. ábra mutatja. A mássalhangzókra jellemző átlagos hangidőtartam 68,8 ms (Olaszy 2007b). Ez az érték 13 hang/s-os artikulációs sebességre vonatkozik. Az időtartameloszlást az 5.16. ábra mutatja. A mérési adatok egyetlen személy felolvasási beszédprodukciójából származnak. A rövid mássalhangzók hangokra lebontott átlagos időtartamértékeit a 5.17. ábrán adjuk meg. A mássalhangzók frekvenciaszerkezeti összesített adatait az 5.18. ábra mutatja.



5.15. ábra. A magyar beszéd mássalhangzóinak artikulációs konfigurációi (betűjelöléssel). A képzési helyeket nyíl jelöli. A csillaggal jelölt hangok variánsok



5.16. ábra. A magyar mássalhangzók időtartam-eloszlása felolvasásban a fonológiai hanghosszúság besorolásától függetlenül (Olaszy 2006a)



5.17. ábra. A rövid mássalhangzók átlagos hangidőtartamai hosszúság szerinti sorrendben (Olaszy 2006a)

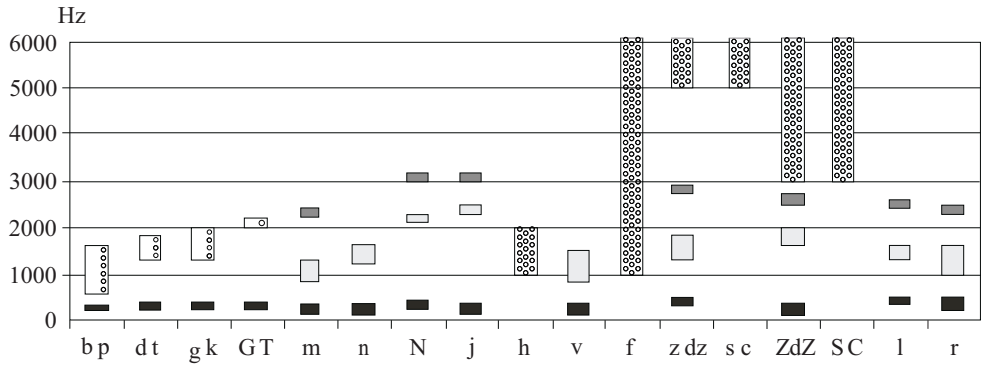
5.11. táblázat. A magyar mássalhangzók osztályozása a gerjesztés, a képzési hely és mód szerint (betűjelöléssel). A variánsokat csillaggal jelöltük

Képzési hely								
Képzési mód	Bilabiális	Labio-dentális	Dential-veoláris	Alveoláris	Palatális	Veláris	Faringális	Gerjesztés
Zárhangok	b							zg
	p							ztl
			d					zg
			t					ztl
					gy			zg
					ty			ztl
						g		zg
						k		ztl
Réshangok		v						zg
		f						ztl
			z					v
			sz					ztl
				zs				v
				s				ztl
					h*			ztl
						h**		ztl
							h	ztl
							h***	zg
					j*			ztl
Zár-rés h.			dz					v
			c					ztl
				dzs				v
				cs				ztl
Közelítő				l				zg
					j			zg
Pergőhang			r					zg
Nazálisok	m							zg
		mv*						zg
			n					zg
						ng*		zg
					ny			zg

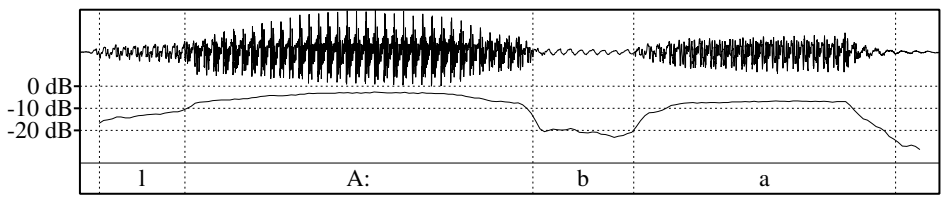
megjegyzés:
zg = zöngés, ztl = zöngétlen, v = zg+ztl gerjesztés egy időben, közelítő = közelítő hangok
h*=[ɦ], j* = [ç], h** = [x], h*** = [ɣ], mv* = [ɱ], ng* = [[ŋ]]

5.1.3.1. Zöngés zárhangok

A magyar zöngés zárhangok a [b d ɟ g]. Szerkezeti szempontból egységesen fojtott zöngéből és zárfelpattanásból állnak (5.19. ábra). A fojtott zönge a zárszakaszban jön létre és minden esetben a hang nagyobb részét teszi ki (70–90%-át). Az orális zárból következik, hogy a fojtott zönge képződése alatt hangkiáramlás nincs, tehát ennek a hangrésznek formánsai nincsenek. A fojtott zönge amplitúdója kicsi, 15–20 dB-lel marad el a magánhangzókétól. CV kapcsolatokban a zárfelpattanás beleolvad a magánhangzó kezdetébe (5.19. ábra). A zárfelpattanásból adódó hangszakasz (összefoglaló nevén zöngés zárfelpattanás) jellemzése a következő. A hossza a kép-



5.18. ábra. A magyar mássalhangzók frekvenciaszerkezeti elemei. A pöttyözött komponensek zörejt jelentenek, a többi zöngés gerjesztésű



5.19. ábra. A [b] hang szerkezete a *lába* (630 ms) szóban. A szó időfüggvénye (fent), az intenzitásfüggvény és a hangszimbólumok az E1-jelű karakterekkel (lent)

zési helytől függően változó [b] < [d] < [ɟ] < [g]. A zöngés zárfelpattanáshoz zörej is adódhat feszesebb képzés esetén. Ez leginkább a palatális, illetve a veláris zárhangnál jelentkezhet. A zöngés zárhangok képzési helyére jellemző akusztikai vetület formánsadatait az 5.12. táblázatban adjuk meg Olaszy (1985) adatai alapján. Egy CV kapcsolatban tehát a csatlakozó V formánsai a táblázatban megadott formánsértékekhez közelítenek a C-hez való kapcsolódási ponton. A megadott adatok tendenciákat fejeznek ki, hiszen a beszédben az akusztikai eredmény függ a beszélő fiziológiai és mentális adottságaitól is.

5.1.3.2. Zöngétlen zárhangok

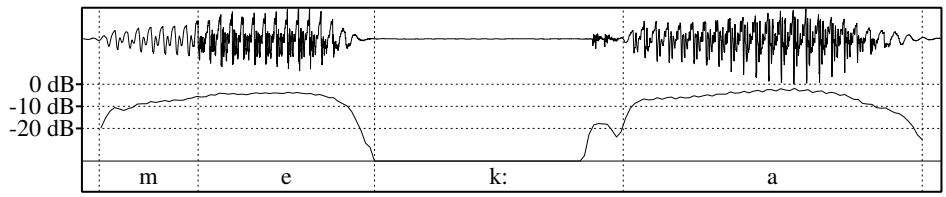
A magyar zöngétlen zárhangok, [p t c k], a zöngés zárhangok zöngétlen párjai. Artikulációs szempontból összetett szerkezetűek, egységesen zárrészből (néma fázis) és egy zárfelpattanással kezdődő rövidebb-hosszabb zörejes hangszakaszból állnak (5.20. ábra). A zárfelpattanáskor keletkező zörejkomponensek jellemző frekvenciaösszetevői az 5.18. ábrán láthatók. A néma fázis képzési szempontból a zöngés zár-

5.12. táblázat. A zöngés zárhangok képzési helyének akusztikai vetülete a hozzájuk kapcsolódó magánhangzó függvényében a kapcsolódási pontra jellemző tájékoztató formánsadatokkal kifejezve. A magánhangzókat az E1-hangszimbólumokkal jelöltük. A számértékek Hz-ben értendők

Képzési hely	Formáns	A:	a	o	u	U	i	E:	O	e
Bilabiális	F1	450	400	350	300	300	300	300	300	350
	F2	1300	900	700	600	1600	1800	1800	1300	1300
	F3	2500	2600	2700	2800	2300	2400	2500	2300	2300
Dentialveoláris	F1	450	400	350	300	300	300	300	300	350
	F2	1500	1400	1250	1000	1600	1800	1800	1500	1600
	F3	2800	2700	2600	2500	2400	2700	2700	2700	2700
Palatális	F1	450	400	350	300	300	300	300	300	350
	F2	2000	1900	1800	1800	2100	2200	2100	2000	2100
	F3	2600	2600	2600	2600	2600	3000	3000	2600	2800
Veláris	F1	400	350	300	300	300	300	300	300	350
	F2	1500	1000	800	600	1600	2200	2200	1600	2100
	F3	2500	2500	2500	2500	2500	3500	2600	2500	2700

hangokban jelen lévő fojtott zöngének a zöngétlen párja. A hang nagy részét ez az elem alkotja (70–90%). Ezekhez a hangokhoz két speciális fogalom is kapcsolható. Az egyik a virtuális hangidőtartam. Ez azt fejezi ki, hogy meghatározott hanghelyzetben ezen hangok időtartama fizikailag nem mérhető, tehát csak virtuálisan vannak jelen a hangsorban. Ilyen hanghelyzet a hangsorkezdő pozíció. A virtuális hangidőtartamot beszédtechnológiai feldolgozásnál figyelembe kell venni (kötött szótáras beszédszintetizátorok).

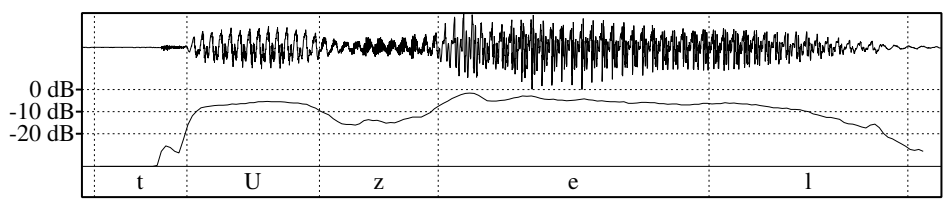
A másik fogalom a zárfelpattanáshoz kapcsolható, és elnevezése zönge kezdési idő (Voice Onset Time, VOT). Könyvünkben csak zöngétlen zárhangokra vonatkoztatjuk ezt a fogalmat, részletes tárgyalása Liebermann–Blumstein (1988) művében olvasható. A VOT-értéke tehát a CV kapcsolatokban a zöngétlen zárhang zárfelpattanásától a következő magánhangzó indulásáig eltelt idő, amit ms-ban adnak meg. A magyarra vonatkozó ilyen VOT értékek több korábbi munkában is megtalálhatók (Olaszy 1985, Gósy 2000b). Az átlagos tendencia az, hogy minél hátrébb van a képzési hely a zöngétlen zárhangban, annál nagyobb a VOT. Olaszy (2007b) átlagolt, számítógépes mérései alapján a [p]-re (16 ms), a [t]-re (23 ms), a [k]-ra (39 ms). A VOT-t értelmezhetjük CC kapcsolatokra is (Olaszy 2007b), lásd az 5.45. ábrán. A VOT értékei nyelvfüggők. A zöngétlen zárhangok akusztikai vetülete megegyezik a zöngés párjukéval, hiszen az akusztikai vetület a képzési helytől függ, nem pedig a gerjesztéstől. A zöngétlen zárhangokban a zárfelpattanási zörej intenzitása szintén hang- és artikulációfüggő, mintegy 15–30 dB-lel alacsonyabb, mint a magánhangzó maximuma.



5.20. ábra. A [kː] hang szerkezete a *Mekka* (602 ms) szóban. A gerjesztési formákat jól meg lehet különböztetni a hanghullámban. A zöngés elemek periódusai szabályos időközönként ismétlődnek, a néma fázisban gyakorlatilag nincs jel, a zárfelpattanás (22 ms) rövid zöreje szabálytalan rezgés. A fonológiai hosszúság akusztikai megvalósulása a néma fázis nyújtásában realizálódik

5.1.3.3. Zöngés réshangok

A magyarban zöngés réshangok a [v z ʒ], és még a [ɦ] hang is idesorolandó, amelyik a [h] zöngés variánsa. Ezek a hangok artikulációs szempontból egyetlen réselemből állnak. A képzett résen keresztül áramlik a hang, amely zöngés elemet és zörejt is tartalmaz. A keverési arány az artikuláció függvénye. Ha a rést szűkítjük, akkor egy pont után a zöngés rezgésből le-leszakadnak turbulens áramlások, amelyek zörejt szuperponálnak rá. A zörej jelenléte a [v]-re intervokális helyzetben nem jellemző (Siptár 1996, Kiss–Bárkányi 2006), inkább CC kapcsolatokban képezzük így a hangot (Bőhm–Olaszy 2007). A zörejkomponens szerkezete hasonló az [f]-éhez. A [z ʒ] viszont minden helyzetben tartalmaz zörejkomponenst (5.21, amelynek frekvenciaszerkezete hasonló az [s ʃ h] hangokéhoz (5.18. ábra). A [ɦ] hang variánsként vesz részt a beszédben, zöngés volta a környezeti hatásból ered. Két zöngés hang közé kerülve nagy valószínűséggel zöngésedik (*nahát*). Ilyenkor a zöngés komponens amplitúdója sokkal nagyobb, mint a zörejé. A zöngés réshangok intenzitása csak kismértékben (4–10 dB) marad el a környezetükben lévő magánhangzókétól. A zöngés



5.21. ábra. A [z] hang szerkezete a *tüzel* (624 ms) szóban. A zöngére szuperponálódott zörejt a feketedések mutatják az időfüggvényen (fent)

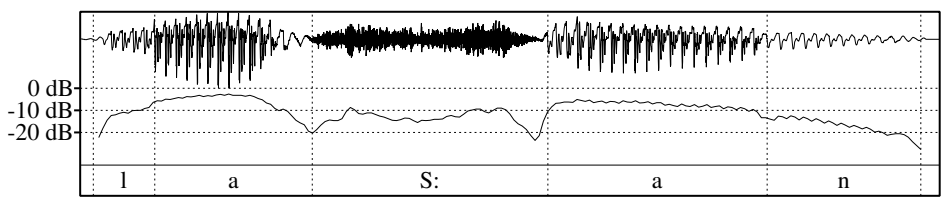
réshangok képzési helyére jellemző akusztikai vetület formánsadatait az 5.13. táblázatban adjuk meg. A zöngés réshanghoz csatlakozó magánhangzó formánsai a táblázatban megadott formánsértékekhez közelítenek a C-hez való kapcsolódási ponton. A megadott értékek tendenciákat fejeznek ki.

5.13. táblázat. A zöngés réshangok képzési helyének akusztikai vetülete a hozzájuk kapcsolódó magánhangzó függvényében a kapcsolódási pontra jellemző tájékoztató formánsadatokkal kifejezve. A magánhangzókat az E1-hangszimbólumokkal jelöltük. A számértékek Hz-ben értendők

Képzési hely	Formáns	A:	a	o	u	U	i	E:	O	e
Labiodentális	F1	450	400	350	300	300	300	300	300	350
	F2	1100	1000	900	800	1600	1700	1700	1500	1500
	F3	2500	2600	2700	2800	2300	2400	2500	2300	2600
Dentialveoláris	F1	300	300	250	200	200	200	300	300	300
	F2	1300	1300	1200	1200	1600	1800	1800	1500	1600
	F3	2800	2700	2600	2500	2400	2700	2700	2700	2700
Alveoláris	F1	350	300	250	200	200	200	300	300	350
	F2	1300	1350	1250	1200	1600	1800	1800	1500	1600
	F2	2800	2700	2600	2500	2400	2700	2700	2700	2700

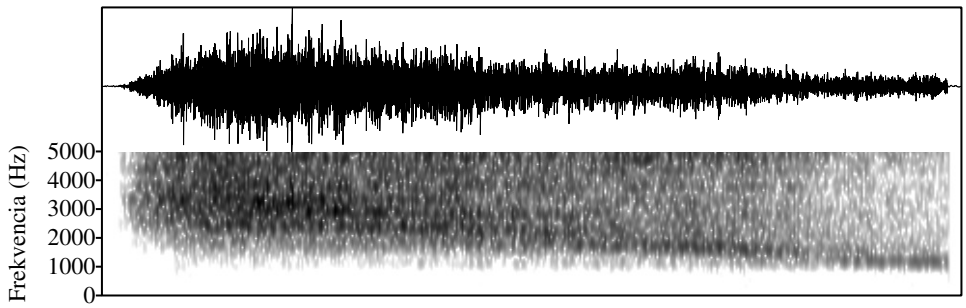
5.1.3.4. Zöngétlen réshangok

A magyar beszédben előforduló, fonémaértékű zöngétlen réshangok a következők: [f s ʃ h]. A [x] és [ç] hangok variánsként vesznek részt a hangsorépítésben. A zöngétlen réshangok képzésekor a hangszalagok nem rezegnek, fúvó, illetve h állásban vannak. A réshang zöreje fokozatosan alakul ki, intenzitása nő a hang közepéig, majd csökken a hang végéig (5.22. ábra). A zörej tehát a hang közepén a legerősebb. A hang intenzitása a középső szakaszon mintegy 5–20 dB-lel kisebb, mint a magánhangzókra jellemző érték. A hangok intenzitási sorrendje a következő: [h] < [x] < [f] < [ç] < [s] < [ʃ]. A réshang zörejének frekvenciaszerkezete karakte-



5.22. ábra. Az [ʃ] hang szerkezete a *lassan* (655 ms) szóban. A gerjesztésváltási pontot az időfüggvényen (fent) A jól meg lehet figyelni, a zörej szabálytalan, sűrű, tömött rezgésképet mutat

resen jellemző a hangra (5.18. ábra). A hangsorba szerveződéskor a magánhangzó akusztikai vetülete kis mértékben hatással van a zörej alsóbb frekvenciakomponenseire, amelyek elmozdulhatnak a V F2-jének a függvényében. A hátul képzett, labiális magánhangzók például mélyítik a réshang színét (más lesz az [ʃ] hangzása a *súg* szóban, mint a *sí*-ben). Az alsó zörejgóc frekvenciájának folyamatos változását mutatja az 5.23. ábra egy célzott artikulációjú hangfelvételből. A zöngétlen réshangok képzési helyére jellemző akusztikai vetület formánsadatai megegyeznek a zöngés párjaikéival, hiszen a képzési helyük azonos. Ez az akusztikai vetület azonban csak elméletileg fejezhető ki formánsértékekkel, mivel zöngés komponens nincs a



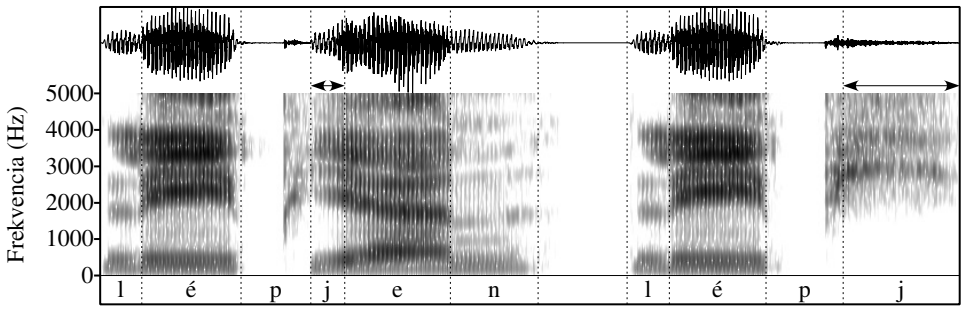
5.23. ábra. Az [ʃ] hang alsó zörejgócát (lent) befolyásolja a hozzá csatlakozó magánhangzó labialitási foka. Az ábra a réshang alsó frekvenciakomponenseinek változását mutatja célzott ejtésből az ajkak folyamatos labializációja függvényében az illabiális helyzetből kiindulva (bal) a labiálisig (jobb). Az amplitúdó alakulásából (fent) pedig az látható, hogy az nagyobb az illabiális ejtésű magánhangzóhoz csatlakozó réshangban, mint ellenkező esetben

réshangban. Hangsúlyozzuk, hogy ez az akusztikai vetület nincs kapcsolatban a zöngétlen réshangban jelen lévő, turbulens áramlásból keletkező zörej frekvenciakonponenseivet, az akusztikai vetület ebben az esetben csak a képzési helyre jellemző, formánsértékei csak a csatlakozó magánhangzó elején mérhető adatokból határozhatók meg. A magánhangzó átmeneti fázisában a formánsok a megadott akusztikai vetület formánsértékeihez közelítenek a C-hez való kapcsolódási ponton. A réshangok zörejelemének lényeges frekvenciakomponenseit a képzett rés tulajdonságai alakítják ki (lásd az 5.18. ábrán). Kirívóan magas frekvenciakomponensekkel az [s] rendelkezik (6000–10 000 Hz), ez a hang torzul leginkább a telefonsávi átvitel során, ahol a szabványos átviteli sávtartomány 300–3400 Hz.

A [h]-nak többféle artikulációs megvalósulása van (Siptár–Szentgyörgyi 2004). Az abszolút szókezdő helyzetű [h] hang a gégében keletkezik (*hal*). Ebben az esetben nem határozható meg rá akusztikai vetület, mivel képzési helye az artikulációs csatorna kezdeti pontján van, így nincs sajátos befolyása a következő hangra, inkább fordítva (*hű, hó*). Ilyen [h] hang egyes CC kapcsolatokban is megvalósul, amikor a [h] a CC kapcsolat második eleme (*népharag*). A gégében keletkezett [h] hang intenzitása igen kicsi, ez a hangzóssági sorrend utolsó hangja a magyarban (Olaszy 1989a, Gósy 2004b). A [x] hang artikulációja a veláris területhez köthető, a rést itt hozza létre a beszélő. A megelőző magánhangzó hatására lehet preveláris (*ihlet, technika)*. Ilyenkor ez a magánhangzó elöl képzett. A képzés veláris, ha a megelőző magánhangzó hátul képzett (*fachpolc, doh)*. Ennek a réshangnak a jellemző frekvenciakomponensei a 2000–3000 Hz-es sávban helyezkednek el.

A [ç] réshang (amit a [j] hang zöngétlen variánsaként tart számon az irodalom) általában hangsorzáró helyzetben jön létre, ha zöngétlen zárhang vagy réshang előzi meg (*lépj*), de hangsor belsejében is kialakulhat (*hívj ki*). A hang frekvenciaszerkezetének lényeges komponensei 3000 Hz környékén koncentrálódnak. (Az 5.24. ábrán

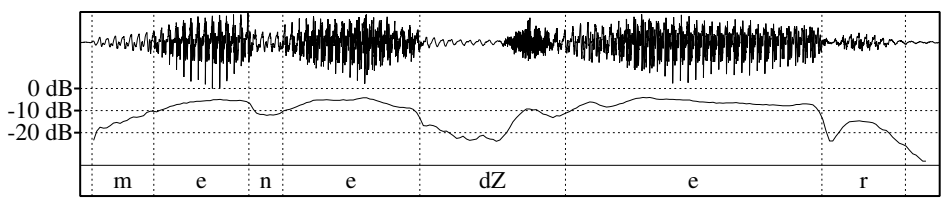
példát láthatunk a hang zöngés és zörejes megvalósulásában létrejövő jellemző frekvenciakomponensekre).



5.24. ábra. A [j] és a [ç] hangok a *lépjen* (800 ms) és a *lépj* (580 ms) szavakban. A nyilakkal jelzett részeket kell összehasonlítani

5.1.3.5. Zöngés zár-rés hangok

A magyarban két zöngés zár-rés hang van, a [d͡z] és a [d͡ʒ]. Önálló előfordulásuk ritka, inkább a [t͡s], [t͡ʃ] hangok zöngésedéséből keletkeznek (Siptár 2006c), intenzitásuk 10–20 dB-lel alacsonyabb, mint a magánhangzó környezeté. A zöngés zár-rés hangok a legbonyolultabb szerkezetű magyar beszédhangok, ugyanis zárszakasz, rövid idejű zárfeloldódás, majd réselem alakítja ki a hangot. Ugyanakkor a gerjesztés is két elemből áll, zöngéből és zörejből. A zárelem fojtott zöngéje hasonló szerkezetű, mint a [d] hangé, a réselem pedig ehhez kapcsolódik a folyamatos zöngeképzésre szuperponálódva. A réselemben a [z], illetve [ʒ] hanghoz hasonló kevert gerjesztésű hang jelenik meg (5.25. ábra).



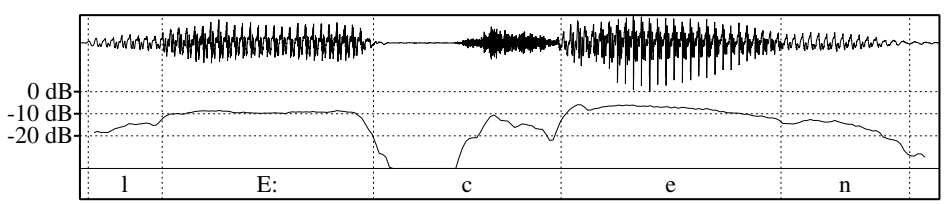
5.25. ábra. A [d͡ʒ] hang szerkezete a *menedzser* (825 ms) szóban

A réselem zörejének frekvenciaszerkezete minden szempontból hasonló az [s] és az [ʃ] szerkezetéhez, hiszen képzési helyük azonos, dentialveoláris, illetve alveoláris. Akusztikai vetület vonatkozásában ezekre a hangokra is ugyanaz vonatkozik, mint a

dentialveoláris, illetve alveoláris zöngés mássalhangzókra. A [d͡z d͡ʒ] hangok átlagos időtartama 100 ms körül mozog.

5.1.3.6. Zöngétlen zár-rés hangok

A magyarban két zöngétlen zár-rés hang van, a [t͡s] és a [t͡ʃ]. Artikulációs szempontból összetett szerkezetűek, néma fázisból, a zár feloldásából, majd zörejes réselemből állnak. A nyelvi megítélésük nem kiforrott, folyamatos vita tárgyát képezik mind az elméleti, mind a fonetikai nyelvészeti kutatások munkáiban (Hegedűs 1958, Szende 1975, Olaszy 1991b, Laver 1994, Kovács 2002). A vita témája annak eldöntése, hogy önálló hangnak kell-e tekinteni őket, vagy egy zárhang és egy réshang kapcsolatának (Siptár 2006b). Az tény, hogy akusztikailag a réselem frekvenciaszerkezete minden szempontból hasonló az [s] és [ʃ] szerkezetéhez, hiszen képzési helyük azonos, dentialveoláris, illetve alveoláris (5.18 ábra). A vizsgálatok szerint azonban a zár-rés hangok és a réshangok réselemének hossza között eltérés van, a valódi zöngétlen réshangok hosszabbak (Olaszy 2007b), mint a zár-rés hangból a réselem. Ebben a könyvben a fenti két hangot egyértelműen zár-rés hangként kezeljük. A zárszakasz néma fázisnak tekintendő (5.26. ábra), a rés hatására létrejövő turbulens áramlás frekvenciakomponenseit pedig a képzési hely befolyásolja. A zörejelem amplitúdója gyors növekedéssel alakul ki a zár feloldódásakor, intenzitása nő a hang közepéig, majd csökken a hang végéig. A zörej tehát a középső szakaszában a legerősebb, itt mintegy 10–15 dB-lel kisebb intenzitású, mint a hozzá csatlakozó magánhangzó szintje. A réselem zörejének frekvencia-összetevői egyértelműen jellemzik a zár-rés hangot. A [t͡s] hang rendelkezik a legmagasabb frekvenciakomponensekkel, hasonlóan a [s] hanghoz. Hangidőtartam szempontjából a zár-rés hangok a leghosszabbak közé tartoznak. Ezeknél a hangoknál is létezik a virtuális hangidőtartam, hasonlóan ahhoz, ahogy a zöngétlen zárhangoknál tárgyaltuk.

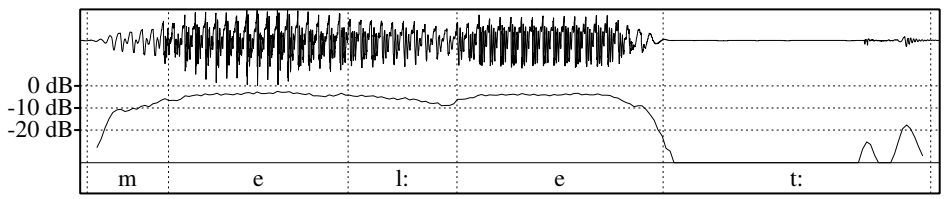


5.26. ábra. A [t͡s] hang szerkezete a *lécen* (790 ms) szóban. A hang mindkét eleme (zár+rés) jól megkülönböztethető az időfüggvényen (fent)

Akusztikai vetület vonatkozásában ugyanazok az értékek vonatkoznak erre a két hangra is, mint a dentialveoláris, illetve alveoláris zöngés mássalhangzókra.

5.1.3.7. Közelítő hangok

A magyar beszédben két olyan orális mássalhangzó van, amelyik a magánhangzókhoz hasonló akusztikai tulajdonságokkal rendelkezik, vagyis formánsaik vannak, intenzitása is nagy. Elnevezésük: közelítő hangok. Ezek az alveoláris képzési helyű [l] és a palatális képzési helyű [j]. Az [l] esetében a nyelv hegye az alveoláris területhez tapad, a nyelv két oldalán azonban szabadon áramlik a levegő. Ezért oldalréshangként is nevezték a korábbi szakirodalomban. Az [l] hang intenzitása hasonló, mint a magánhangzóké. VCV kapcsolatban az [l] határát csak spektrogram alapján lehet meghatározni, mivel az időfüggvényen összemosódhat a magánhangzó rezgésképével (5.27. ábra).



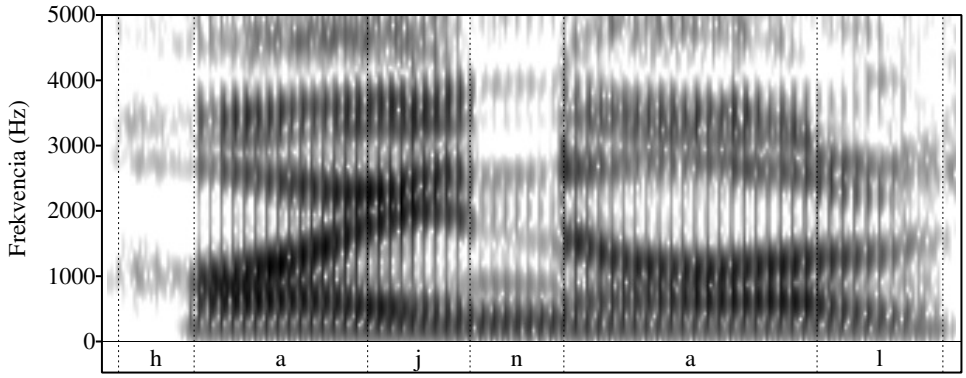
5.27. ábra. Az [l:] hang szerkezete a *mellett* (686 ms) szóban. A hang intenzitása (lent) hasonló, mint a magánhangzóké

Az [l]-re jellemző akusztikai vetület a következő: F1 = 400, F2 = 1300–1600, F3 = 2700 Hz. Az F2 értéke a csatlakozási ponton az [l]-hez csatlakozó hang akusztikai vetületében szereplő F2 értékétől függ. Ha ez alacsonyabb, mint 1300 Hz körüli, akkor az [l]-ben 1300 Hz lesz, ha magasabb, mint 1600 Hz körüli, akkor az [l]-ben 1600 Hz lesz, más esetekben megegyezik a két F2.

A [j] hang palatális zöngés hang, a nyelv az [i] magánhangzóhoz közel álló pozíciót vesz fel, a nyelv kissé közelebb van a kemény szájpadhoz, mint az [i]-nél, ezzel valósítjuk meg a rést. A mássalhangzók közül ez a legintenzívebb az egymáshoz viszonyított sorrendben, ez valósul meg a legnagyobb amplitúdóval, intenzitása hasonló a magánhangzókéhoz. A [j] akusztikai vetülete: F1 = 250, F2 =2000, F3 = 3000 Hz (5.28. ábra). A palatális hangok – mint már láttuk – akusztikai vetület szempontjából stabilnak mondhatók, tehát magukhoz idomítják az őket megelőző, illetve követő hangokat. A [j] is ezt teszi mind VCV, mind CC kapcsolatokban (5.28. ábra). A zöngétlen variánsát korábban tárgyaltuk.

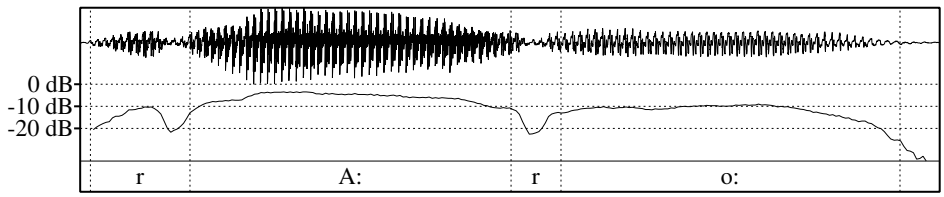
5.1.3.8. Pergőhang

A magyar [r] alveoláris zöngés pergőhang a legrövidebb beszédhangunk. Általánosságban egyetlen perdülettel valósul meg a beszédben. Az [r] hangzásának kialaku-



5.28. ábra. A [j] akusztikai vetületei a *hajnal* (612 ms) szóban. Látható, hogy a palatális mássalhangzó magához idomítja a környező hangok formánsait. Ez főleg az F2-k mozgásában látványos a példában

lásához két fizikai paraméter egyidejű, összehangolt működése szükséges. Az első az alveoláris képzési helyből adódó akusztikai vetület, ami az [r]-nek a formánsstruktúráját adja. A másik a nyelvhegy gyors, egyszeri közelítése az alveoláris területhez, majd hirtelen távolodása. Ez utóbbi mozgás a hangintenzitásra van lényeges hatással, ugyanis a nyelvhegy igen rövid időre elzárja a hangkiáramlás útját, így a hangintenzitás akár nagyon is lecsökkenhet, ekkor nincsenek formánsok a hangban. A pergetett [r] legfontosabb jellemzője tehát a hangintenzitás gyors csökkenése, ami egy intenzitásminimumban végződik legalább 10–15 ms-os időtartamban. Ilyenkor a hang 10–20 dB-lel alacsonyabb intenzitású, mint például az őt követő magánhangzó. Az alveoláris résztől való távolodással az intenzitás ismét elkezd növekedni. Ha nincs intenzitásminimum, akkor nem érzékeljük a pergő hangot. Az intenzitásminimum tehát a magyar, apikális [r] legfontosabb sajátossága. Nézzük meg részletesen a hangsorkezdő [r] szerkezetét is, mivel annak felépítése más, mint a hangsor belsejié. Az alapvető követelmény a #CV helyzetben sem változik, létre kell hozni a már említett intenzitásminimumot. Belátható, hogy ezt csak egy már hangzó hangból lehet megvalósítani. Az ellentmondás ott van, hogy hangsor eleji helyzetben vagyunk, és ilyenkor általában nulla intenzitásról indítjuk a hangokat, legyen az bármilyen hang. Nulla intenzitásról pedig nem lehet intenzitásminimumot létrehozni, ami az [r] alapkritériuma. A megoldást az artikulációs mechanizmusunk úgy oldja meg, hogy a #CV helyzetű [r] szerves része lesz egy rövid előke, egy svá jellegű indító hang (Olaszy 1985, Gósy 2006). #CV helyzetben tehát az [r] szerkezete három részből áll: a) indító svá (ebből alakítja ki a beszélő az intenzitásminimumot), b) az intenzitásminimum, c) az [r] hangnak a következő hanghoz kapcsolódó része. A hangsorkezdő svá elem rövid, hossza az ejtéstől függ, intenzitása 5–10 dB-lel kisebb, mint a csatlakozó magánhangzóé. Az [r]-nél ebben a hanghelyzetben tehát a svá a beszédhang szerves eleme (5.29. ábra). Az artikulációban a fonológiai [r:] gyakorlati megvalósí-



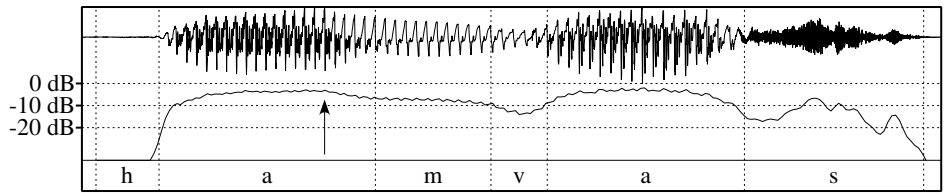
5.29. ábra. Az [r] hangok szerkezete a *Ráró* (851 ms) szóban. A #CV helyzetű [r]-nél a kezdő svá elem nyolc periódusnyi. Hangsor belseji helyzetben az [r] első része a V1 végén alakul ki, ezután következik az intenzitásminimum, majd a pergő hang befejező része következik, amely a V2-höz kapcsolódó zöngés periódusban valósul meg folyamatos átmenettel

tása is több formát mutat. Ezekről részletesen lásd Olaszy (2007b) munkáját. Az [r]-re jellemző akusztikai vetület: F1 = 450 Hz, F2 = 1500 Hz, F3 = 2700 Hz. Az [r]-hez csatlakozó hang akusztikai vetülete kismértékben befolyásolhatja az [r]-re jellemző F2 mozgását. Ha például palatális hang követi az [r]-t, akkor az F2 az 1500 Hz-es értékről felfelé fog elmozdulni, az F1 pedig csökkenni fog.

5.1.3.9. Nazális hangok

A magyar nazális hangok az [m], [n], [ɲ]. Ez a három hang fonémákat képvisel. A koartikulációs hatások következtében létrejönnek variánsok is. Az [m] variánsa a [ɱ] labiodentális nazális zárhang (*hamvas*). Az [n] variánsa a veláris nazális [ŋ] hang (*hangos*). Az [n]-nek létezik olyan formája is, amikor a nazális hang túlnyomórészt az őt megelőző magánhangzó nazalizációjában testesül meg. Ez akkor jön létre, amikor az [n]-t [h j s ʃ] mássalhangzó követi (*színház, fenség*). Ezekben az esetekben nem jön létre a zár a dentialveoláris területen, a nazális hang egybeolvad a megelőző magánhangzóval. Artikulációs szempontból a nazális hangok különlegesnek számítanak, hiszen az orális üreg mellé egy második hangképzési csatorna lép be párhuzamosan a beszédképzés folyamatába. Ennek sok esetben az a következménye, hogy a nazális hangok „megzavarják” egyes orális mássalhangzók képzési folyamatait, amelyekről a hangkapcsolódásoknál lesz szó. Fontos megjegyezni, hogy a nazális csatorna méretei nem változnak a beszédképzés során. Ezért a nazális csatorna hatásának vizsgálata egyszerűbb, mint az orálisé. A nazális mássalhangzók képzése a következő mozzanatokból áll: zár a szájüregben, ugyanakkor az orrüreget kinyitjuk a nyelvcsappal, ezen keresztül fog áramlani a zöngés hangelem. A beszédhang kialakításában tehát a szájüreg mint zárt rendszer, az orrüreg mint nyitott csatorna vesz részt. A keletkezett nazális hang mindhárom mássalhangzónál egyaránt tiszta, zörejmentes, zöngés periódusokból áll. A nazális hang akusztikai vetülete egységes, hiszen az orrüreg méreteit az artikuláció során nem tudjuk változtatni. A nazális hangok hatása kiterjed a magánhangzókra is, ezek is nazalizálódhatnak (5.30. ábra).

Ilyenkor a nazális üregrendszer bekapcsolódásának hatására a magánhangzó formánsainak struktúrája változhat (Horváth 2005). A nazális mássalhangzók intenzitása



5.30. ábra. Az [ɔ] magánhangzók szerkezete a *hamvas* (723 ms) szóban. Az első magánhangzó nazalizációja már elkezdődik a hang vége előtt. Ez látható is az ottani hangintenzitás csökkenésén (nyíl)

5–10 dB-lel alacsonyabb, mint a csatlakozó magánhangzóé. A képzési helyükre jellemző akusztikai vetület formánsadatait az orrüreg és a szájüreg együttesen alakítja ki. A jellemző értékeket az 5.25. táblázatban adjuk meg. A csatlakozó hang formán-

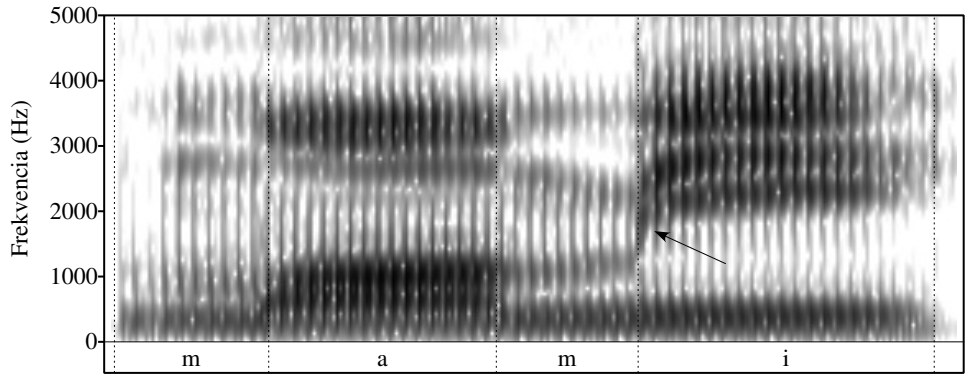
5.14. táblázat. A nazális mássalhangzók képzési helyének akusztikai vetülete a hozzájuk kapcsolódó magánhangzó függvényében a kapcsolódási pontra jellemző tájékoztató formánsadatokkal kifejezve. A magánhangzókat az E1-hangszimbólumokkal jelöltük. A számértékek Hz-ben értendők

Képzési hely	Formáns	A:	a	o	u	U	i	E:	O	e
Bilabiális	F1	250	250	250	250	250	300	250	250	250
	F2	1000	900	900	900	1400	1600	1300	1200	1200
	F3	2200	2200	2200	2000	2200	2400	2300	2200	2300
Alveoláris	F1	250	250	250	250	250	250	250	250	250
	F2	1350	1300	1300	1300	1500	1600	1500	1400	1400
	F3	2600	2600	2600	2500	2800	2900	2700	2700	2700
Palatális	F1	350	350	350	350	350	350	350	350	350
	F2	2100	2100	2100	2100	2100	2100	2100	2100	2100
	F2	3500	3500	3500	3500	3500	3500	3500	3500	3500

sai a táblázatban megadott formánsértékekhez közelítenek a kapcsolódási ponton. A megadott értékek tendenciákat fejeznek ki. Mivel az üregváltás gyors, a kapcsolódási ponton az F2-t illetően formánsugrás is létrejöhet, ha a nazális hang és a kapcsolódó hang akusztikai vetülete között nagy különbség van. Erre láthatunk példát a *mami* szó [i] hangjának elején (5.31. ábra). A VCV kapcsolatokban a nazális hang F2-je a megadott szűk sávban mozoghat a hangon belül, ha a közrefogó V-k akusztikai vetülete nagyon eltérő.

5.2. A hangkapcsolódások típusai és szerkezeti sajátosságaik

Az előző fejezetekben ismertetett beszédhangok csak önmagukban, izoláltan ejtve tartalmazzák a rájuk megállapított akusztikai jellemzőket. Ha két hangot összekap-



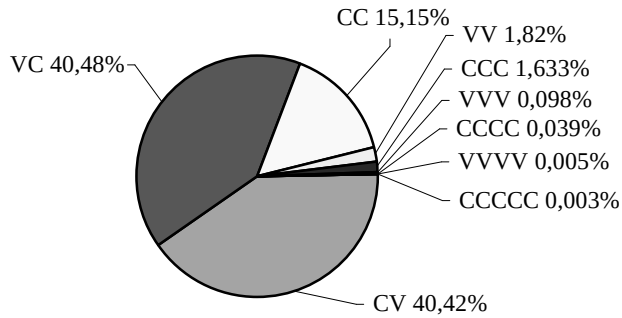
5.31. ábra. A *mami* (523 ms) szó frekvenciaszerkezete. Az[m]+[i] kapcsolódási ponton formánsugrás jöhet létre (nyíl)

csolva ejtünk (*só*), akkor a koartikuláció következtében létrejön egy köztes fázis, ami akusztikailag összekapcsolja őket (3.1.2.1. fejezet). Ezt a hangrészt hangátmenetnek nevezzük. A hangátmeneti rész tehát minden beszédhangban jelen van a folyamatos beszédben, ha két különböző hang követi egymást. A hangátmenet létrejötte a folyamatos artikulációból adódó akusztikai következmény, tehát koartikulációs hatás. Az artikulációs mozgás minden pillanatának (mondhatjuk úgy is, hogy filmkockájának) megfelel a rá jellemző akusztikai szerkezet, vagyis az akusztikai vetület. Az átmeneti fázisokban legtöbbször mindkét hang akusztikai szerkezete valamelyest megváltozik, közelít egymáshoz. A kérdés az, hogy ez milyen mértékű. Az átmeneti fázis akusztikai jellemzői minden hangkapcsolódásban más képet mutatnak, ez függ a nyelvtől, a beszélőtől és a beszédformától is. Spontán beszédben lazább artikuláció valósul meg, mint felolvasásban, ami a hangzásban is érzékelhető. Ezt a kiejtés modellezésében figyelembe is veszik (Lindblom 1990). A hangátmeneti szakaszok jellemzéséhez az akusztikai rezgésekre vonatkozó három fizikai paramétert kell leírnunk az átmeneti fázisokban is, vagyis a frekvenciaszerkezet változását, az időszerkezeti tényezőket és végül a hangkapcsolatra jellemző intenzitásfüggvényben történő módosulásokat. E három elem valamelyike (vagy kombinációjuk) megváltozhat a hangok kapcsolódási fázisában. Az átmeneti fázisra jellemző akusztikai szerkezeti változások függnek a két hangra jellemző artikulációs képzési helytől, a módtól és a gerjesztés fajtájától is (zöngés, zöngétlen, kevert gerjesztésű). Az átmeneti fázisnak lényeges eleme az időtényező. A két beszédhang közötti artikulációs mozgást, a koartikulációt a beszélőnek meg kell valósítania, ez időt vesz igénybe. Ha a koartikuláció nem igényel bonyolult mozgássort – azaz az egyes mozgások nem akadályozzák egymást –, akkor rövidebb idő alatt jöhet létre az átmenet, ha igen, akkor hosszabb idő alatt. Például az [ʃ] hang átlagos hossza a *kismadár* szóban 122 ms, míg az *áskálódik* szóban csak 95 ms (vö. Olaszy 2007a). A legegyszerűbb koartikulációs helyzetnek tekinthetjük, amikor ugyanazon hangok találkoznak

egymással. Ilyenkor többnyire csak az időszerkezetben találunk változást, a hangra jellemző időtartam (vagy időszerkezeti elem) megnyúlik (*rá**á**ll, legjo**bb** **b**arátom*). Ha a koartikuláció összetett mozgássort igényel, akkor egyrészről az átmeneti fázisra fordított idő lesz hosszabb, másrészről a kapcsolatban résztvevő hangok szerkezetében is olyan változások állhatnak be, amelyek jelzik, hogy bonyolult, nehezen megvalósítható egymásra hatásról van szó (töltelékhang jelenik meg, a zörejelem megnyúlik, azaz spirantizálódás jön létre stb.). Vegyük ezeket sorra. Egyszerű, egymást nem zavaró mozgások zajlanak le például a különböző magánhangzók találkozásánál (*k**ie**nged, rá**é**r, k**ő**omlás*). A magánhangzók közötti átmeneteknél a nyelv, az ajkak, az állkapocs mozgása folyamatos. Mindhárom változhat akadálytalanul, hiszen egymástól függetlenek, ezért az átmenet képzése sima, az akusztikai szerkezet folyamatosan változik a két magánhangzó között (lehetséges gégezárral is ejteni az ilyen kapcsolatokat, de most a folyamatos artikulációt vizsgáljuk). Látszólag bonyolultabb a helyzet a CV, VC kapcsolatoknál, mivel az ilyen hangkapcsolódásoknál a mássalhangzókra jellemző képzési helyet és módot, valamint a két hang közötti esetleges gerjesztésváltást is meg kell valósítani a hangkapcsolódásban. Az artikulációs mozgások azonban itt sem ütköznek, nem akadályozzák egymást. A két hang egymásrahatása csupán az átmeneti fázis súlyozásában valósul meg (kölcsönösen illeszkedik mindkét hang, vagy az egyik rákényszeríti a hatását a másikra). Ez a súlyozás azonban nem jelent torzulást egyik hang akusztikai szerkezetében sem, csupán némileg megváltoztathatja az azokra egyébként jellemző akusztikai paraméterek értékét (formánsok, zörejgócok mozognak). Belátható, hogy a legbonyolultabb koartikulációs helyzet a mássalhangzó-kapcsolódásoknál lép fel, amikor az egyik C képzési helyéből, módjából és gerjesztéséből kell a másik (esetleg a harmadik) kapcsolódó mássalhangzó képzésére jellemző artikulációs helyzetbe vezérelni a beszédszerveket (*a**br**ak, fe**lspr**iccel*). Ezért ezekre a hangkapcsolódásokra az jellemző, hogy a bennük részt vevő hangokban lényeges akusztikai változások jöhetnek létre a koartikuláció hatásaként. Ezeket később tárgyaljuk.

A hangsorok felépítése során (amikor beszélünk) alapvetően négyféle hangkapcsolódási forma jöhet létre: magánhangzóból magánhangzóba (VV), mássalhangzóból magánhangzóba (CV), magánhangzóból mássalhangzóba (VC) és mássalhangzóból mássalhangzóba (CC). A beszédben ezek további, egymással való kombinációi is lehetségesek (Szende 1976). Az egyes hangkapcsolódások részvételi aránya nyelvfüggő, a magyarra vonatkozó adatokat az 5.32 ábra mutatja. A hangsorépítésben tehát zömmel a CV és VC kapcsolatok dominálnak, a CC elemek is gyakran fordulnak elő. A többelemű magánhangzó és mássalhangzó-kapcsolatok csak kis mértékben fordulnak elő. Ennek ellenére ezek vizsgálata is fontos, hiszen szerves részei a hangsorépítésnek.

Amennyiben az abszolút hangsorkezdő, illetve hangsorzáró pozíciót is hangkapcsolatnak tekintjük (hangindítás, hanglezárás), akkor a következő előfordulási gyakoriságokat kapjuk ugyanarra a vizsgálati korpuszra: #C = 1,69%, #V = 1,42%,



5.32. ábra. A hangsorépítő elemek előfordulási gyakorisága a Magyar Nemzeti Szövegtár szövegei alapján

C# = 2,08% és V# = 1,02%. Ezek szerint mássalhangzóval gyakrabban kezdünk is és fejezünk is be közlést, mint magánhangzóval.

A hangátmeneti fázis és a teljes hang viszonya. A beszédhangsor minden beszédhangjában két átmeneti fázis jön létre (kivételt képez a hangsort indító és befejező hang, ahol csak egy). A kérdés az, hogy időszerkezeti szempontból hogyan alakul a beszédhang szerkezete. Elméletileg három fázist különböztethetünk meg: az átmeneti fázist a hang kezdeti szakaszában, a hangra jellemző úgynevezett tiszta fázist, és harmadikként a hang befejező szakaszára jellemző átmeneti fázist (a gyakorlatban ezek nemigen válnak szét). Ezt a felosztást nem minden beszédhangra tudjuk alkalmazni, például a zár- és zár-rés hangoknál nem lehet értelmezni. Percepciós szempontból az átmeneti fázisokat az adott hanghoz tartozónak tekintjük, függetlenül attól, hogy esetlegesen olyan akusztikai komponenseket is mérhetünk bennük, amelyek nem jellemzőek magára a hangra (például a *jó* hangkapcsolatban az [oː] hangban az F2 formáns felvehet akár 1800 Hz-es értéket is a mássalhangzóhoz való kapcsolódási fázisában (vö. a függelékben bemutatott hangkapcsolatokkal), noha a rá jellemző F2 érték 1000 Hz körül van. A magánhangzó formánsstruktúrája tehát lényegesen megváltozhat az átmeneti fázisban, ennek ellenére ugyanahhoz a magánhangzóhoz tartozónak tekintjük (vö. 4.5. fejezet). A beszédértési mechanizmusunk a hangátmenetekkel együtt észleli és különbözteti meg a beszédhangokat, sőt a hangátmeneti részek információt közölhetnek a szomszédos hangról. A beszédhangok egymásra hatásának fokozatai vannak és ezek hangfüggőek, ezekről később részletesen szólunk. Az egyes hangok megváltoztathatják a belső szerkezetüket is a hangkapcsolódás során a koartikuláció következtében. Ez főleg bizonyos mássalhangzók találkozásánál jön létre A beszédtechnológiai modellkészítéseknél fontos ismernünk a hangátmenetek jellemzőit.

5.2.1. Magánhangzó-magánhangzó kapcsolódások

A VV típusú hangkapcsolódások vizsgálata mind fonológiai, mind fonetikai szempontból ingoványos terület. Ha két magánhangzó találkozik, akkor fonológiai szempontból a hangűr (hiátus) jelenségével állunk szemben (Siptár 2002a). A magyarban a hiátus kitöltődhet egy [j]-hez hasonló hangelemmel (*fiú)* bizonyos magánhangzók találkozásánál (Siptár 2002b). Ennek a jelenségnek a vizsgálata beszédtechnológiai szempontból is fontos, az ezzel kapcsolatos legújabb mérési eredményeket az 5.2.1.1 fejezetben tárgyaljuk. A VV kapcsolatra fonetikai szempontból azt mondhatjuk, hogy a két magánhangzót összekötő artikulációs mozgások (nyelv, állkapocs, ajakállás) folyamatosan zajlanak le (semmilyen akadályozó körülmény nem lép fel), tehát folyamatos lesz a hangátmenet (a gégezárral történő ejtést itt nem tárgyaljuk). Fizikai szempontból ez azt jelenti, hogy mind a frekvenciaszerkezet (formánsok), mint a hangintenzitás változása folyamatos. Ilyen szempontból nézve tehát ez a legproblémamentesebb hangátmeneti forma, ami jól is modellezhető. A formánsmozgások annál nagyobbak a hangátmeneti részben, minél távolabb vannak a két csatlakozó hang formánsai egymástól. Az intenzitás módosulását csak a formánsok távolsági változása határozza meg. Az egyetlen vizsgálható paraméter az ilyen kapcsolatokban az időtényező, vagyis azt lehet vizsgálni, hogy mennyi az átmenet létrehozására fordított idő. További kérdés, hogy a kapcsolatban résztvevő hangok időtartama hogyan alakul egymáshoz viszonyítva, illetve a hangsorban máshol előforduló ugyanazon magánhangzókhoz hasonlítva. Ezek az időparaméterek erősen függhetnek az ejtéstől (egyéni jellemző), valamint nyelvi tényezőktől (hangsúly, hangkörnyezet, hanghelyzet). Az egyéni ejtésben például előfordulhat, hogy a beszélő különválasztja a két magánhangzót a hangsorban (gégezár), de az is, hogy folyamatos hangátmenettel hozza létre (*fa-ág / faág)* a hangkapcsolatot. Az előbbi gondos artikulációnál (például színészi beszéd) fordulhat elő.

Az első rendszerezett fonetikai vizsgálatot az ilyen hangkapcsolatokra vonatkozóan Menyhárt (2006) közölte. A vizsgálat anyagát 40-féle VV kapcsolat képezte, amelyeket rövid mondatokba ágyazott. A mondatokat 10 beszélő olvasta fel. A beszédanyagban az átlagos artikulációs sebesség 11,8 hang/s volt. A szerző a hangátmeneti osztályozásra három fő csoportot határozott meg: a) nem észlelhető hangátmenet (45,6%); b) van hangátmeneti fázis (13,2%); c) hiátustöltés van jelen a két magánhangzó között (34,6%). Az osztályozást auditív-vizuális (mérés hangspektrogramon) megítélés alapján alakította ki. Menyhárt szerint a VV kapcsolatok nagy részében nem észlelhető hangátmeneti rész. A leggyakoribb kapcsolatok ezek közül az [ɔaː] [aːɔ] [ɛɔ]. Észlelhető hangátmenetek csak kis számban fordultak elő, átlagos időtartamuk 14,9 ms volt (szórás 3,1 ms). A kis szórás azt mutatja, hogy az ilyen típusú hangátmenetek függetlenek a magánhangzó típusától, a rövid időtartam pedig azt, hogy a percepcióban valószínűleg nincs nagy szerepük. Gósy (2002) szükségtelen hangátmeneteknek nevezi ezeket, mivel a magánhangzók percepcióját a rövid

hangátmenetek jelenléte nem befolyásolja. A leggyakoribb ilyen hangátmenetek a vizsgált anyagban az [ɛu], [øaː], [øo] kapcsolatok voltak. A rövid magánhangzók átlagos hossza 73 ms, a hosszúaké 94 ms volt, függetlenül a kapcsolatban elfoglalt helyüktől.

Néhány VV kapcsolat időtartamát Olaszy (1994b) hasonlította össze CVC kapcsolatok magánhangzóinak időtartamaival. A nyelvi anyagot 50 rövid kijelentő mondat képezte, egyetlen férfi felolvasásában. Itt cél volt a VV kapcsolatokra vonatkozó modell kialakítása, tehát az algoritmizálhatóság. Az eredmények azt mutatták, hogy a VV kapcsolatban a hangok megnyúlnak, vagyis az adott VV kapcsolat összesített időtartama hosszabb (1,3-szoros), mint az azt felépítő hangok más, CVC kapcsolatokban mért hangidőtartamainak összege. A nyúlás mértéke nagyobb, ha hosszú magánhangzók találkoznak (*ráértek*). A mondat belsejében a nyúlás kisebb, mint a mondat első, illetve utolsó szavában. Az eredményeket beszédszintetizátort vezérlő algoritmusba építették (Olaszy et al. 2000a). A mai kutatást hatékonyan támogatják a nyílt, szabadon hozzáférhető beszédadatbázisok, ezeket felhasználva is végeztünk méréseket egy magyar szóadatbázison (8.4.1. fejezet) és mondatadatbázison (8.4.2. fejezet). A vizsgált V1V2 kapcsolatok egyes hangjainak időtartamait mértük meg. A szóadatbázisban 54-féle ilyen kapcsolatra kaptunk adatot egy férfi és egy női bemondó ejtéséből (összesen 108 hangkapcsolat). A mintaszavak 3–5 szótagot tartalmaztak (*ráöntötte, faágon, almaérés, adóalap, vágóüzem* stb.). A hangidőtartamok átlagait az 5.15. táblázat mutatja. Hasonló mérést végeztünk a mondatokat tartalma-

5.15. táblázat. A mért magánhangzó-kapcsolatok hangjainak átlagos időtartama ms-ban a szóadatbázisban a VV kapcsolatban elfoglalt hanghelyzet függvényében az E1-hangszimbólumokkal jelölve

V1	Átlag	V2	Átlag
A:	177	A:	185
a	100	a	119
o:	133	o	111
u	100	u	102
U	114	U	94
E:	143	E:	146
O:	146	O:	99
e	92	e	125

zó beszédadatbázisban (5.16. táblázat), amelynek artikulációs sebessége 13 hang/s (egy férfi bemondó ejtése). A hangidőtartamok a két felolvasási formát jól tükrözik. A szóolvasásban a gondos artikulációból adódóan hosszabbak a hangok, mint a folyamatos felolvasásban. A mondatadatbázis általános célú, nem arra készült, hogy VV kapcsolatokat vizsgáljanak, ez megmutatkozik abban is, hogy nem minden VV kapcsolatra találhatók benne adatok. Az itt mért adatok fejezik ki legjobban a folyamatos beszédben lezajló artikulációs történéseket. Ebben az adatbázisban már találhatók szóhatáron fellépő VV kapcsolatok is. A hangsúlyozási esetleges nyújtás, illetve az esetenkénti redukált ejtés miatt a mondatokban nagy a mért hangidőtartamok

5.16. táblázat. A mondatadatbázisban mért magánhangzó-kapcsolatok hangjainak átlagos időtartama a VV kapcsolatban elfoglalt hanghelyzet szerint ms-ban. Ahol nincs adat, olyan kapcsolat nem volt az adatbázisban

V1	Átlag	V2	Átlag
A:	95	A:	148
a	70	a	79
o	–	o	101
u	–	u	77
U	–	U	50
E:	97	E:	100
O	–	O	93
e	56	e	103

szórása, akár ugyanabban a hanghelyzetben is. Hangsúlytalan esetben az átlaghoz képest nagyon lerövidülhet a magánhangzó, amennyiben a V1, illetve V2 hangsúlyt kap, akkor megnyúlhat.

Nézzünk néhány példát a VV kapcsolat V1-helyzetű hangjára:

*. . . kifüggesztv**e a** . . .*	[ɛ]= 20 ms	[ɔ]= 70 ms
*. . . d**e a** múlt . . .*	[ɛ]= 29 ms	[ɔ]= 99 ms
*. . . felsegítén**e a** . . .*	[ɛ]= 67 ms	[ɔ]= 66 ms
*. . . vacsor**a e**lőtt . . .*	[ɔ]= 34 ms	[ɛ]= 110 ms
*. . . rózs**a u**tca. . .*	[ɔ]= 111 ms	[u]= 65 ms
*. . . tette hozz**á A**nna néni.*	[aː]= 69 ms	[ɔ]= 93 ms
*. . . szól r**á a**z anyja.*	[aː]= 134 ms	[ɔ]= 62 ms

Néhány példa a VV kapcsolat V2-helyzetű hangjára:

*. . . m**a e**ste. . .*	[ɔ]= 84 ms	[ɛ]= 68 ms
*. . . lopt**a e**l. . .*	[ɔ]= 54 ms	[ɛ]= 79 ms
*. . . h**a e**lmondod. . .*	[ɔ]= 88 ms	[ɛ]= 139 ms
*. . . h**a e**lromlik. . .*	[ɔ]= 76 ms	[ɛ]= 157 ms

A példákból látható, hogy a VV kapcsolatokban megvalósuló hangidőtartamok sok tényezőtől függnek. A mondatadatbázisban a rövid magánhangzókat tekintve a legrövidebb VV kapcsolatban részt vevő hang az [ɛɔ] kapcsolat [ɛ] hangja volt, 17 ms értékkel az alábbi mondatban.

*Úgy állítom össz**e a** mondatokat, hogy minden típus előforduljon.*

A kirívó rövidség a laza ejtésből is adódhat, az [ɛ] hang szinte nincs is jelen a hangsorban, a redukált ejtés miatt. Ugyanakkor a leghosszabb rövid magánhangzó is az [ɛ] volt, 157 ms az [ae] kapcsolatban az alábbi mondatban.

*Mit csinálsz, h**a e**lromlik a gép?*

Az [ɛɔ] és [ɔɛ] kapcsolatra pontosabb vizsgálatot is tudtunk végezni, mivel több előfordulása is volt ezeknek a kapcsolatoknak (5.17. táblázat). Az [ɛ] hang a számadatok szerint szignifikánsan rövidebb a V1 helyzetben, mint a V2-ben. A fenti mérések

5.17. táblázat. Az [ɛ] hang időadatai az [ɛɔ] és [ɔɛ] kapcsolatban

Hang	V1=e	V2=e
Átlag	36	106
Minimum	17	67
Maximum	67	157
Szórás	23	31
Darab	14	11

mindkét beszédadatbázisban megismételhetők. Megjegyezzük, hogy a VV kapcsolatok vizsgálatánál az is befolyásoló tényező lehet, hogy a kapcsolat hangjai morfémahatáron helyezkednek el, továbbá a kérdés nyelvjárásfüggő is. Ezekkel az esetekkel nem foglalkoztunk.

5.2.1.1. A hiátustöltés jelensége

A magyar nyelvben bizonyos VV kapcsolatok ejtésekor megjelenik a két magánhangzó között a hiátustöltő hang (Siptár 2002a,b), amelyik minden esetben a [j]-hez hasonló akusztikai szerkezettel rendelkezik. A *stúdió* szó kiejtése például [ʃtuːdijoː].

A hiátustöltő hangot helyesírásunk nem jelöli. A hiátustöltés elemzése és definiálása nehéz. Általában percepciós alapon döntik el, hogy van-e [j] hanghoz hasonló hangrész a két kapcsolódó magánhangzó között (5.2.1. fejezet), avagy nincs. A jelenség objektív vizsgálata a pontos fonetikai átíráshoz segít hozzá, ami a gépi beszédfelismerés, valamint a beszédszintézis szempontjából is fontos. Ezekben a technológiákban az akusztikai jel minél pontosabb elemzése, illetve megvalósítása a cél. A hiátustöltés jellemzően olyan magánhangzók között jelenik meg, amelyekben az F2 formánsok távolsága nagy. A leggyakoribb hiátustöltést generáló hang az [i]. Menyhárt (2006) mérésében az [i] a vizsgált hiátustöltéses kapcsolatok 64,2%-ában fordult elő (*dió*), ezt követte az [ɛ] 19,6%-kal (*tea*), majd az [y] következett 13,6%-kal (*menüért*).

Ebben a fejezetben csak az [i]+V és V+[i] kapcsolatok legújabb, számítógépes vizsgálatáról adunk meg adatokat (Olaszy 2010), valamint összehasonlítást végzünk korábbi vizsgálatokkal.

Az artikulációs mozgások VV kapcsolatban. Ha két magánhangzó találkozik, akkor az artikulációs mozgások a hangokra jellemző nyelvállások és ajakartikulációk között folyamatosan zajlanak le (átmeneti fázist hozunk létre). Ez a mozgás kényszermozgás, mindenképpen mozgatni kell a nyelvet is és az ajkakat is az egyik pozícióból a másikba. Az átmeneti fázis időtartama mondja meg, hogy a feladat végrehajtása mennyire bonyolult az artikulációs szervek részéről. A hiátustöltés esetében tehát a töltelékhang időtartamát érdemes meghatározni. Az [i] és a [j] artikulációja között az a különbség, hogy az [i]-nél a nyelv távolabb van a kemény szájpadtól, mint a [j] esetében. A V+[i] kapcsolatban – ha van hiátustöltés – irány szerint az [i] felé artiku-

lálunk a V-ből, de tovább emeljük a nyelvet a [j] pozíciójáig (*mindmáig*), majd onnan engedjük vissza az [i]-re jellemző magasságra. Ez pluszráfordítás. Ha [i]+V kapcsolatról van ugyanígy szó, akkor az [i]-ből először felfelé emeljük a nyelvet, majd csak utána vesszük az irányt a magánhangzóra jellemző pozíció felé. Tehát itt is plusz-energiát fektetünk be. A kérdés tehát az, hogy mikor döntünk úgy, hogy alkalmazzuk ezt a plusz-energiaráfordítást.

A célkitűzésünk beszédtechnológiai jellegű. Az próbáljuk feltárni, hogy mely esetekben jön létre hiátustöltés a vizsgált hangkapcsolatokban. Erre szabályokat alkotunk. Törekszünk továbbá az összehasonlításra is, vagyis ugyanazon hangkörnyezetre megállapítjuk, hogy van-e fizikailag mérhető szignifikáns különbség a következő három eset között: a két magánhangzó közötti hiátustöltés van (*fiának*), illetve valódi [j] hang van közöttük (*kijárat*, illetve nincs hiátustöltés közöttük csak hangátmenet (*kiáltás)*. Az eredményeket magyar nyelvű szöveg-hang átíró algoritmusnál is használni kívánjuk.

A módszert három tényező köré építjük. Vizsgáljuk a hiátustöltés hangkörnyezetét, mérjük a töltelékhang időtartamát és felhasználjuk a korábbi kutatásokból származó eredményeket. A hanghosszúság mérésére többtényezős módszert alkalmazunk. A hiátustöltés fizikai azonosítása nehéz, hiszen az [i]-t követően kell a [j]-szerű hangot azonosítani. Hangspektrogramon ez nehéz, auditív értékelés is kell. Ez utóbbit alkalmaztuk, méghozzá rugalmasan változtatható hangablakozással jobbról is és balról is közelítve, az ablakot szélesítve, keskenyítve hallgattuk a jelet (Olaszy 2001a). Ezeken felül bevezettünk egy új módszert is. Úgy döntöttünk, hogy a fent kifejtett artikulációs szempontok figyelembevételével célzott kiejtési vizsgálattal is megpróbáljuk támogatni a döntést. Ennek lényege a következő. A hiátustöltés meglétét úgy ellenőrizzük, hogy az ejtés során tudatosan beékeljük a [j] hangot a két magánhangzó közé. Ha könnyedén sikerül kiejteni (például: *kijabál*), és nem érződik az artikulációnkban, hogy erőltetett lenne, akkor lehet ott hiátustöltés, ha nem, akkor inkább nem jellemző (például: *kijáltás, menüji, alájírás*). A vizsgálat során a korábbi kutatások azon feltételezését, hogy morfémahatáron nem jellemző a hiátustöltés (*kiállítás, kienged, férfiarc, beleilleszt, almaillat*) átvettük és alkalmaztuk. A hiátustöltés hang- és betűjelölésére a j+ jelet használjuk a következőkben. Ezzel azt akarjuk kifejezni, hogy ez a hang nem egyezik meg a [j] hanggal.

A vizsgálatokhoz két beszédadatbázist is felhasználtunk egy régebbit és egy újat, amit kifejezetten erre a vizsgálatra terveztünk A régebbi a 8.4.1. fejezetben ismertetett szólista-adatbázis, az újabb szintén szólistából áll továbbá egyetlen mondatból. Ez utóbbiban célzott hangszerkezetű szavakat állítottunk össze (256 szó). A vizsgálati mondat a következő: *A fiának a kiáltását hallotta a kijárati ajtó mellett.*. Az anyagot 12 személlyel felolvastattuk, ez a hanganyag képezte a vizsgálati teret.

Az időtartammérések eredményei azt mutatják, hogy a hiátustöltés időtartama széles skálán mozog a hangkapcsolatot felépítő magánhangzók függvényében (5.18. táblázat). A mérésből kapott teljes értékű [j] hang időtartamátlaga 57,4 ms volt. Lát-

5.18. táblázat. A V+[i] és az [i]+V betűkapcsolatok kiejtésénél megjelenő hiátustöltések időtartamátlagai a mért hangkapcsolatokban.

Betűkapcsolat	ái	ai	ói	ui	üi	éi	öi	ei	j+ átlag ms
A hiátustöltés időtartama ms	53,5	42,5	34,5	24	35	34	40,25	38,2	37,7
Betűkapcsolat	iá	ia	io	Ió:	iu	iü		ie	j+ átlag ms
A hiátustöltés időtartama ms	37,2	37,4	36,1	44,7	45,4			38,5	39,8

ható, hogy a hiátustöltés rövidebb, mint a [j] hang. Ez lényeges akusztikai különbség.

A mérések eredményeiből hat szabályt alakítottunk ki a hiátustöltés algoritmizálhatóságának meghatározására két-, három- és négyelemű magánhangzó-kapcsolatokra.

a) Kételemű kapcsolatoknál nem jellemző a j+, ha a V1-et [j] előzi meg (*maszkjai, tokajiak, napjáig, tetejéig, tagjainak*). Ennek valószínű magyarázata a gazdaságos artikulációra való törekvés. A helyesírásban előírt hang kap előnyt, azt mindenképpen ki kell ejteni.

b) Kételemű kapcsolatoknál kis valószínűséggel valósul meg a j+, ha a V2-t [j] hang követi. Mivel a palatális mássalhangzót mindenképpen ki kell ejteni, ezért az őt megelőző hiátustöltés létrejötte opcionlális, meg is maradhat (*mánij+ája, plébánij+ája, szérij+ája, kémij+áját*), de el is maradhat (*parókiája, koncepciójáról, koncentrációja*). A hiátustöltés megvalósulását valószínűleg a képzési helyek, illetve az orális-nazális váltás is befolyásolják.

c) Az [øː]+ [i] kapcsolatokban a hiátustöltés megvalósulása változó képet mutat mivel a két magánhangó artikulációja nagyon közel áll egymáshoz, a nyelvállás ugyanaz, csak az ajakállás változik az átmenetben (*erdőidet, küllőire*).

d) Hármas magánhangzó-kapcsolatokban, ha nem [j] előzi meg közvetlenül a hármas magánhangzó-kapcsolatot, a hiátustöltés általában az első két hang közé ékelődik be (*unokáj+ié, nőj+iesen, kémij+ai, mij+eink*). Ha [j] előzi meg, akkor eggyel jobbra tolódik a hiátustöltés megvalósulása, tehát a második és harmadik magánhangzó között jön létre (*filmjeij+ért, darabjaij+ért, forintjaij+ért*).

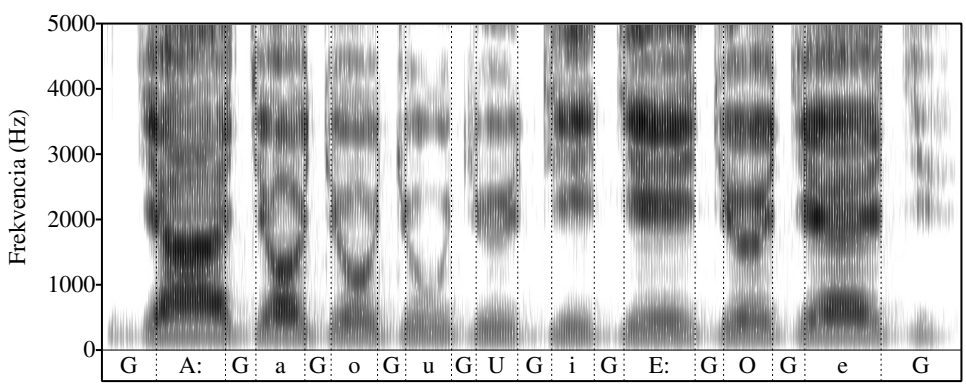
e) Négy magánhangzó kapcsolódása esetén két j+ nem lehet közvetlen egymás utáni szótagban. Ez azt jelenti, hogy az első két magánhangzó és az utolsó két magánhangzó közé ékelődhet a hiátustöltés (*biblij+aij+ak, itálij+aij+é*). Ezekre a kapcsolatokra is értelemszerűen vonatkozik a d) szabály.

f) Nem jön létre hiátustöltés, holott meg kellene valósulnia (*kiáltás, március, június, július, protein, ateista*). Az ilyen eseteket szabállyal nem tudjuk magyarázni, ezek nyelvi kivételként kezelendők.

A hiátustöltéssel kapcsolatos fenti szabályokat a 2010-ben nyilvánossá tett magyar szavak kiejtési szótárában (8.4.3. fejezet) már alkalmazták.

5.2.2. Mássalhangzó-magánhangzó-mássalhangzó kapcsolódások

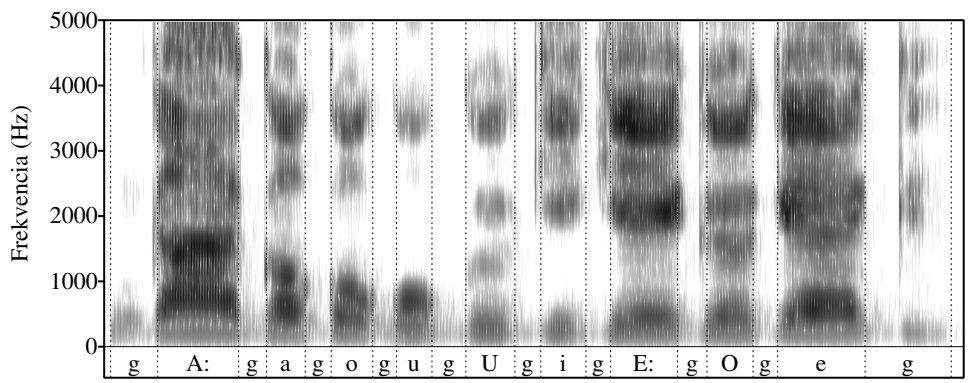
A CV és VC típusú hangkapcsolódásokat együtt tárgyaljuk a CVC kapcsolódásokban. A kapcsolódások akusztikai jellemzésénél a csatlakozó hangok három fizikai paraméterének esetleges változását kell leírni a kapcsolódási ponton és annak közvetlen közelében, vagyis a hangátmeneti fázisban. Mivel itt mássalhangzó és magánhangzó találkozik, bonyolultabb artikulációs mozgássorozatok zajlanak le, mint a VV kapcsolatokban. A frekvenciaszerkezeti változások tekintetében a C-re és a V-re korábban megadott akusztikai vetületek adatai a mérvadóak. Ezekből kiolvasható, hogy ugyanazon hang adott frekvenciakomponensére több jellemző Hz-érték is vonatkozhat (formánssáv), attól függően, hogy milyen a hozzá csatlakozó másik hang akusztikai vetülete. A CV, VC kapcsolatokban a frekvenciailleszkedési hajlandóság szempontjából megkülönböztetünk illeszkedő és nem illeszkedő hangokat (ezeknek koartikulációs okai vannak). Az illeszkedő idomul a hozzá csatlakozó hanghoz, a nem illeszkedő pedig nem, inkább saját magához idomítja azt. Ez utóbbi hangok frekvenciaszerkezete közel állandó, függetlenül attól, hogy milyen hang kapcsolódik hozzá. A magánhangzók az illeszkedő kategóriába tartoznak, a formánsaik elmozdulhatnak az átmeneti fázisban. A mássalhangzók közül nem illeszkedő típusok a palatális hangok, erősen illeszkedők a velárisok, a többi az illeszkedő kategóriába sorolható. A CVC kapcsolatokban végbemenő frekvenciailleszkedések spektrális képei minden hangra megtalálhatók a függelékben. A palatális mássalhangzók akusztikai vetülete stabil frekvenciákkal jellemezhető (F2 = 2000 Hz és környéke), és minden hozzájuk csatlakozó magánhangzó frekvenciaszerkezetét magukhoz idomítják (5.33. ábra), sőt még a mássalhangzókét is (lásd a CC kapcsolatoknál). A példában látható, hogy igen erőteljes formánsmozgások zajlanak le azokban a kapcsolatokban, ahol a V F2-je távol esik a 2000 Hz-es értéktől. Ugyanakkor azoknál az eseteknél, ahol



5.33. ábra. A palatális zárhang kapcsolódása a magánhangzókhoz. A frekvenciailleszkedés során csak a V-k formánsai illeszkednek a 2000 Hz körüli frekvenciaponthoz

nincs ilyen különbség, formánsmozgás gyakorlatilag nincs ([y i eː ɛ]). A formánsok mozgása ilyen képet mutat bármely palatális mássalhangzó esetében függetlenül a gerjesztéstől és a képzés módjától. A palatális mássalhangzó hatással van a magánhangzóra jellemző tiszta fázisú formánsértékek kialakításában is. Ahol nagymértékű F2 mozgás van, ott a magánhangzóra jellemző tiszta fázisú formánsértékek felfelé tolódnak (az akusztikai vetület változik). Például az [o] F2-je feljebb tolódik palatális hang utáni helyzetben, mint a [ko] kapcsolatban. Az előbbire számadatokat is megadtunk az 5.1.2, fejezet 5.6. táblázatában.

A veláris mássalhangzók erősen illeszkednek a magánhangzókhoz. Ez azt jelenti, hogy frekvencia-összetevőik a hozzájuk csatlakozó magánhangzó jellemző akusztikai vetületéhez idomulnak (a magánhangzóban számottevő formánsmozgás nem lesz) (5.34. ábra). A veláris területen tehát más-más helyen képződik a zár a magánhangzó nyelvállásától függően (*gu, gü, gi*). Amennyiben hátsó nyelvállású a magánhangzó, a [g] artikulációja is hátra tolódik a veláris területen, és fordítva. A [g]-hez



5.34. ábra. A veláris zöngés zárhang zárfelpattanásának frekvenciakomponensei erősen illeszkednek a magánhangzó F2-jéhez. Ez a mássalhangzó befolyásolja legkevésbé a V-k frekvenciaszerkezetét

kapcsolódó magánhangzókban minimális formánsmozgás látható, mivel a zárhang zárfelpattanásának frekvenciaértéke minden esetben közel van a V F2-jéhez, igazodik ahhoz. Ezt ki lehet használni a beszédszintézishez tervezett hangelembázisok készítésénél (lásd a 8.2.2. fejezetben).

A többi mássalhangzó a kölcsönösen illeszkedő kategóriába sorolható, azaz az illeszkedés mértéke megosztódik a két hang között. Például a zöngétlen rés- és zár-rés hangok alsó frekvenciakomponenseit az ajakartikuláció befolyásolja, ezért mélyebb zörejkomponenseket (tónust) hallunk az ajakkerekítéses magánhangzóknál (*sú, só, szú*), magasabbat az ajakréseseknél (*sí,, szí, szé*). A CVC kapcsolatokban a magánhangzóban lezajló formánsmozgásokból tehát következtetni lehet a mássalhangzóra, illetve egyes mássalhangzók jellemző frekvenciaértéke utalhat arra, hogy melyik magánhangzó következik utánuk a hangsorban. A hangátmeneti frekvenciaszerkezet

tehát előrejelző funkciót is betölt, ezt az emberi percepciós rendszer fel is használja a beszédértésnél. A függelékben megadjuk a magyar CVC kapcsolatok magánhangzójában létrejövő jellemző formánsmozgásokat a C képzési helyének függvényében minden magánhangzóra, férfi és női ejtésre a korábban tárgyalt formáns adatbázisból származtatott adatok alapján. Ezek a formánsmozgási tendenciák jellemzőek a hangkapcsolódásra, függetlenül a beszélőtől.

Szólnunk kell még a hangkapcsolódás intenzitásviszonyairól is. Itt a specifikus intenzitások a mérvadóak. Ahol olyan hangok találkoznak, amelyek között nagy különbség van, ott hirtelen intenzitásváltozás lesz jellemző a kapcsolódási pont környékére. Hol lehet kicsi a hangintenzitás? Egyrészről a zöngés-zöngétlen gerjesztésváltási pontokon (*sas, kék*). Ennek oka, hogy a gégében a fúvó állásból a zönge állásba kell váltani, ami azzal jár, hogy a hangszalagokat összezárjuk. A zárás pillanatában semmiféle levegőkiáramlás nem lesz a gége szintjén, intenzitásminimum keletkezik. A CV kapcsolatban az intenzitásminimum után hirtelen elkezd növekedni a V intenzitása. Ellenkező esetben, VC kapcsolatban az intenzitás hirtelen fog csökkenni a minimumra. Mindkét változásból következtethetünk a hanghatár pontos helyére is.

Amennyiben a C gerjesztése zöngés, akkor a képzési módtól függően két állapot alakulhat ki, A zár- és zár-rés hangoknál a fojtott zönge szakaszhoz csatlakozó V-nél szintén fokozatosan csökkenni fog az intenzitás, és a minimumot a fojtott zönge indulásánál éri el (*eb, ág*). Ugyanilyen intenzitásminimum keletkezik a pergő hangban is, amikor magánhangzóhoz kapcsolódik (*ára*). Az [r] hangnak ez a legjellemzőbb tulajdonsága (lásd az 5.1.3. fejezetben). A többi zöngés mássalhangzó esetében a két hangra jellemző specifikus intenzitás értéke fog egymáshoz kapcsolódni folyamatos átmenettel (*lé, el, nő, ön, jó, hű*).

Végül a hangkapcsolódásra jellemző átmeneti fázis időviszonyait elemezzük. A kérdés az, hogy mennyi idő szükséges a CV és VC hangátmenet megvalósításához. A válasz, hogy annál hosszabb, minél távolabbi artikulációs konfigurációt kell áthidalni. Ez kihat a beszédhangok hosszára is (specifikus időtartamok). A hangátmenetre fordított időt befolyásolja az artikulációs sebesség is. Gyors beszédűeknél kevesebb idő marad a hangátmenet megvalósítására. A magyarra jellemző 13 hang/sos artikulációs sebességnél a CV hangátmenetre fordított jellemző érték a V-ben a hang teljes hosszának 30–40%-a a rövid hangokban mérve. A VC átmenetre is hasonló érték adható meg (Olaszy 1991b). Ebből az is látszik, hogy a magánhangzók esetén a hang nagy része csak átmeneti fázisokból áll. A fonológiailag hosszú magánhangzóknál a hang tiszta fázisára hosszabb idő juthat (ejtéstől függ), ezekben az esetekben az átmeneti fázis a hang teljes hosszának 20–25%-a, amennyiben a beszélő tényleg hosszan ejti a hangot. A magyar CV és VC hangkapcsolódások részletes tárgyalása Olaszy (1985, 1989a) munkáiban található meg.

5.2.3. Mássalhangzó-mássalhangzó kapcsolódások

A mássalhangzók kapcsolódásánál arra vagyunk kíváncsiak, hogy a kapcsolódó hangok hogyan hatnak egymásra, mely esetekben történnek változások a kapcsolódásban résztvevő hangok akusztikai szerkezetében. A mássalhangzó-kapcsolódásoknál az egyik C képzési helyéből, módjából és gerjesztéséből kell a másik (esetleg a harmadik) kapcsolódó C képzésére jellemző artikulációs helyzetbe vezérelni a beszédszerveket. Ez bizonyos esetekben a mássalhangzó-kapcsolatra jellemző akusztikai módosulásokkal is jár. A következőkben a kettős, hármas és négyes mássalhangzó-kapcsolódásokban létrejövő koartikulációs jellemzőket tárgyaljuk, valamint azok hatását a kapcsolódó hangok akusztikai szerkezetére. A mássalhangzó-kapcsolódások részletes tárgyalása Olaszy (2007b) munkájában található. E fejezetben is abból vettünk át számadatokat. A közölt időadatok 10,5 hang/s-os beszédsebességre vonatkoznak. A vizsgálatok alapját képező szóadatbázis közvetlenül is tanulmányozható a `http://fonetika.nytud.hu/cccc` honlapon. A mássalhangzó-kapcsolódások elemzésének jobb megértéséhez előrevetítünk néhány fogalmat.

A koartikulációs néma fázis. Egyes mássalhangzó-kapcsolódásban mérhető egy a némafázisszerű hangelem a két hang kapcsolódási határán, amelyet koartikulációs néma fázisnak nevezünk (Olaszy 2006a, 2007b). Ez jellemzően a zöngétlen rés-, zárrés hangok réseleme és a nazális mássalhangzó közé ékelődik (*Sza**tym**az, ki**sny**úl*). A koartikulációs néma fázis mint hangelem nem kapcsolható a tradicionális hangleírások (zárhang, zár-rés hang stb.) egyetlen belső szerkezeti eleméhez sem, általában a zöngétlen mássalhangzóhoz tartozónak vallják (például a hanghatár kijelölésekor). Vizsgálata fontos a beszéd akusztikai szerkezetének precíz leírása, továbbá a beszédtechnológiai alkalmazások szempontjából. Részletesebben az 5.2.3.4 fejezetben tárgyaljuk.

Zöngekezdési idő CC kapcsolatokban. Magyarra eddig nem adtak meg VOT adatokat CC kapcsolatokra. Vizsgálatainkból ilyeneket is megadunk és azokat VOT-CC-vel jelöljük

Akusztikai vetület CC kapcsolatokban. A korábban bevezetett akusztikai vetület fogalmát értelmezzük a mássalhangzókra (pontosabban azok képzési helyére) CC kapcsolatokban is. A mássalhangzók közötti hangkapcsolódási pontokat az adott képzési helyekre jellemző akusztikai vetülettel hozzuk kapcsolatba (5.19. táblázat). Ha a kap-

5.19. táblázat. A magyar orális mássalhangzók képzési helyeinek akusztikai vetületei CC kapcsolatokra, tájékoztató formánsadatokkal kifejezve. A számértékek Hz-ben értendők. A megadott értékek tendenciákat fejeznek ki, hiszen a beszédben az akusztikai eredmény függ a beszélő fiziológiai és mentális adottságaitól is.

Mérés	Formáns	Bilabiális, Labiodentális	Dentialveoláris, Alveoláris	Palatális	Veláris
Akusztikai vetület	F1	450	450	450	450
	F2	1300	1700	2000	1400
	F3	2500	2800	2800	2500

csolódó hangok akusztikai vetületei különböznek, akkor frekvenciamozgás jön létre a kapcsolódási pont környékén. A mássalhangzó-kapcsolódások akusztikai szerkezetének részletes vizsgálatához ad segítséget a `http://fonetika.nytud.hu/cccc` honlap, ahol az összes ilyen magyar kapcsolatra találunk akusztikai példát és diagramokat, valamint rövid jellemzést.

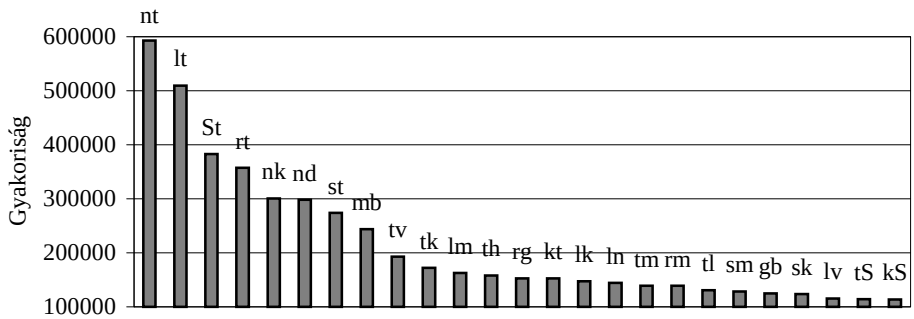
5.2.3.1. Kettős mássalhangzó-kapcsolódások

A CC kapcsolatokat úgy tekintjük, mint a mássalhangzó-kapcsolódások alapváltozatát. A magyarban kiejthető CC kapcsolatok száma 373 (5.20. táblázat). A CC

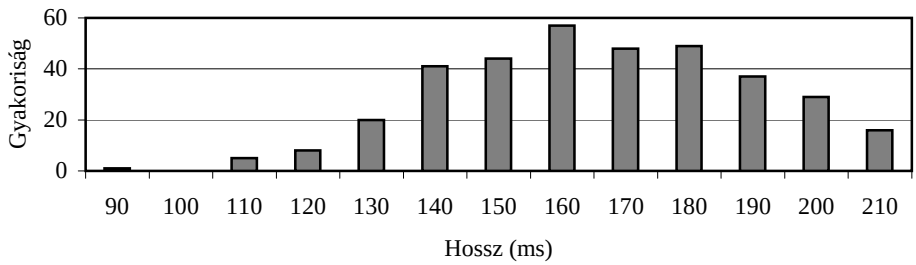
5.20. táblázat. A magyar beszédben előforduló CC hangkapcsolatok. Az első oszlop a C1-et, az első sor a C2-t jelenti. A hangokat az E1-jelű számítógépes hangszimbólumokkal adjuk meg. Jelmagyarázat: + = létező hangkapcsolat; +h = zöngésségi hasonulás van a kapcsolatban; - nincs ilyen kapcsolat

CC	b	d	G	g	p	t	T	k	m	n	N	j	h	v	f	z	s	Z	S	dz	dZ	c	C	l	r
b	-	+h	+h	+h	-	-	-	-	+	+	+	+	-	+	-	+h	-	+h	-	-	-	-	-	+	+
d	+h	-	-	+h	-	-	-	-	+	+	+	+	-	+	-	-	-	-	-	-	-	-	-	+	+
G	+	+	-	+	-	-	-	-	+	+	+	+	-	+	-	+	-	+	-	-	-	-	-	+	+
g	+h	+h	+h	-	-	-	-	-	+	+	+	+	-	+	-	+h	-	+h	-	-	-	-	-	+	+
p	-	-	-	-	-	+h	+h	+h	+	+	+	+	+h	+	+h	-	+h	-	+h	-	-	+h	+h	+	+
t	-	-	-	-	+h	-	-	+h	+	+	+	+	+h	+	+h	-	+h	-	+h	-	-	-	-	+	+
T	-	-	-	-	+h	+h	-	+h	+	+	-	-	+h	+	+h	-	+h	-	+h	-	-	+h	-	+	+
k	-	-	-	-	+h	+h	+	-	+	+	+	+	+h	+	+h	-	+h	-	+h	-	-	+h	+h	+	+
m	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+
n	-	+	-	+	-	+	-	+	-	-	-	+	+	-	-	+	+	+	+	-	-	+	+	+	+
N	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	-	-	+	+	+	+
j	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	-	-	+	+	+	+
h	-	-	-	-	+	+	+	+	+	+	+	+	-	-	+	-	+	-	+	-	-	+	+	+	+
v	+h	+h	+h	+h	-	-	-	-	+	+	+	+	-	-	-	+h	-	+h	-	-	-	-	-	+	+
f	-	-	-	-	+h	+h	+	+h	+	+	+	+	+h	+	-	-	+h	-	+h	-	-	+h	+h	+	+
z	+h	+h	+h	+h	-	-	-	-	+	+	+	+	-	+	-	-	-	-	-	-	-	-	-	+	+
s	-	-	-	-	+h	+h	+	+h	+	+	+	+	+h	+	+h	-	-	-	-	-	-	+h	+h	+	+
Z	+h	+h	+h	+h	-	-	-	-	+	+	+	-	-	+	-	-	-	-	-	-	-	-	-	+	+
S	-	-	-	-	+h	+h	+h	+h	+	+	+	+	+h	+	+h	-	+	-	-	-	-	+h	+h	+	+
dz	+	+	+	+	-	-	-	-	-	+	-	-	-	+	-	-	-	+	-	-	-	-	-	+	-
dZ	+	+	+	+	-	-	-	-	+	+	+	+	-	+	-	-	-	-	-	-	-	-	-	+	+
c	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-	-	-	+	+	+
C	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-	-	+		+	+
l	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	-	-	+	+	-	+
r	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-

kapcsolatok leggyakoribb elemeinek gyakorisági grafikonját az első 25 kapcsolatra a 5.35. ábrán mutatjuk be. A listát 2 millió szót tartalmazó szövegtestből származtattuk. A teljes gyakorisági listát a függelékben adjuk közre. A CC kapcsolatok kiejtésének átlagos teljes időtartama 162,5 ms (5.36. ábra), az eloszlás két szélső értékénél (80 ms, illetve 220 ms) csak kis számban fordulnak elő. A legrövidebb



5.35. ábra. A leggyakoribb magyar CC kapcsolatok előfordulása



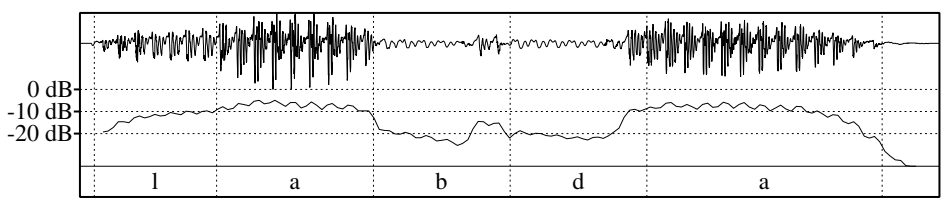
5.36. ábra. A CC mássalhangzó-kapcsolódások teljes időtartamainak eloszlása a mért szóadatbázisban 10,5 hang/s artikulációs sebességre (Olaszy 2007b)

időtartamú CC kapcsolatoknál azt látjuk, hogy a kapcsolat mindkét eleme az átlagok hierarchiájában is rövidnek számít, ilyenek az [l]+[r], az [r]+[l], az [n]+[l] kapcsolatok. Ezekben a kapcsolatokban a rövid összidőtartam kialakulásában két tényező vesz részt, az azonos vagy közeli képzési hely, valamint az, hogy a kapcsolatot felépítő hangok hol helyezkedik el az általános hanghosszúsági sorrendben. A magyarban az [r], [l] a legrövidebb hangok közé tartozik (Magdics 1966, Olaszy 2006a), átlagosan a leghosszabbak pedig a zöngétlen rés- és zár-rés hangok. A leghosszabb összidőtartamot azokra a CC kapcsolatokra mértük, amelyeknél az előbbi két tényező mindegyike hosszú időtartamot ad önmagában is. Ilyenek a [t͡s], [t͡ʃ], [t], [s], [ʃ] hangokból felépülő CC kapcsolatok (*mala**cs**ág, vála**szcs**apás, ő**scs**ótány*).

Zöngés zárhangok CC kapcsolatokban. A zöngés zárhangok akusztikai szerkezete mássalhangzó-kapcsolatokban, bizonyos hangokhoz való kapcsolódásukkor és bizonyos hanghelyzetekben egy svá töltelékelemmel egészül ki. A svá töltelékelem jelenléte a CC kapcsolatban azt fejezi ki, hogy a kapcsolatban megvalósítandó artikulációs változásokhoz nagy artikulációs energia szükséges, hiszen a hang-

kapcsolódás csak egy betoldott zöngés hangelem, a svá közbeiktatásával hozható létre. A svá általában a következő formánsszerkezettel rendelkezik: F1 = 500 Hz, F2 = 1500 Hz, F3 = 2500 Hz. Ezt a kapcsolódó mássalhangzó akusztikai vetülete megváltoztathatja. Három jelölésformát használunk a CC kapcsolatokban előforduló svá erősségének jellemzésére, továbbá jelölünk egy palatalizált képzési helyhez illeszkedő koartikulációból adódó változatot. A svá+ jelzés azt fejezi ki, hogy a töltelékelem kialakulása nem függ a beszélőtől, minden esetben megjelenik, még pongyola ejtésnél is, és nagy energiával van jelen. A svá- jelzés az jelenti, hogy a töltelékhang megvalósulása erősen beszélőfüggő, csak gondozott ejtésnél jön létre, akkor is kisebb energiával. A svá jelzés pedig azt jelenti, hogy a hangbetoldás létrejötte többnyire beszélőfüggetlen, csak igen laza ejtésnél nem valósul meg. A svá* jelzést akkor használjuk, amikor a svá képzése magasabb nyelvállású lesz, ilyenkor az F2 a 2000 Hz-es frekvencia felé tolódik. Ez akkor fordul elő, amikor palatális mássalhangzó következik a svát tartalmazó mássalhangzó után.

A [b] *hang CC kapcsolatokban.* Általánosságban kimondható, hogy a [b]-t tartalmazó CC kapcsolatokban a [b]-re a legkomplexebb hatást az [m] gyakorolja, mivel a [b]+[m] kapcsolatban a zárfelpattanási rész elmarad (*lá**bm**elegítő*), tehát a [b] rövidül, az [m]+[b] kapcsolatban pedig a [b] hang fojtott zönge részét rövidíti le a nazális mássalhangzó (*szá**mb**ól*). Ezek a hatások egyrészről az azonos képzési helyből, másrészről a nazális-orális üregváltásból erednek. Hasonló rövidítő hatása van továbbá az [ɲ]-nek is (*fé**nyb**en*). Bizonyos [b]+C kapcsolatokban a [b] zárfelpattanása módosul, vagyis helyette egy svá elem jelenik meg (5.21. táblázat), ennek időtartama akár 30–40 ms is lehet (5.37. ábra). A [b]+C esetekben az összes lehetséges ilyen CC kapcsolatra nézve mintegy 50%-ban van jelen a svá (Olaszy 2007b). Ez a szám a bilabiális és a többi képzési hely artikulációs kapcsolatát fejezheti ki közvetett formában. A [b] hang időtartama átlagosan 78 ms



5.37. ábra. A svá megvalósulása a *labda* (610 ms) szóban a két zöngés zárhang találkozási pontján. A svá 6–8 dB-nyit emelkedik ki a fojtott zöngék szintjéből

a [b]+C kapcsolatokban, a megvalósulási időtartamsávja 64-től 89 ms-ig terjed. A legrövidebb a [j] előtt, a leghosszabb az [r] előtt. A svá jelenléte jellemzően hosszítja a [b] időtartamát. A zöngésedésből létrejövő [b]-re 68 ms-os az átlag, ami azt mutatja, hogy a zöngésedésnek akusztikai következménye is van. A C+[b]

5.21. táblázat. A [b] mássalhangzó zárfelpattanásaként realizálódó svá elemek a kapcsolódó C függvényében [b]+C kapcsolatokban

b+C	d	G	g	m	n	N	j	v	z	Z	l	r
Svá jelenléte	svá	svá	svá	-	svá-	svá-	-	-	-	svá	-	svá+

kapcsolatokban a [b] időtartama átlagosan 69 ms. Az időtartam-különbségek az átlagok között szignifikánsak. A [b] időtartamsávja az ilyen kapcsolatokban 41 ms-tól 80 ms-ig terjed. Legrövidebb az [m] után, a leghosszabb, amikor [d d͡z] előzi meg.

A [d] *hang CC kapcsolatokban.* A [d]-re a legösszetettebb hatást az [n] gyakorolja, mivel a [d]+[n] kapcsolatban a [d] zárfelpattanása elmarad (*va**dn**ak*). Az [n]+[d] kapcsolatban a [d] fojtott zönge része lényegesen lerövidül egyrészről a megelőző C nazális volta miatt, másrészről az azonos képzési hely következtében (*poro**nd**on*). Hasonló rövidítő hatása van továbbá az [m],[ɲ] hangnak is (*háre**md**al, fé**ny**duda*). A [d] zárfelpattanása elmarad a [d]+[l] kapcsolatban is (*va**dl**iba*). A [d] zárfelpattanása módosul, vagyis helyette egy svá elem jelenik meg egyes [d]+C kapcsolatokban (5.22. táblázat). Ez a legkonzekvensebben a [d]+[r], valamint a [d]+[v] kapcsolatban valósul meg (*va**dr**a, a**dv**a*). A [d]+C esetekben az összes

5.22. táblázat. A [d] mássalhangzó zárfelpattanásaként létrejövő svá elemek a kapcsolódó C függvényében a [d]+C kapcsolatokban

d+C	b	g	m	n	N	j	j	l	r
Svá jelenléte	svá	svá	svá	-	-	svá-	svá+	-	svá+

lehetséges ilyen CC kapcsolatra nézve mintegy 60%-ban van jelen svá elem. Ez az érték a legmagasabb a zöngés zárhangok sorában, tehát a [d] képviseli a legbonyolultabb artikulációs helyzetet a zöngés zárhangokat illetően a CC kapcsolatokban. A [d]-re számított összesített átlagos hangidőtartam a [d]+C helyzetre 77 ms. A megvalósulási időtartamsávja 69-től 85 ms-ig terjed. A legrövidebb a [g] előtt, a leghosszabb, amikor [r] követi. A zöngésedésből létrejövő [d]-re 68 ms-os az átlag, ami azt mutatja, hogy a zöngésedésnek akusztikai következménye is lehet. Itt a kevés számú minta korlátozta a statisztikai megerősítést. C+[d] kapcsolatban a [d] átlagos időtartama 61 ms, ez lényegesen és szignifikánsan rövidebb, mint a [d]+C helyzetű hangé. A [d] időtartamsávja az ilyen kapcsolatokban 33 ms-tól 75 ms-ig terjed. Legrövidebb az [n] után, a leghosszabb, amikor [b g] előzi meg.

A [ɟ] *hang CC kapcsolatokban.* A [ɟ] akusztikai vetülete stabil a beszédfolyamatban, nem illeszkedik a szomszédos hangokhoz, inkább azokat kényszeríti magához. Ebből következik, hogy ez a hang erősen hat a hozzá kapcsolódó bármely hang frekvenciaszerkezetére, mintegy magához kényszeríti annak akusztikai vetületét a kapcsolódási ponton. Ez a hatás a zárfelpattanás helyett kialakuló svá formánsszerkezetében is érvényesül, az F2 a svá-ban felfelé csúszik a 2000 Hz-es

pont környékére. Ez a svá tehát egy magasabb nyelvállású svá variánsnak (svá*) tekinthető minden esetben (5.23. táblázat). A [ɟ]+C esetekben az összes lehetséges

5.23. táblázat. A mássalhangzó zárfelpattanásaként létrejövő svá elemek a kapcsolódó C függvényében a [ɟ]+C kapcsolatokban

G+C	b	d	g	m	n	N	j	v	z	Z	l	r
Svá jelenléte	svá*	-	svá*-	svá*	-	-	-	svá*	-	-	-	svá*+

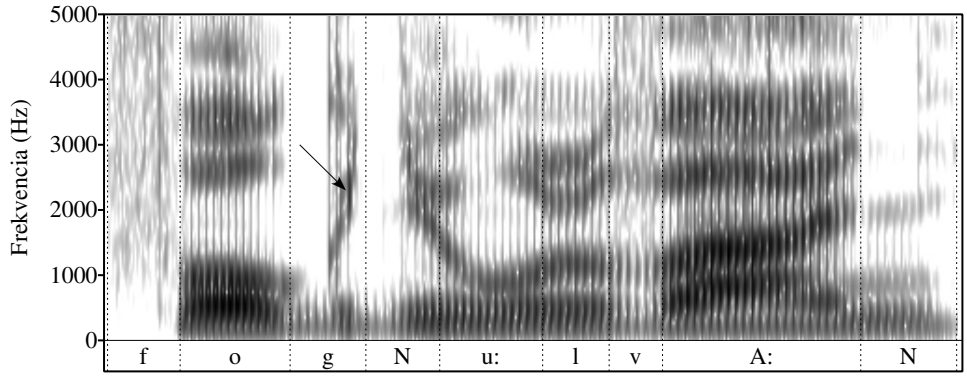
CC kapcsolatra nézve mintegy 35%-ban van jelen svá elem. Ez a legalacsonyabb érték a svá előfordulására a zöngés zárhangok köréből. Ez azt mutatja, hogy a [ɟ] artikulációs pozíciójából egyszerűbben lehet eljutni a vele kapcsolódó mássalhangzók artikulációs pozíciójába, mint a többi zöngés zárhangnál. A [ɟ]-re számított összesített átlagos hangidőtartam gyakorlatilag ugyanannyi a [ɟ]+C helyzetre, mint a C+[ɟ] kapcsolatra (71 ms, 73 ms), a hang időtartamára tehát nem hat a CC kapcsolatban elfoglalt helyzete. A [ɟ]+C helyzetben az időtartamsáv 51 ms-tól 90 ms-ig terjed. A [ɟ] a legrövidebb a [z] előtt, a leghosszabb, amikor [r] követi. A C+[ɟ] helyzetben az időtartamsáv 47 ms-tól 87 ms-ig terjed. Legrövidebb a hang az [m ɲ] után, a leghosszabb, ha [r] előzi meg.

A [g] *hang CC kapcsolatokban.* A [g]-vel kapcsolatos legjellemzőbb szerkezeti változás az n+[g] kapcsolatban valósul meg. Az egymásra hatás mindkét irányba végbemegy. Egyrészről a [g] megváltoztatja az [n] képzési helyét, a nazális hang velarizálódik, másrészről a fojtott zönge lényegesen lerövidül a megelőző [n] miatt. Hasonló rövidítő hatása van továbbá az [m], [ɲ] hangoknak is (*sze**mg**ödör*). A zárfelpattanás helyett megjelenő svá megvalósulásait az 5.24. táblázatban adjuk meg. A palatálisokhoz való kapcsolódásban a svá magasabb nyelvállású variánsa

5.24. táblázat. A mássalhangzó zárfelpattanásaként létrejövő svá elemek a kapcsolódó C függvényében a [g]+C kapcsolatokban.

g+C	b	d	G	m	n	N	j	v	z	Z	l	r
svá jelenlét	svá	svá	svá+*	-	svá	svá*	-	svá+	-	-	-	svá+

jön létre (5.38. ábra). A [g]+C kapcsolatokban az összes lehetséges CC esetre vetítve mintegy 50%-ban van jelen svá elem, amely a legkonzekvensebben a [g]+[r]-ben jön létre. A [g]-re számított összesített átlagos hangidőtartam a [g]+C helyzetben 71 ms, a C+[g] kapcsolatban 65 ms. A két átlag közötti eltérés nem szignifikáns, a zöngés zárhang időtartama tehát nem függ a hangkapcsolatban elfoglalt helyzettől. Ugyanez a helyzet a zöngésedésből létrejövő [g]-nél is, aminek átlaga 70 ms. A [g]+C helyzetben az időtartamsáv 59 ms-tól 83 ms-ig terjed. A [g] időtartama a legrövidebb az [l] előtt, a leghosszabb, amikor [v] követi. C+[g] helyzetben a [g] időtartamsávja 42 ms-tól 75 ms-ig terjed, a legrövidebb az [n] után, a leghosszabb, amikor [v l] előzi meg.

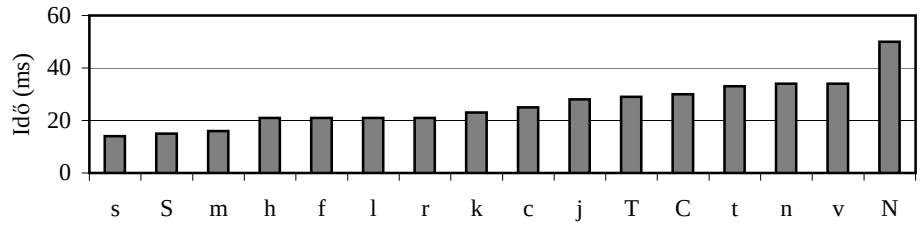


5.38. ábra. A *fognyúlvány* (921 ms) szóban a svá (nyíllal jelölve) magasabb nyelvállású variánsa jön létre, az F2-je fokozatosan csúszik a 2000 Hz-es érték felé, azaz illeszkedik a következő palatális mássalhangzó akusztikai vetületéhez

Zöngétlen zárhangok CC kapcsolatokban. A zöngétlen zárhangok a zöngések párjai. Csak a gerjesztésben van különbség közöttük, a zárszakaszban itt a néma fázis alkotja a hang első és döntő részét. A zöngétlen zárhangok szerkezetében a következő változások jöhetnek létre CC kapcsolatban. Ha a zárhang a kapcsolat első eleme, akkor a zárfelpattanása csatlakozik a következő hanghoz, és három realizációs formában jön létre a hozzá kapcsolódó mássalhangzó függvényében: a) a zárfelpattanás a VCV helyzetihez hasonló (*lép**j**en*), b) a hang zárfelpattanási eleme elmarad, vagy nagyon kis intenzitással van jelen (*né**pm**ese*), c) a zárfelpattanás zörejeleme megnyúlik (*né**pt**anító*). Ha a zöngétlen zárhang a kapcsolat második tagja, akkor a néma fázisa kapcsolódik a megelőző hanghoz, az egymásra hatásban csupán a néma szakasz rövidülhet. Ez döntően a nazálisok után jöhet létre (*rá**mp**a*). Az akusztikai vetület tekintetében a zöngétlen zárhangoké megegyezik a zöngés párjukra jellemzővel, hiszen az akusztikai vetület a képzési helytől függ, nem pedig a gerjesztéstől. Megadjuk a zöngétlen zárhang zárfelpattanása és a következő hang kezdete közötti időt is. A VOT-CC a definíció szerinti kapcsolódásokban tehát leolvasható.

A [p] *hang CC kapcsolatokban.* A [p]-re a legkomplexebb hatást az [m] és az [ɲ] gyakorolja. A [p]+[m] kapcsolatban a zárfelpattanás el is maradhat (*ké**pm**ás*), az [m]+[p] kapcsolatban pedig a [p] hang néma fázisa rövidül (*rá**mp**a*). Ez a hatás egyrészről az azonos képzési helyből, másrészről az üregváltásból ered. A [p]+[ɲ] kapcsolatban a [p]-nek a zárfelpattanási zöreje a leghosszabb (*né**pny**omor*), az [ɲ]+[p] kapcsolatban pedig a [p] néma fázisa rövidül (*fé**nyp**edál*). A [p] zárfelpattanásának hossza a [p]+C kapcsolatokban erősen függ a C-től (5.39. ábra), és jellemzően hosszabb, mint a VCV helyzetű érték, ami 12–16 ms körüli (Olaszy 2007b). A nyúlás mértéke kifejezi az artikulációs kapcsolódási mozgások bonyo-

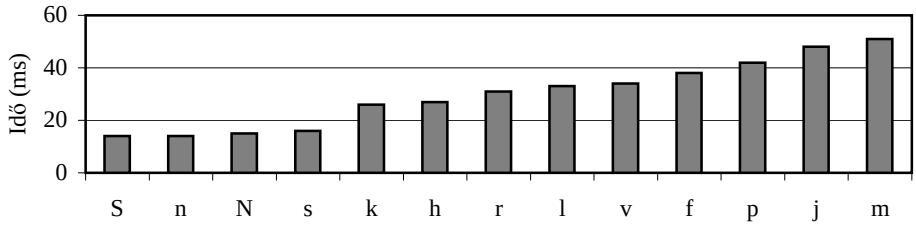
lultságát. Az ábrából leolvashatók a jellemző VOT-CC értékek is. A néma fázis



5.39. ábra. A [p] zárfelpattanási zörejének kezdete és a mássalhangzó kezdőpontja közötti idő [p]+C kapcsolatokban a C függvényében

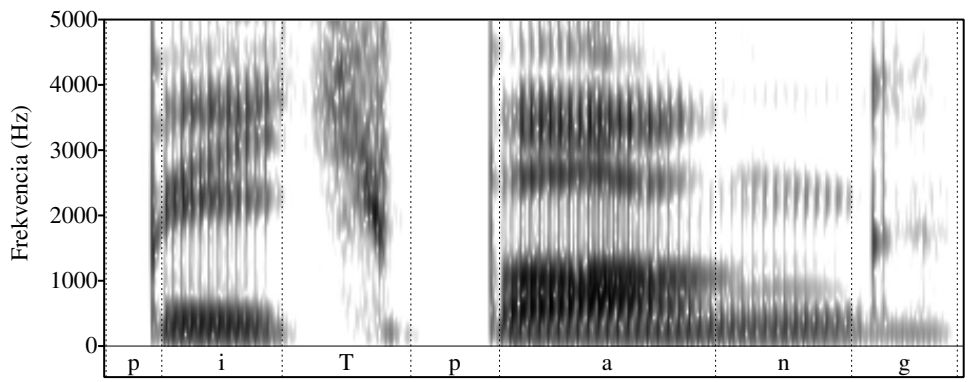
lerövidülése akkor következik be, ha a [p] nazális mássalhangzó után következik. A leghosszabb néma fázis az [l]-lel való kapcsolódásra jellemző (***Alp****ok*). Ezek szerint itt a legbonyolultabb az artikulációs mozgássorozat a két hang összekapcsolásánál. A [p]-re jellemző átlagos hangidőtartam 87 ms, nem függ a [p] pozíciójától a CC kapcsolatban. A zöngétlenedésből keletkezett [p] (*do****bt****ető*) időtartama viszont jellemzően rövidebb, 77 ms. A rövidülés szignifikáns. A [p]+C kapcsolatokban a [p] időtartamsávja 64 ms-tól 104 ms-ig terjed, a legrövidebb az [s] előtt, a leghosszabb, amikor [ɲ] követi. A C+[p] helyzetben a [p] időtartamsávja 68 ms-tól 109 ms-ig terjed. Legrövidebb az [ɲ] után, a leghosszabb, amikor [l] előzi meg.

A [t] *hang CC kapcsolatokban.* A [t]-re a CC kapcsolatokban a legkomplexebb hatást az [n] gyakorolja. A [t]+[n], esetében (*át****tn****evel, át****tny****úl*) a zárfelpattanás zöreje a legrövidebb, ezzel egyidejűleg a [t] teljes időtartama viszont a leghosszabb. A zárfelpattanás rövidülése az azonos, illetve közel azonos képzési helyből adódik. A különlegesen hosszú néma fázis (jellemzően 100 ms) pedig azt mutatja, hogy az üreg-, továbbá a gerjesztésváltásra hosszabb idő kell ezekben a kapcsolatokban, mint egyébként. Az [n]+[t] esetében az azonos képzési hely miatt a [t] időtartama a legrövidebb. A [t] zárfelpattanásának hossza a [t]+C kapcsolatokban erősen függ a C-től (5.40. ábra). A VCV helyzetre jellemző átlag 23 ms, ennél rövidebb és hosszabb értékek is előfordulnak a CC kapcsolatokban. Az ábrából leolvashatók a jellemző VOT-CC értékek is. A [t] jellemző időtartama független a CC kapcsolatban elfoglalt helyétől, átlagosan 86 ms. A zöngétlenedéskor létrejött [t] időtartama (*va****dk****an*) szignifikánsan rövidebb (74 ms), mint a nem zöngétlenedésből keletkezetté. A zöngétlenedés tehát akusztikai információban is kifejeződhet, a hanghosszúság jellemezheti a koartikulációs hatást. A [t]+C helyzetben az időtartamsáv 51 ms-tól 113 ms-ig terjed. Legrövidebb a hang az [s] előtt, a leghosszabb, amikor [n] követi. A C+[t] kapcsolódásban a hang időtartamsávja 52 ms-tól 106 ms-ig terjed. Legrövidebb az [n] után, a leghosszabb, ha [r] előzi meg.



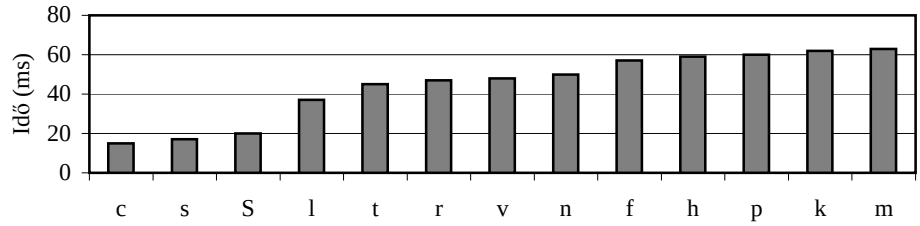
5.40. ábra. A [t] zárfelpattanási zörejének kezdete és a mássalhangzó kezdőpontja közötti idő [t]+C kapcsolatokban a C függvényében

A [c] *hang CC kapcsolatokban.* A [c] akusztikai vetülete – hasonlóan a zöngés párjához – stabil a beszédfolyamatban. Frekvenciaszerkezeti szempontból nem illeszkedik a szomszédos hangokhoz, inkább azokat kényszeríti magához. A [c] képzési helyére jellemző akusztikai vetület azonban csak a zárfelpattanási zörej elején érvényesíti hatását, utána a zörej frekvenciakomponensei a követő mássalhangzóra jellemző akusztikai vetület frekvenciaértékei felé mozdulhatnak el (*pi**typ**ang*). Ez főleg azokra az esetekre jellemző, amikor a zörejkomponens elég hosszú, tehát van idő arra, hogy a stabil, 2000 Hz-es frekvenciapontról a követő mássalhangzóra jellemző akusztikai vetület elmozdítsa a frekvenciakomponenseket, azaz hatást gyakoroljon a zörej alakulására (5.41. ábra). A [c] zárfelpattanásának



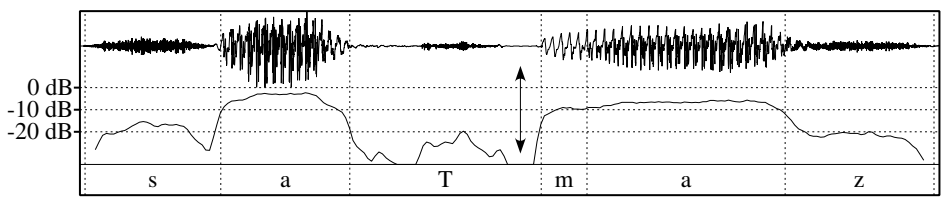
5.41. ábra. A *pitypang* (816 ms) szó frekvecniaszerkezete. A [c] zárfelpattanási zörejében elmozdulnak a frekvenciakomponensek a követő mássalhangzó akusztikai vetülete felé. A zörejgóc alsó frekvenciaértéke fokozatosan csökken 2000 Hz-ről 1300 Hz-re a [p]-hez való közeledéskor

hossza a [c]+C kapcsolatokban erősen függ a C-től (5.42. ábra). Ennek a sávnak közel a közepén helyezkedik el a VCV helyzetű hang zörejének hossza (23 ms). A rövid időtartamú zörejelem a legkisebb energiát igénylő artikulációs mozgásoknál jön létre. A többi esetben a zár-rés hangban különböző mértékű spirantizálódás tapasztalható, azaz a réselem zöreje megnyúlik. Ennek mértéke fejezi ki az artiku-



5.42. ábra. A [c] zárfelpattanási zörejének kezdete és a mássalhangzó kezdőpontja közötti idő [c]+C kapcsolatokban a C függvényében

lációs mozgások bonyolultságát a kapcsolatokban. A leghosszabb időtartam akkor jön létre, amikor a következő mássalhangzó bilabiális nazális zárhang (*Szatymaz*). Ebben a hangkapcsolatban a nagy VOT-CC értéket nem csupán a zörejelem nyúlása okozza, hanem a koartikulációs néma fázis kialakulása is Olaszy (vö. 2006a), amely a [c] és az [m] közé ékelődik (5.43. ábra). Az ábrán látható, hogy a 109 ms-os hangból a zárfelpattanás zörejeleme 50 ms-os, a lecsengés utáni néma rész 20 ms-nyi, ezután indul a nazális hang rezgése. Hasonló szerkezeti módosulás tapasztalható a [c]+[n] kapcsolatban is (*poron**tyn**ak*). A [c] jellemző időtartama [c]+C kapcsolatban

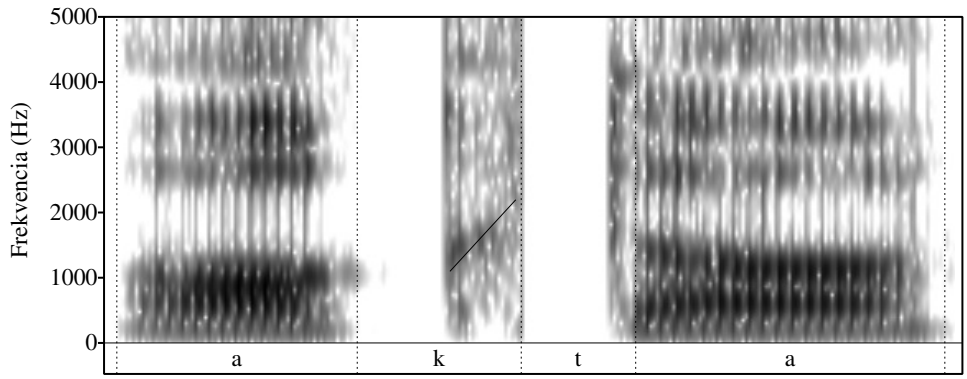


5.43. ábra. A *Szatymaz* (747 ms) szó szerkezeti elemei. A [c] és az [m] közé ékelődik a koartikulációs néma fázis. A VOT-CC értéke a példában 70 ms, amiből 20 ms a koartikulációs néma fázis (nyíl)

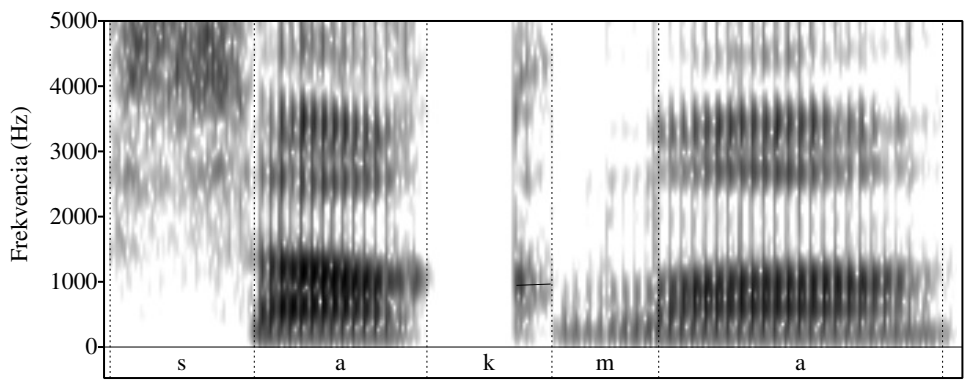
átlagosan 94 ms, az időtartamsáv 78 ms-tól 105 ms-ig terjed, tehát meglehetősen szűk. Legrövidebb a hang az [s] előtt, a leghosszabb, amikor [n] követi. A C+[c] kapcsolódásban a [c] hang átlagos időtartama 102 ms, az időtartamsáv 91 ms-tól 114 ms-ig terjed, szintén szűk. Legrövidebb a hang az [t͡s] után, a leghosszabb, amikor [r] előzi meg.

A [k] *hang CC kapcsolatokban.* A kapcsolódó C-k hatása a [k]-ra összetett. A [k] artikulációja illeszkedő jellegű, ami azt jelenti, hogy idomul a hozzá kapcsolódó hanghoz. A zárképzési mozzanat a veláris terület különböző pontjaira tolódik el a kapcsolódó hang artikulációs vetületének a függvényében. Ez azt eredményezi, hogy a [k] hang akusztikai vetülete is rugalmasan alakul, erős frekvenciailleszkedést mutat a CC kapcsolatokban. Például az *akta* szóban a [k] zárfelpattanási zörejének alsó zörejgóca az 1300 Hz körüli értékről a [t]-re jellemző magasabb frekvenciaérték

felé (1700 Hz) mozdul el (5.44. ábra), míg a *szakma* szóban nem látható ilyen elmozdulás (5.45. ábra). Ugyanakkor tudnunk kell, hogy a [t] hatásaként előbb említett frekvenciamozgás nem jön létre például az *iktató* szó [k] hangjában, mivel a [k] az őt megelőző hanghoz is frekvenciailleszkedést hajt végre, és ebben az esetben a zörejének alsó frekvenciagóca eleve magas frekvencián lesz (1700 Hz körül) az [i] miatt. A [k] zárfelpattanásának hossza a [k]+C kapcsolatokban függ a C-től

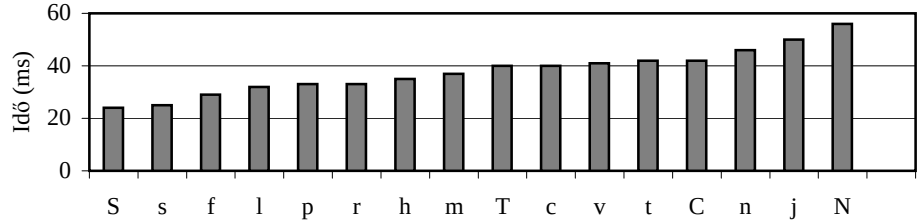


5.44. ábra. Az *akta* (521 ms) szó szerkezeti elemei. A [k] hang illeszkedik az őt követő mássalhangzó akusztikai vetületéhez. A [t] előtti helyzetben az alsó zörejgóc felfelé mozog (vonallal jelölve)



5.45. ábra. A *szakma* (624 ms) szó szerkezeti elemei. A [k] hang illeszkedik az őt követő mássalhangzó akusztikai vetületéhez. Az [m] előtti helyzetben az alsó zörejgóc alacsonyan marad (vonallal jelölve)

(5.46. ábra). A VCV helyzetre jellemző átlaga 39 ms. Ennél rövidebb és hosszabb értékek is előfordulnak a CC kapcsolatokban. A [k]-ra a legkomplexebb hatást a nazális mássalhangzók gyakorolják. Az ilyen [k]+C kapcsolatokban kialakul a koartikulációs néma fázis, a C+[k] kapcsolatokban pedig a néma fázis rövidül a nazális hang hatására. A [k]+C kapcsolatokban a nazális hangok előtti zárfelpattanási



5.46. ábra. A [k] zárfelpattanási zörejének kezdete és a mássalhangzó kezdőpontja közötti idő [k]+C kapcsolatokban a C függvényében

zörej viszonylag hosszú. Ebben jelentős szerepet vállal a koartikulációs néma fázis jelenléte is. A C+[k] kapcsolatokban az [ŋ] rövidítő hatása a legmarkánsabb. A [k] összesített időtartamátlaga nem függ a CC kapcsolatban elfoglalt helyétől, mindkét esetben 88 ms körüli érték jellemző a hangra. A zöngétlenedésből keletkezett [k] hangok időtartamátlaga 72 ms, ami szignifikánsan rövidebb, mint a nem zöngétlenedésből keletkezett hangé. A [k]+C helyzetben a hang időtartamsávja 73 ms-tól 105 ms-ig terjed. Legrövidebb az [s] előtt, a leghosszabb, amikor [ɲ] követi. A C+[k] helyzetben a hang időtartamsávja 64 ms-tól 104 ms-ig terjed. Legrövidebb az [ŋ] után, a leghosszabb, amikor [h] előzi meg.

Zöngés réshangok CC kapcsolatokban. A magyar zöngés réshangok szerkezeti felépítése CC kapcsolódásokban nem változik. Ez abból adódik, hogy a zöngés elem és a rá szuperponálódott turbulens áramlás frekvenciaszerkezete olyan, amire nemigen tud hatni a szomszédos mássalhangzó.

A [v] *hang CC kapcsolatokban.* A [v]-t tartalmazó CC kapcsolatokban a [v], mint réshang szerkezetileg egységesnek mondható. A hangban mindig van némi zörejkomponens is, ennek erőssége a csatlakozó mássalhangzó függvénye. A legzörejesebb a [v] a palatális zárhangok előtt (*évgy**ű**rű*), legkevésbé zörejes a dentialveoláris zöngés réshang, valamint a [j] hang előtt (*évj**á**rat*). Meg kell jegyeznünk, hogy a [v] zörejesedése beszélőfüggő is. A [v]-ben kialakulhat svá a kapcsolódási ponton (ez VCV helyzetre egyáltalán nem jellemző). A legjellemzőbb svá-generáló csatlakozó mássalhangzók a [d], [z], [r], ezek közül a legerősebb az [r]. A [v]-ben megjelenő svá elem kifejezi, hogy mely kapcsolódásokhoz kell nagy artikulációs energia. A [v]-re számított összesített átlagos hossz a C+[v] kapcsolatokban szignifikánsan rövidebb (59 ms), mint a [v]+C helyzetben (73 ms). A [v]+C helyzetben a hang időtartamsávja 57 ms-tól 86 ms-ig terjed. Legrövidebb a [z] előtt, a leghosszabb, amikor [b] követi. A C+[v] helyzetben a hang időtartamsávja 34 ms-tól 72 ms-ig terjed. Legrövidebb az [ɲ] után, a leghosszabb, ha [d] előzi meg.

A [z] *hang CC kapcsolatokban.* A [z]-t tartalmazó CC kapcsolatokban a [z], mint zöngés-zörejes réshang szerkezetileg egységesnek mondható, a környezeti mássalhangzók lényegileg csak a hang időtartamára vannak hatással, szerkezeti komponenseire és frekvenciaszerkezetére nem. Svá-szerű elemek nemigen fordulnak elő a hangban a CC kapcsolat csatlakozási pontján. A hangra jellemző zörejkomponensek CC kapcsolatban is ugyanúgy megtalálhatók, mint a VCV helyzetű hangnál. A [z]-re számított összesített átlagos hossz a C+[z] kapcsolatokban hasonló értéket mutat, mint a [z]+C helyzetben (83–86 ms). Ez szignifikánsan hosszabb, mint ami VCV helyzetű [z]-re jellemző érték (71 ms). A [z]+C helyzetben a hang időtartamsávja 71 ms-tól 94 ms-ig terjed. Legrövidebb az [l] előtt, a leghosszabb, amikor [ɲ] követi. A C+[z] helyzetben a hang időtartamsávja 67 ms-tól 103 ms-ig terjed. Legrövidebb az [ɟ] után, a leghosszabb, amikor [r l] előzi meg.

A [ʒ] *hang CC kapcsolatokban.* A hang stabil tulajdonságokkal rendelkezik a CC kapcsolatokban. Összetett, zöngés-zörejes frekvenciaszerkezetére nincs lényeges hatással a hozzá csatlakozó mássalhangzó. Svá-szerű elemek nemigen fordulnak elő a hangban a CC kapcsolat csatlakozási pontján. A hangidőtartamok átlagaiban nemigen tükröződik a hangkapcsolatban megvalósuló artikulációs mozgások bonyolultsága. A hang átlagos időtartama 83 ms, pozíciótól függetlenül. Hasonló hosszúságú a zöngésedésből keletkező is. A [ʒ]+C helyzetben a hang időtartamsávja 76 ms-tól 95 ms-ig terjed, ami meglehetősen szűk. Legrövidebb a hang a [ɟ] előtt, a leghosszabb, amikor [j] követi. A C+[ʒ] helyzetben a hang időtartamsávja 74 ms-tól 98 ms-ig terjed, a legrövidebb az [m] után, a leghosszabb, amikor [j] előzi meg.

Zöngétlen réshangok CC kapcsolatokban. A magyar zöngétlen réshangok szerkezeti felépítése CC kapcsolódásokban változó képet mutat. Egyrészről bizonyos hasonulások miatt időszerkezeti változások léphetnek fel, másrészről a nazálisok előtti helyzetben létrejön a koartikulációs néma fázis a réshang után.

Az [f] *hang CC kapcsolatokban.* Az [f] hang stabil tulajdonságokkal rendelkezik a CC kapcsolatokban, frekvenciaszerkezetére nincs hatással a hozzá csatlakozó mássalhangzó. A hang időtartama átlagosan 92 ms, függetlenül a CC kapcsolatban lévő helyzetétől. A zöngétlenedésből keletkezett [f] időtartama viszont szignifikánsan rövidebb (73 ms). Az [f]+C helyzetben lévő hang időtartamsávja 69 ms-tól 110 ms-ig terjed, a legrövidebb az [s] előtt, a leghosszabb, amikor [m] követi. A C+[f] helyzetben a hang időtartamsávja 68 ms-tól 113 ms-ig terjed, a legrövidebb a [t͡s] után, a leghosszabb, amikor [l] előzi meg.

Az [s] *hang CC kapcsolatokban.* Az [s] a CC kapcsolatokban szerkezetileg egységesnek mondható, a környezeti mássalhangzók lényegileg csak a hang időtartamára vannak hatással, frekvenciaszerkezetére nem. Fontos tudni, hogy az [s]+[m]

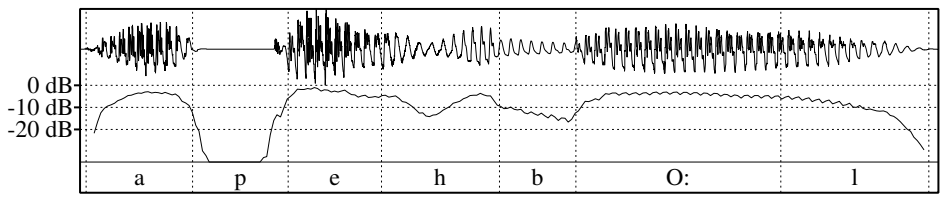
és [s]+[n] kapcsolatokban következetesen kialakul egy 30–40 ms-os koartikulációs néma fázis a réshang után. További fontos kérdés, hogy a [t]+[s] kapcsolatban mikor jön létre hasonulás (*játs**z**ótárs*), aminek folytán [t͡sː] hang lesz a kiejtésben, illetve mikor marad meg a zárhang + réshang karakter (*áts**z**ervez*). Az időszerkezeti tulajdonságok vizsgálata kimutatta, hogy jogosan jelöljük két hangnak az ilyen kapcsolatokat morfémahatáron (*szé**tsz**ór*), mivel az [s] réselem szignifikánsan hosszabb az ilyen kapcsolatokban, mint a [t͡s] réseleme VCV helyzetben (Olaszy 2007b). Harrington (1988) angol percepciós kísérletei szerint is a zörejelem időtartama adja az elsődleges felismerési kulcsot annak megkülönböztetésében, hogy ugynazon képzési helyű (homorgán) zárhang és réshang-e a hallott szekvencia vagy zár-rés hang. Az [s]-re számított összesített átlagos hossz az [s]+C kapcsolatokban hasonló értéket mutat, mint a C+[s] helyzetben (111/113 ms). Az [s]+C helyzetben a hang időtartamsávja 95 ms-tól 131 ms-ig terjed, a legrövidebb a [p] előtt, a leghosszabb, amikor [n] követi. C+[s] helyzetben a hang időtartamsávja 103 ms-tól 127 ms-ig terjed, a legrövidebb a [t͡ʃ] után, a leghosszabb, amikor [l] előzi meg.

Az [ʃ] *hang CC kapcsolatokban.* Az [ʃ] hang stabil tulajdonságokkal rendelkezik a CC kapcsolatokban. A frekvenciaszerkezetére nincs lényeges hatással a hozzá csatlakozó mássalhangzó, az időtartamára viszont igen. A hangra a legkomplexebb hatást az [n] gyakorolja, hiszen az [ʃ]+[n] kapcsolatban a leghosszabb az időtartama (130 ms), az [n]+[ʃ] kapcsolatban pedig a legrövidebb (90 ms). Ki kell emelni, hogy az [ʃ]+[m] és [ʃ]+[n] kapcsolatokban is létrejön a 30–40 ms-os koartikulációs néma fázis az [ʃ] után, ami az üreg- és gerjesztésváltás komplex következménye. Ezt figyelembe kell venni a beszédtechnológiai feldolgozásoknál. A [t]+[ʃ] kapcsolat morfémahatáron két hangként kezelendő, mivel nem jön létre a hasonulás (*áts**i**klik*). A mérési eredmények szerint a morfémahatáron lévő réselem szignifikánsan hosszabb az ilyen kapcsolatokban, mint a [t͡ʃ] réseleme VCV helyzetben. Jogosan jelöljük tehát két hangnak az ilyen kapcsolatokat, ha morfémahatáron vannak. Más helyzetekben hasonulással kell számolni (*hát**s**ó*). A [ʃ] hang időtartama átlagosan 112 ms függetlenül a kapcsolatban elfoglalt helyzetétől, valamint attól is, hogy zöngétlenedésből jött-e létre. A [ʃ]+C helyzetben a hang időtartamsávja 90 ms-tól 132 ms-ig terjed, a legrövidebb a [t͡s] előtt, a leghosszabb, amikor [n] követi. C+[ʃ] helyzetben a hang időtartamsávja 82 ms-tól 129 ms-ig terjed, a legrövidebb időtartamú a [t͡ʃ] után, a leghosszabb, amikor [r] előzi meg.

A [h] *hang CC kapcsolatokban.* Ez a réshang változatos képet mutat a CC kapcsolatokban. Többféle variánsa is kialakul, mind a gerjesztést, mind pedig a képzési helyet illetően. Ha a hang a mássalhangzó-kapcsolat második eleme akkor [h]-t ejtünk. Zöngétlen (*la**kh**ely)*, valamint zöngés változatban is (*sza**jh**a*) előfordul. A gerjesztési formát a csatlakozó mássalhangzó határozza meg. Ha ez zöngés, akkor [ɦ]-t ejtünk. A [h] hang átlagos hossza a C+[h] kapcsolatokban 64 ms,

az időtartamsávja 44 ms-tól 81 ms-ig terjed. Legrövidebb a hang a [t͡s] után, a leghosszabb, amikor [n] előzi meg.

A veláris [x] változat a CC kapcsolat első elemeként jöhet létre. Ilyenkor a [x] hang idomul az őt követő magánhangzó akusztikai vetületéhez, melynek következtében a zörejgóc jellemző frekvenciája a csatlakozó magánhangzó F2-jének magasságában jelenik meg a spektrumban. Ennek megfelelően különböztetünk meg velári képzési helyű (*sa**h**nak*), illetve palatoveláris [x] variánst (*te**ch**nika*). A [x] hangnak is létezik zöngés változata (*APE**H**ből*), melyet a [ɣ] IPA hangszimbólummal jelölnek. Az 5.47. ábrán jól látható, hogy folyamatosan zöngés gerjesztés van jelen a CC kapcsolatban. A [x] hang átlagos időtartama 93 ms, időtartamsávja



5.47. ábra. A [x] hang zöngés változata [ɣ] az *APEHből* (810 ms) szó ejtésekor

69 ms-tól 111 ms-ig terjed. Legrövidebb a hang az [f] előtt, a leghosszabb, amikor [j] követi. Az adat [x]+C kapcsolatokra vonatkozik. A C+[h] kapcsolat [h] hangja átlagosan szignifikánsan rövidebb, mint a [x]+C kapcsolat [x] hangja.

Zár-rés hangok CC kapcsolatokban. A [d͡z] és [d͡ʒ] zöngés zár-rés hangok ritkán fordulnak elő CC kapcsolatokban (*e**dzl**ek*) a zöngésedésből keletkezett változatuk kicsit sűrűbben (*lé**cb**e*). Szerkezetükre nincs hatással, hogy CC kapcsolatban vannak. A [d͡z] időtartama átlagosan 101 ms, a hang időtartamsávja 81 ms-tól 124 ms-ig terjed a [d͡z]+C kapcsolatokban. A hang a legrövidebb a [g] előtt, a leghosszabb, amikor [l] követi. A [d͡ʒ] időtartama átlagosan 107 ms, a hang időtartamsávja 94 ms-tól 132 ms-ig terjed [d͡ʒ]+C kapcsolatban. A hang a legrövidebb a [v] előtt, a leghosszabb, amikor [r] követi. A [t͡s] és [t͡ʃ] zöngétlen zár-rés hangok réselemének frekvenciaszerkezetére nincs hatással a CC kapcsolat, a néma fázis időtartamára azonban igen. A legkomplexebb ilyen hatást a kapcsolódó nazális hangok okozzák. A C+[t͡s] és C+ [t͡ʃ] kapcsolatokban a nazális hang lerövidíti a néma fázist (*ka**nc**a, ka**ncs**al*), a [t͡s]+C és [t͡ʃ]+C kapcsolatokban pedig kialakulhat a 30–40 ms-os koartikulációs néma fázis (*Vá**cn**ak, o**csm**ány*), ami a réselem vége és a nazális hang kezdete közé ékelődik (hasonlóan, ahogyan a réshangoknál láttuk). Általános szerkezeti tulajdonság, hogy a réselem a kapcsolatok többségében hosszabb, mint a zárszakasz. A réselem a teljes hang 60–70%-át teszi ki. A [t͡s] és [t͡ʃ] átlagos teljes hossza 110 ms a CC kapcsolatokban. A [t͡s] hang időtartamsávja 86 ms-tól 141 ms-ig terjed a [t͡s]+C kapcsolatokban. A hang a legrövidebb a [t͡ʃ]

előtt, a leghosszabb, amikor [n] követi. A C+[t͡s] helyzetben a hang időtartamsávja 88 ms-tól 128 ms-ig terjed, a legrövidebb a [c] után, a leghosszabb, amikor [p] előzi meg. A [t͡ʃ] hang időtartamsávja 78 ms-tól 134 ms-ig terjed [t͡ʃ]+C kapcsolatban. A hang a legrövidebb az [s] előtt, a leghosszabb, amikor [n] követi. A C+[t͡ʃ] helyzetben a hang időtartamsávja 88 ms-tól 128 ms-ig terjed, a legrövidebb a [h] után, a leghosszabb, amikor [r] előzi meg.

Nazális mássalhangzók CC kapcsolatokban. A CC kapcsolatokban a nazális hangok különlegesnek számítanak, hiszen a képzésük során az orális csatorna mellé (amelyben alapvetően zárat hozunk létre háromféle képzési hellyel) belép a nazális csatorna, ezen keresztül sugárzódik ki a hang. A nazális mássalhangzóknak így egyedi akusztikai vetületük van (5.25. táblázat). Erre az akusztikai vetületre van hatással a kapcsolódó mássalhangzó.

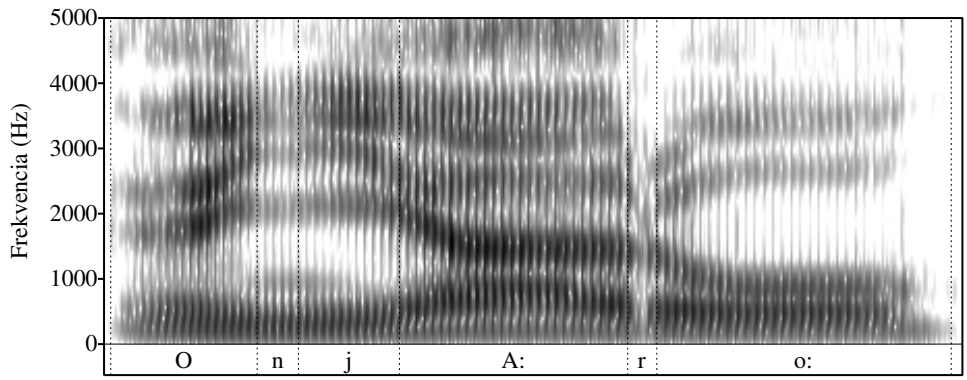
5.25. táblázat. A magyar nazális mássalhangzók képzési helyének akusztikai vetületei CC kapcsolatokra, tájékoztató formánsadatokkal kifejezve. A számértékek Hz-ben értendők. A megadott értékek tendenciákat fejeznek ki, hiszen a beszédben az akusztikai eredmény függ a beszélő fiziológiai és mentális adottságaitól is

	Formáns	Bilabiális	Dentialveoláris	Palatális
Akusztikai vetület	F1	250	250	250
	F2	1000	1350	2000
	F3	2200	2600	3500

Az [m] *hang CC kapcsolatokban.* Az [m] hang szerkezete nem változik lényegesen CC kapcsolatokban, folyamatos zöngés hangként van jelen. Egyetlen kapcsolatnál fordul elő, hogy a képzési helye megváltozik, bilabiálisból labiodentálissá válik, ha labiodentális mássalhangzó követi (*ha**mv**as, ká**mf**or*). Ennek a változásnak az eredménye az [ɱ] variáns, melynek akusztikai szerkezete nem tér el lényegesen a bilabiális [m] hangétól. Amennyiben másik nazális mássalhangzó kapcsolódik az [m]-hez (*háro**mn**ak*), a rezgésképen szinte nem is lehet megállapítani a kapcsolódási pontot, az átmenet folyamatos a két hang között. A frekvenciaszerkezet szempontjából a csatlakozó mássalhangzók nincsenek különösebb hatással az [m]-re, a hang időtartama viszont meglehetősen függ a hozzá csatlakozó mássalhangzótól. Az [m]+C kapcsolatokban az [m] átlagos időtartama 87 ms, az időtartamsávja 64 ms-tól 113 ms-ig terjed. A legrövidebb a hang a zöngétlen réshangok előtt, a leghosszabb a zöngés zárhangok előtt. A C+[m] kapcsolatban a hang átlagos időtartama 66 ms, ami szignifikánsan rövidebb, mint az [m]+C kapcsolatra jellemző érték. Ebben a kapcsolati formában az időtartamsávja 45 ms-tól 78 ms-ig terjed, a legrövidebb a [p] után, a leghosszabb, amikor [v] előzi meg.

Az [n] *hang CC kapcsolatokban.* Az [n] változó viselkedést mutat CC kapcsolatokban. Az [n]+C kapcsolatokban az [n] dentialveoláris képzésű akkor, ha a C képzési helye közeli, tehát ugyanaz, vagy alveoláris ([t d t͡s t ͡ʃ s l]). A többi

esetben az [n] képzési helye és módja megváltozik,variáns jön létre. Ezek közül a fonetikai irodalom a veláris képzésű hangot tartja számon (*ha**ng**os, e**ng**em*), mint legjellemzőbbet, azonban más formációk is kialakulnak. Ilyen például, amikor az [n] zár része nem jön létre. Ez akkor következik be, amikor [r s ʃ z ʒ j h] követi az [n]-t (*ö**nz**ő, ö**ns**zervező, be**ns**őséges*). Ezekben az esetekben az [n]-t megelőző magánhangzó utolsó harmada erősen nazalizálódhat, összeolvad a nazális hanggal. Példát mutatunk be a palatális réshang ilyen hatására az 5.48. ábrán. Az [n] palatalizálódhat is (*mondja*), ilyenkor az [ɲ]-hez hasonló hang jön létre. Az [n]



5.48. ábra. Az *önjáró* (828 ms) szó frekvenciaszerkezete. Az [n] akusztikai vetületének az F2-jét az őt követő palatális közelítő hang befolyásolja, magához idomítja

időtartama széles sávban mozog [n]+C kapcsolatokban, ahol átlagos hossza 76 ms, a legrövidebb az időtartama (jellemzően 50 ms) az [l], valamint a dentialveoláris és alveoláris zöngétlen réshangok előtt (*börtö**nl**akó, be**ns**őséges*), a leghosszabb, mintegy 100 ms a zöngés zárhangok előtt (*poro**nd**on*). A C+[n]kapcsolatokban az átlagos időtartama 65 ms, az időtartamsávja 50 ms-tól 77 ms-ig terjed. A hang a legrövidebb a [t͡ʃ] után, a leghosszabb, amikor [r] előzi meg.

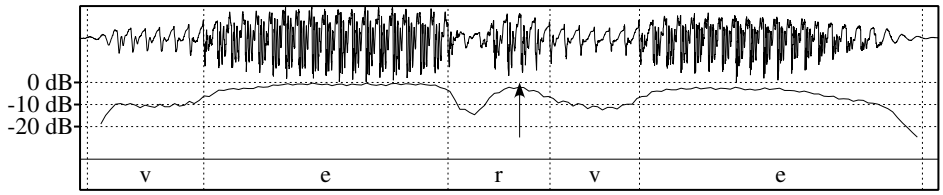
Az [ɲ] *hang CC kapcsolatokban.* Az [ɲ] hang viselkedése CC kapcsolatokban stabil. Akusztikai vetülete a kapcsolódó mássalhangzókat kényszeríti frekvenciailleszkedésre (hasonlóan, mint a [ɟ]. Az [ɲ]-re számított összesített átlagos hangidőtartam nem függ a CC kapcsolatban elfoglalt pozíciótól, átlagosan 79 ms. Az [ɲ]+C kapcsolatokban a hang időtartamsávja 59 ms-tól 95 ms-ig terjed. A hang legrövidebb az [s] előtt (*fé**nys**zóró*), a leghosszabb, amikor [g] követi (*fé**ny**galaxis*). A C+[ɲ] kapcsolatokban a hang időtartamsávja 62 ms-tól 97 ms-ig terjed, a legrövidebb a [t͡s] után, a leghosszabb, amikor [r] előzi meg.

A [j] *hang a CC kapcsolatokban.* A zöngés palatális mássalhangzó egységesen viselkedik a CC kapcsolatokban, hasonlóan, mint a magánhangzók a VC,

illetve CV helyzetekben. A frekvenciaszerkezetére a csatlakozó mássalhangzók nincsenek hatással, az időtartamában sem jönnek létre komoly változások. A hang átlagos időtartama 70 ms, függetlenül a kapcsolatban elfoglalt helyzetétől. A [j]+C helyzetben a hang időtartamsávja 58 ms-tól 77 ms-ig terjed, meglehetősen szűk. Legrövidebb a [d] előtt, a leghosszabb, amikor [n] követi. A C+[j] helyzetben az időtartamsáv 55 ms-tól 92 ms-ig terjed. Legrövidebb a [ɟ] után (*ágy**j**elenet*). A rövidülés oka lehet, hogy ebben a helyzetben azonos a képzési hely is és a gerjesztés is, tehát minimális artikulációs mozgásra van szükség a két hang közötti átmenetben. A leghosszabb a hang, amikor [r] előzi meg. A [j]-nek létezik zöngétlen variánsa C+[j] kapcsolatban, ez a [ç] hang, amely csak speciális hangkörnyezetben jön létre, általában abszolút hangsor végén, ha a C zöngétlen (*lé**pj***), illetve hangsor belsejében, ha a C zöngétlenedett mássalhangzó (*hí**vj** ki*). A palatális zöngétlen réshang akusztikai vetülete megegyezik a más palatális hangokéival. Részletes adatokat erről a hangról Olaszy (1985) munkájában találhatunk.

Az [l] *hang CC kapcsolatokban.* A CC kapcsolatokban az [l] mint zöngés közelítőhang, egységesen és magánhangzószerűen viselkedik. Minden mássalhangzóval kapcsolatba tud lépni. A hangidőtartamok tekintetében nincs szignifikáns eltérés a kapcsolatban elfoglalt helyzete szerint, az átlagos hossza 58 ms. Az [l]+C helyzetben a hang időtartamsávja 33 ms-tól 68 ms-ig terjed, a legrövidebb az [r] előtt, a leghosszabb, amikor [v] követi. Az [l]+C kapcsolatokban tehát, ha a C képzési helye megegyezik, vagy közel áll az alveoláris területhez (*e**lr**ont, e**ls**odor, o**ld**alra*), akkor az [l] időtartama rövidebb átlagú, mint más esetekben. Ez valószínűleg a kevesebb artikulációs ráfordítási energia következménye. A C+[l] helyzetben az időtartamsávja 45 ms-tól 76 ms-ig terjed. Legrövidebb a hang a [h] után, a leghosszabb, amikor [g] előzi meg.

Az [r] *hang CC kapcsolatokban.* A CC kapcsolatokban az [r] változó képet mutat, a hang részletes elemzése Gósy (2008b) munkájában található. Az [r] szerkezetéhez az esetek nagy többségében szorosan hozzátartozik a svá jelenléte, függetlenül attól, hogy az [r] a CC kapcsolatnak az első, illetve a második eleme. A svá elem elhelyezkedése a hangon belül attól függ, hogy az [r] melyik tagja a CC kapcsolatnak. Pontosabban arról van szó, hogy a svá kialakulhat a pergő hangra jellemző intenzitásminimum előtt (*vádra*, illetve után is (*kardok*). Az [r] megvalósulási időtartamsávja ilyen kapcsolatokban 43–69 ms. Az [r]+C kapcsolatokban az [r] négyféle szerkezeti formát vehet fel a csatlakozó mássalhangzótól függően. Az első és a legtöbbször létrejövő formátum, amikor az intenzitásminimum után, a hang végén egy teljes svá elem keletkezik. Ez olyan kapcsolódásokban jön létre, amelyekben a C hangintenzitása kicsi, illetve amikor nazális üregváltás van a két hang kapcsolódási pontján, tehát a [b p d t g k ɟ c v f m n] hangok előtt (5.49. ábra). Ilyenkor az [r] hang relatíve hosszú lesz a svá miatt.



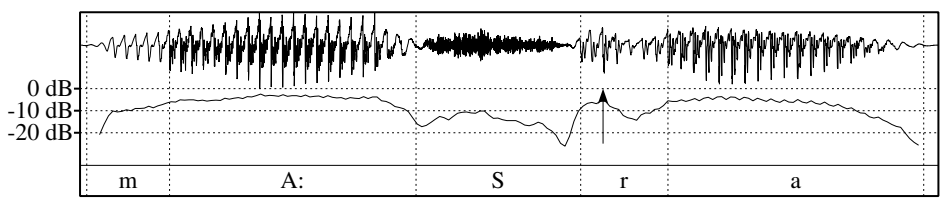
5.49. ábra. A *verve* (558 ms) szó szerkezeti elemei. Az [r]+C kapcsolatban teljes svá elem jön létre (nyíl)

A második lehetőség az [r]+[z] kapcsolatokra jellemző. Ilyenkor a közeli képzési hely miatt az [r] intenzitásminimum utáni zöngés, svá-szerű szakasza felépül ugyan, de a lecsengése már nem valósul meg, mivel folytatódik a homorgán zöngés mássalhangzó.

A harmadik szerkezeti formáció az [r]+[s t͡s t] kapcsolatokban jön létre. Ekkor svá elem nincs, a gerjesztésváltás már az r-ben is érezteti hatását, és a hang második felében zörejes elem keletkezik.

A negyedik megvalósulási forma az [r]+[j l] kapcsolatokban jön létre, itt svá elem nincs, a VCV kapcsolatokra jellemző intenzitásminimum jön csak létre, mivel két nagy intenzitású zöngés hang fogja közre az [r]-t.

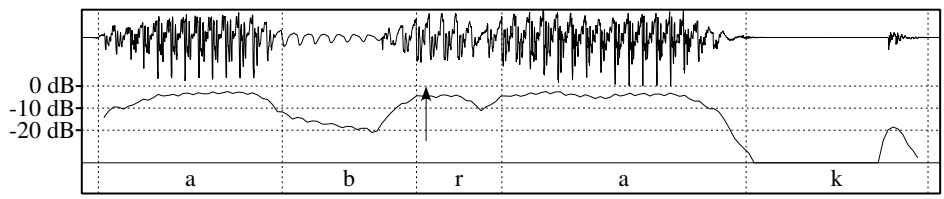
A C+[r] kapcsolatokban a korábban bemutatott szerkezeti formák hangelemei ugyanúgy létrejönnek, mint az [r]+C kapcsolatban, csak helyenként tükröződéssel. Itt három realizációs forma jöhet létre. A svá elem ebben a kapcsolatban is többféle formációt ölthet: a) a svá elem teljes egészében létrejön, b) a svá elem létrejön, de összeolvad az előző mássalhangzó zárfelpattanásával, c) a hangkapcsolat szerkezete hasonló a VCV kapcsolatéhoz. A első szerkezeti forma a leggyakoribb. Ekkor a svá elem teljes egészében létrejön, mégpedig az [r] hang elején, az intenzitásminimum előtt. Ez a formáció a zöngétlen mássalhangzók után jellemző. Ilyenkor a gerjesztésváltás miatt kötelezően kialakul a svá (5.50. ábra), a pergő hang ezzel kezdődik. A második szerkezeti forma akkor jön létre, amikor az [r]-t megelőző mássalhangzó



5.50. ábra. A *másra* (613 ms) szó szerkezeti elemei. Az [r] összes szerkezeti eleme megvalósul (svá és intenzitásminimum), ha zöngétlen hang előzi meg. A svá-t a nyíl mutatja

zöngés zárhang. Ilyenkor egy közös svá jellemzi a hangkapcsolódási pontot. Ezekben az esetekben a hanghatárt a svá belsejében célszerű kijelölni, jelezve ezzel, hogy

annak első részét a zárfelpattanáshoz számítjuk a többit pedig az [r] indulási részének tekintjük (5.51. ábra). A harmadik szerkezeti forma akkor jön létre, amikor az r-t megelőző mássalhangzó magánhangzószerű tulajdonságokkal bír, ilyenkor csak intenzitásminimum jellemzi a pergő hangot. Az [r] időtartama nem függ a CC kapcsolatban elfoglalt helyétől, a rá jellemző átlagos időtartam 54 ms, ami szignifikánsan hosszabb, mint egy VCV helyzetű pergő hang átlaga (38 ms). Az [r] jellemzően hosszabb, ha van benne svá elem. A hang időtartamsávja [r]+C helyzetben 36 ms-tól

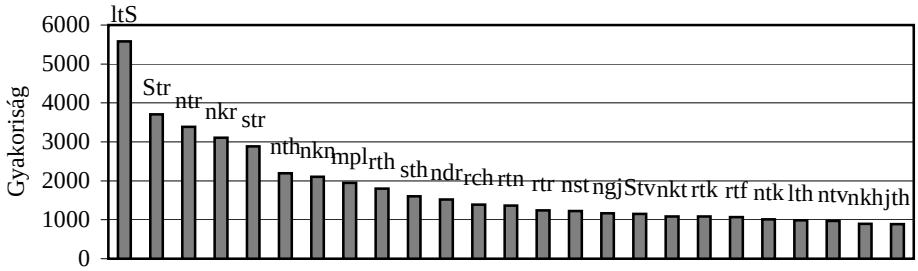


5.51. ábra. Az *abrak* (553 ms) szó szerkezeti elemei. A zöngés zárhang zárfelpattanásaként megvalósuló svá és az [r] kezdeti svá eleme összeolvad a kapcsolatban egyetlen svá elemmé

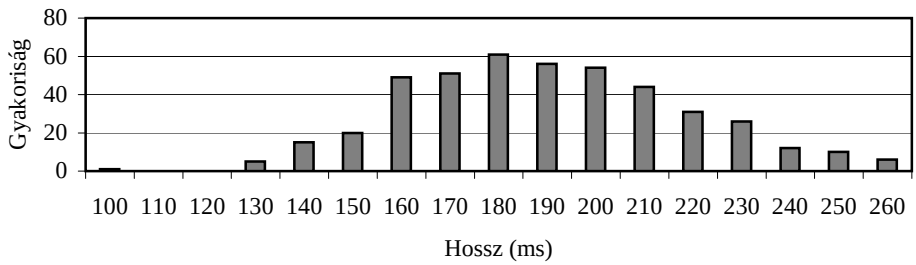
69 ms-ig terjed. Legrövidebb az [l] előtt, a leghosszabb, amikor [p] követi. A C+[r] helyzetben a hang időtartamsávja 41 ms-tól 61 ms-ig terjed. Legrövidebb az [j] után, a leghosszabb, amikor [z] előzi meg.

5.2.3.2. Három elemű mássalhangzó-kapcsolatok

A három elemű mássalhangzó-kapcsolatok még ritkábban fordulnak elő a nyelvben, mint a kételeműek. Az ilyen hangkapcsolódások szerkezeti felépítését és tulajdonságait Olaszy (2007b) munkájában lehet megtalálni részletesen, 445-féle hármas kapcsolatra. Az ilyen hangkapcsolatok előfordulási gyakorisági listájának első 25 elemét az 5.52. ábra mutatja. A vizsgálati adatok azt mutatják, hogy ezekben a kapcsolatokban az artikuláció az esetek többségében egyszerűsödik a CC kapcsolatokhoz képest. Ez leglátványosabban a kapcsolódások időszerkezetében mutatkozik meg. Míg a CC kapcsolatok eloszlási adataiból az ilyen kapcsolat teljes időtartamára 162 ms-os átlagot kaptunk, addig az összes hármas mássalhangzó-kapcsolatból számolva ez az érték 186 ms. A rövidülés tehát egyértelmű. A CCC kapcsolatok átlagos hossza alig 15%-kal haladja meg a CC kapcsolatokét. A CCC kapcsolatok hosszeloszlását 10 ms-os sávokra bontva az 5.53. ábrán szemléltetjük. A szórás 29,17. A vizsgált CCC kapcsolatok leggyakoribb építőeleme az [r], majd a [t] és az [n]. Az [r] gyakori előfordulása a pergetett hangszerkezetből ered, tulajdonképpen minden mássalhangzóhoz tud kapcsolódni, illetve mindegyik tud hozzá is. Érdekesség viszont, hogy az [r] csak C1, illetve C3 pozícióban fordul elő. Ez azt mutatja, hogy az [r] artikulációja megköveteli, hogy egyik szomszédja ne mássalhangzó legyen, két mássalhangzó kö-



5.52. ábra. A hármas mássalhangzó-kapcsolódások leggyakoribb 25 eleme

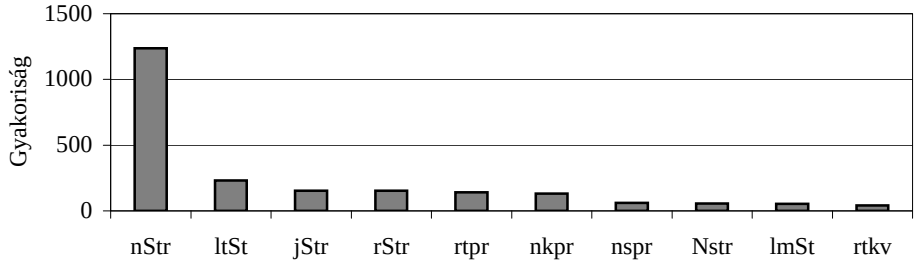


5.53. ábra. A CCC kapcsolódások teljes időtartamainak gyakorisági eloszlása

zött az [r] artikulációja olyan bonyolult, hogy a magyarban a hang nem tud létrejönni. A C2 pozícióban a leggyakoribb építőelem a [t]. A CCC kapcsolatok hangelemeinek tételes eloszlási listája a függelékben található.

5.2.3.3. Négyelemű mássalhangzó-kapcsolatok

Négy mássalhangzó kapcsolódása nagyon ritka a magyarban, összesen 74-féle ilyen kapcsolatról vannak egységes adatok (Olaszy 2007b). A vizsgált kapcsolatok gyakorisági listájában az [nʃtr] kiugróan vezet (5.54. ábra). Ennek az adatnak az ellenőrzésére kontrollmérést is végeztünk. Más szövegkorpuszon is ellenőriztük az előfordulási gyakoriságot. Az eredmény mindkét esetben egybevág az ábra adataival, tehát a négyes mássalhangzó-kapcsolódások első és második gyakorisági jelöltje között ténylegesen közel hatszoros a különbség. A négyelemű mássalhangzó-kapcsolatok időszerkezeti viszonyaival kapcsolatosan azt várnánk, hogy folytatódik a CCC kapcsolatoknál kimutatott rövidülési tendencia. Ez nem igazolódott. A vizsgált CCCC kapcsolatok átlagos hossza 234 ms. Ha a hangrövidülés tovább folytatódott volna, akkor arányosan számítva körülbelül 200 ms körüli értéket kellett volna kapnunk. Feltételezzük, hogy ez az eredmény annak a következménye, hogy az artikuláció már nemigen bírja el, hogy 4 mássalhangzót kell összekapcsolni. Míg a CCC kapcsola-



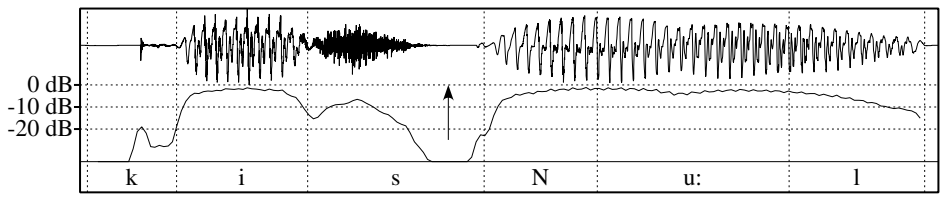
5.54. ábra. A vizsgált CCCC kapcsolatok előfordulási gyakorisági listájának első tíz eleme

toknál lazult az artikuláció, a CCCC kapcsolatoknál feszesebbé válik, ez lehet az oka, hogy nem rövidülnek a kapcsolat elemei. A négyes mássalhangzó-kapcsolatok teljes gyakorisági listáját a függelék tartalmazza.

5.2.3.4. A koartikulációs néma fázis jelensége

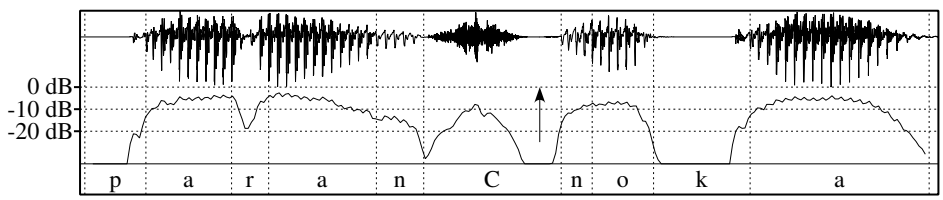
A koartikulációs néma fázis (KNF) mássalhangzó-kapcsolatokban jön létre, nevezetesen a nazális hang és az őt megelőző zöngétlen mássalhangzó zörejkomponense közé ékelődik egy rövid kis intenzitású hangszakasz (*más**sn**ak, esz**n**ek, ács**n**ak, kis**ny**úl*). Kialakulása csak kismértékben függ a beszélő artikulációs aktivitásától, az esetek nagy többségében létrejön. Időtartama 20–50 ms. A koartikulációs néma fázis mint hangelem nem kapcsolható a tradicionális hangleírások (zárhang, zár-rés hang stb.) egyetlen belső szerkezeti eleméhez sem, általában a zöngétlen mássalhangzóhoz tartozónak vallják (például a hanghatár kijelölésekor), tehát az időtartamok mérésénél ez a hangelem növeli az ilyen helyzetű zöngétlen mássalhangzók időtartamát. A KNF-hez hasonló jelenségről Stevens (1998) számolt be, és azzal jellemezte ezt a hangszakaszt, hogy tartalmazhat gyenge intenzitású zörejt is (i. m. 563. o.). Részletes vizsgálatokat a magyarra és más nyelvekre Olaszy (2007a) végzett a jelenséggel kapcsolatosan. A KNF kialakulását két tényező együttes hatásának tulajdoníthatjuk. A néma hangszakasz valószínűsíthetően két egyidejű artikulációs mozzanat összegzett eredménye, vagyis a gerjesztésváltásé és az azzal azonos időben létrejövő üregváltásé. Ez utóbbi során az orális üreget lezárjuk, a nazálist kinyitjuk. E két hangképzési elem együttes megvalósításához ezek szerint ennyi időre van szüksége a beszélőnek. A KNF mérésére Olaszy (2006a) állított fel három kritériumot: minimumhossz 10 ms, az intenzitás a hangszakasz közepén legalább 20 dB-lel alacsonyabb, mint a környező magánhangzók erőssége, a hangszakasz zöngétlen gerjesztésű. Ezekkel a küszöbértékekkel egyrészről egy hozzávetőleges viszonyítási alapot fejezünk ki, másrészről lehetővé tesszük a KNF egyértelmű meghatározását az időfüggvényen. Lássunk néhány példát a KNF jelenségére. Az első példában (*kis**ny**úl*) a KNF egy réshang-nazális mássalhangzó kapcsolódásában látható a réshang végén, a nazális

mássalhangzó rezgésének indulása előtt (5.55. ábra). Az [ʃ]+[ɲ] kapcsolatra vonatkoztatott hullámforma kép alapján egy megfordított affrikátára asszociálhatunk (a réselemet követi a néma fázis). Zár-rés hang is mutathat az előbbi példához hasonló rezgésképet, ha például egy adott CCC kapcsolat középső mássalhangzója (5.56. ábra). A *para**ncsn**oka* szóban a zár-rés hang szerkezete (időbeli lefolyása) szinte

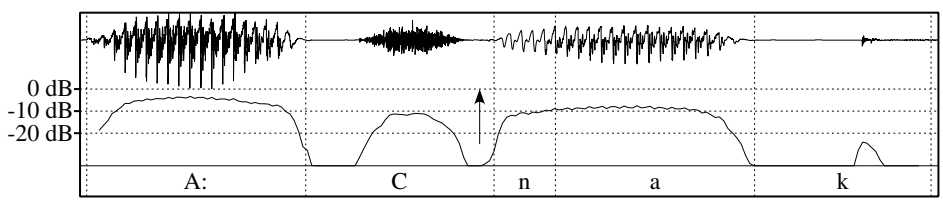


5.55. ábra. A *kisnyúl* (624 ms) szó szerkezeti elemei. A koartikulációs néma fázist a nyíl jelöli, időtartama a példában 41 ms

ugyanolyan képet mutat, mint az 5.55. ábrán. A zárszakasz itt nem jön létre az első nazális hang hatására. A koartikulációs néma fázis viszont létrejön a második nazális hang hatásaként, időtartama 35 ms. Amennyiben egy CC kapcsolatban zár-rés



5.56. ábra. A *parancsnoka* (884 ms) szó szerkezeti elemei. A zár-rés hang szerkezeti változása és a KNF együttesen a megfordított affrikátára emlékeztető rezgésképet eredményez



5.57. ábra. Az *ácsnak* (633 ms) szó szerkezeti elemei. A rezgésképében a zár-rés hang végén megjelenik a koartikulációs néma fázis, hossza a példában 31 ms

hangot követ a nazális mássalhangzó, akkor a rezgésképen két néma fázis szakasz is létrejöhet (5.57. ábra). A koartikulációs néma fázis jelenléte több kérdést vet fel, amelyek tisztázása a jövő feladata. Fonetikai szempontú vizsgálatok végezhetők ab-

ban a tekintetben, hogy mennyire fontos ennek az elemnek a léte a hanghullámban a percepció szempontjából. Vitát lehet nyitni arról, hogy a KNF melyik hanghoz tartozik. Gyakorlati szempontból felmerülhet, hogy a KNF-et tartalmazó hangot hogyan kell kezelni a beszédfelismerés és a beszédszintézis szempontjából (például: megtévesztheti a hanghatárokat kijelölő algoritmust, mint láthatjuk a 8.2.3. fejezetben).

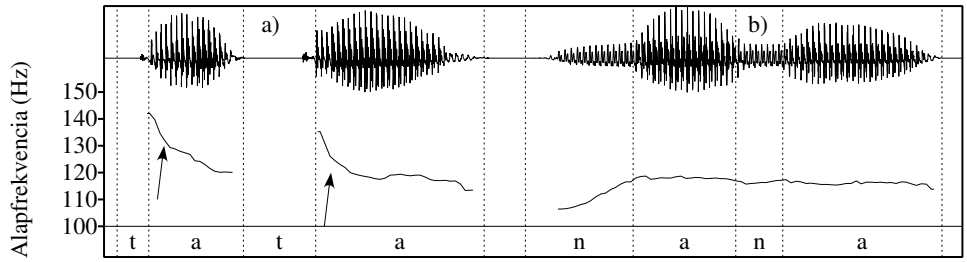
5.3. Szegmentális jelenségek a gége szintjén

A beszédhullámban minden gégeszintű történés nyomot hagy. A szegmentális szintű gégeműködés normál, egészséges fonáció esetén jól formált beszédhullámot hoz létre. Ebben a fejezetben olyan, eddig nem tárgyalt jelenségeket mutatunk be, amelyek szintén előfordulnak a természetes beszédben, és tudatos, illetve nem tudatos tevékenységből származnak. Egészséges gége és zöngeállás esetén jön létre a mikrointonáció és a glottalizáció. Mindkettő tipikusan szegmentális szintű változás. Amennyiben a gége fiziológiai állapota eltér az egészségestől (megfázás, gyulladás stb.), akkor létrejöhetnek olyan köztes működési állapotok, amelyek rekedtséghez vezethetnek. Szólunk a suttogott beszédről is, ami lehet tudatos is és kóros elváltozás következménye is.

5.3.1. Mikrointonáció

A mikrointonáció a hangszalagok zönge állású működésének kezdeti szakaszában jön létre. Maga az elnevezés is utal rá, hogy kismértékű alapfrekvencia-változásról van szó, ami csak pár zöngeperiódusnyi idő alatt zajlik le. A jelenség a következő: a hangkapcsolódások hangjainak belsejében a hangkapcsolatra jellemző néhány Hz-es (5–15) alaphangrezgés-ingadozás van jelen, általában a hangkapcsolódási pontokból kiindulva. Ez a változás automatikusan jön létre, még viszonylagos monoton beszédben is. Függ a személy fiziológiai adottságaitól, valamint a kapcsolódó hangoktól. A mikrointonációs változásoknak fiziológiai magyarázata van. A legjellemzőbb ilyen alapfrekvencia-mozgásokat gerjesztésváltáskor figyelhetjük meg CV kapcsolatokban (az 5.58. ábra a) képe).

Ilyenkor a zöngeképzés indulásakor és utána néhány periódusban a hangszalagokat nagyobb erővel szorítjuk össze, mint az ezt követő további szakaszban. Ez azt eredményezi, hogy nagyobb nyomás szükséges a szétnyitásukhoz, aminek következménye, hogy az első induló periódus magasabb frekvenciájú alaphangot hoz létre, mint a későbbiek. Az F_0 értéke néhány periódus után, fokozatosan és gyorsan csökkenve beáll a magánhangzóra jellemző alapfrekvencia-értékre. Amennyiben zöngés hangok találkoznak, ott nincsenek ilyen különbségek, hiszen a hangszalagok végig

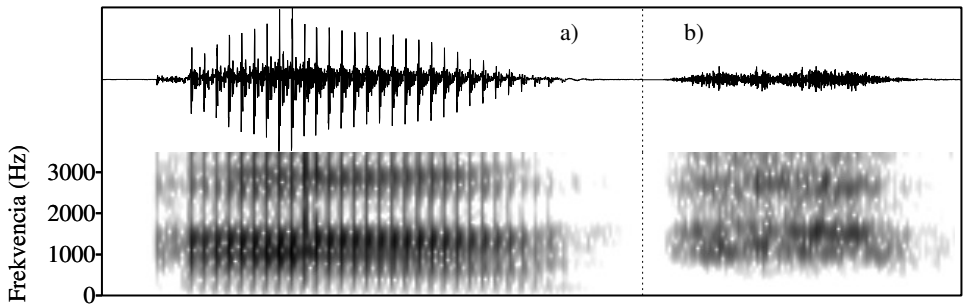


5.58. ábra. A *tata* (626 ms) és *nana* (658 ms) szavak szerkezeti elemei. A mikrointonációs változások a zöngétlen zárhang utáni magánhangzó elején figyelhetők meg (nyílak)

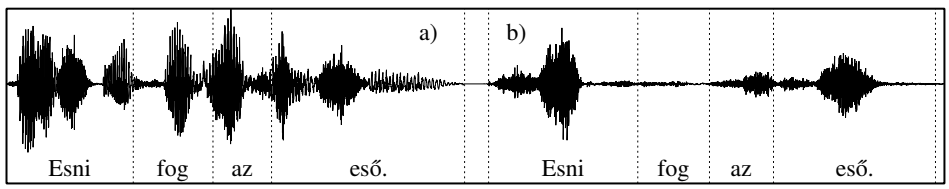
zönge állásban vannak (az 5.58. ábra b) képe). A mikrointonációs változásokat minden magyar CV kapcsolatra Olaszy (1989a) mérte meg és rendszerezte egyetlen beszélő ejtése alapján. A mikrointonáció szegmentális szintű alapfrekvenciaváltozás. A szupraszegmentális szinthez oly módon kapcsolódik, hogy ráépül a beszéddallamra, azon apró fűrészfogszerű formációkat hoz létre.

5.3.2. Suttogás

A suttogás többnyire tudatos tevékenység. Ilyenkor a beszéd gerjesztőjele csak turbulens áramlásból áll, a hangszalagokat folyamatosan suttogó állásban tartja a beszélő (3.1. fejezet). A suttogott beszéd jellemzői a normál beszédhez viszonyítva a következők. A zöngés hangok frekvenciaszerkezete megváltozik, felhangok nem lesznek benne, mivel a gégében keletkezett gerjesztési jel turbulens áramlással jön létre, tehát zörejes. A suttogási alapzörejben lévő frekvenciakomponensek kitöltik a 200–4000 Hz-es tartományt. Az artikulációs csatorna ugyanúgy működik, mint a normál beszédben. A zörej meghatározott frekvenciakomponensei a formánsok helyén felerősödnek, így hallhatóvá válnak például a suttogott magánhangzók. Tehát a suttogott beszédben is megtalálhatók a formánsoknak megfeleltetett úgynevezett zörejgócok. Erre mutat példát az 5.59. ábra, ahol egy normál és egy suttogott magánhangzó spektrális komponenseit hasonlíthatjuk össze. A normál beszédétől élesen különbözik a suttogás intenzitásszerkezete is. Míg a normál beszédben a magánhangzók képviselik a nagy energiájú hangokat, a suttogott beszédben ez ellentétes irányba fordul. A magánhangzók intenzitása lényegesen kisebb lesz, mint a zöngétlen réshangoké. Ez utóbbiak intenzitása nem fog változni a normál beszédhez képest, mivel képzésük ugyanúgy zajlik (5.60. ábra). A fizikai adatok szerint tehát a suttogott beszédben a zörejes réshangok dominálnak. Ezt azonban a percepcióban nem érezzük, kiegyenlítettnek, de halkabbnak halljuk az ilyen beszédet. A suttogott beszédre folyamatosan jellemző, hogy a hangereje alacsonyabb szintű, mint a normál beszédé. A különbség



5.59. ábra. Egy férfi ejtésű [a:] hang formánsai a) és a nekik megfelelő zörejgócok a suttogott ejtésű ugyanazon magánhangzóban b). Jól látható a két ejtésforma közötti intenzitáskülönbség is



5.60. ábra. Az *Esni fog az eső.* mondat normál ejtésben a) (1,14 s) és suttogva b). A két ejtésben élesen különböznek az intenzitásviszonyok. Ez főleg a magánhangzókon látszik

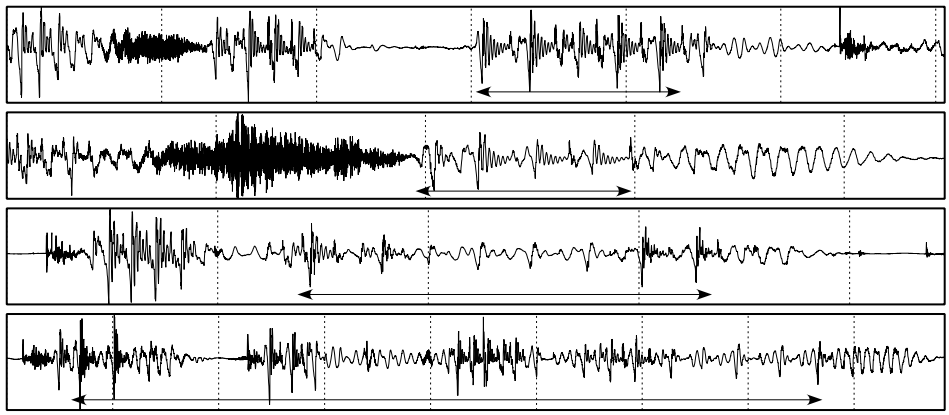
10 dB körül van. A hangerőt sem lehet olyan tág határok között változtatni, mint a normál beszédben (lényegesen hangosabban nem tud suttogni a beszélő). Másik jellemző, hogy a zöngés gerjesztés hiánya miatt a hangszínezet kialakulása elveszik, nem lehet felismerni a suttogó személyt a hangja alapján. Mindezekből következik, hogy a suttogott beszéd gépi felismerése a normál beszédre kidolgozott módszerekkel nem végezhető el. A beszédszintézisben pedig csak olyan módszerekkel lehet suttogott beszédet előállítani, amelyeknél külön van választva a gerjesztés és az azt módosító szűrőrendszer. Ilyen eljárások a formánsszintézis és az olyan statisztikai módszerek, amelyek paramétersorozatból generálják a beszédet egy kódoló visszaállító segítségével (HMM-alapú és LPC).

5.3.3. Irreguláris zöngeképzés, glottalizáció, rekedtség

Bőhm Tamás

Reguláris fonációnál a hangszalagok kváziperiodikusan mozognak (felnyílnak, majd összezáródnak). Irreguláris zöngeképzésről akkor beszélhetünk, ha a periodicitástól való eltérés nagymértékű, és ez az eltérés ép hallású emberek számára határozot-

tan érzékelhető. Néhány példát láthatunk irreguláris zöngével képzett beszédhangrészletre az 5.61. ábrán. Az irreguláris fonációjú beszédet érdes, rekedtes hangként észleljük (Hollien–Wendahl 1968). A fenti meghatározás az olyan irregularitásokra



5.61. ábra. Példák irreguláris zöngére (nyíllal jelzett szakaszok) egy magyar férfi beszédében. A kiejtett szavak fentről lefelé: *az ebéd*, *eső*, *(Ka)talinnak*, *(ku)tyákat bevinni.* A vízszintes tengelyeken a függőleges szaggatott vonalak 100 ms-os osztást jelentenek

összpontosít, amelyek a hangszalagok irreguláris mozgásából erednek – így például nem tekintjük irreguláris fonációnak a levegős zöngét, amikor nincs tökéletes összezáródás a hangszalag valamely részénél és ezért ott surlódási zörej keletkezik, ami szuperponálódik a zöngehangra. Ahhoz, hogy egy beszédrészletet irregulárisnak tekintsünk, egyrészt a rezgésképen/spektrogramon jól látható eltéréseket kell mutatnia a normál zöngétől, másrészt hangszínezetben hallhatóan különböznie kell tőle (Dilley–Shattuck-Hufnagel 1996).

Valójában az irreguláris zönge egy gyűjtőfogalom, amely például Titze (1995) zöngeminőség-rendszerében magában foglalja többek között a diplofóniát, az aperiodicitást és a periódusduplázódást. Számos kutató különböző irregularitáskategóriákat különböztet meg (Redi–Shattuck-Hufnagel 2001, Batliner et al. 1993, Hedelin–Huber 1990). A szakirodalomban a jelenséget recsegő, érdes, rekedtes, nyikorgó zöngének, vagy éppen glottalizációnak nevezik. Az angol szaknyelv is számos különböző kifejezést használ a jelenségre (creaky voice, pulsed phonation, vocal fry, laryngealization, glottalization). Az irreguláris fonáció produkcióját gyakran a hangszalagok szoros összeszorításával magyarázzák, ami a rezgést instabillá teszi. A tüdőből kipréselt levegő ezt az erős zárat csak ritkábban és rövidebb ideig tudja felfeszíteni, és akkor se teljes hosszában. Így a reguláris fonációhoz képest jelentősen kevesebb levegő áramlik át a hangrésen egy időegység alatt (Laver 1980). Ilyen állapot fordul elő a közlések végén, amikor a levegőnyomás jelentősen csökken. A patológiás hangszalag-elváltozások (például aszimmetrikus működés, csomó, bénulás)

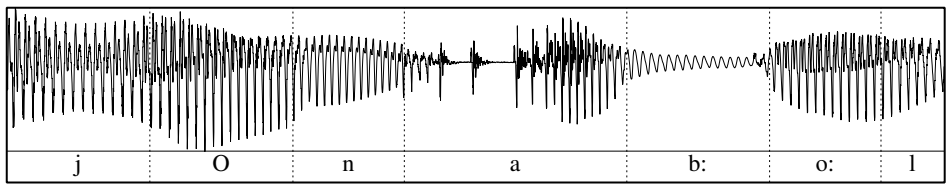
egyik lehetséges tünete az állandósult irreguláris hangszalagmozgás (Hirano 1981). A normális, egészséges hangszalagokkal képzett beszédben előforduló irregularitás csak időszakosan jelentkezik, de sokrétű szerepe lehet a közlésfolyamatban a következők szerint.

A nyelvi üzenet (szegmentális és szupraszegmentális) kiegészítője. Egyes nyelveken az irreguláris fonáció fonológiailag kontrasztív: például a mexikói jalapa mazatecben (Gordon–Ladefoged 2001). A dánban a jelentésmegkülönböztető irreguláris fonációt stodnak hívják (Laver 1980). Az amerikai angolban zöngétlen felpattanó zárhangok allofónja lehet (Pierrehumbert–Talkin 1992). Szintén az amerikai angolban prozódiai szerepe is lehet: gyakran előfordul intonációs frázishatárokon és hangsúlyos szókezdő magánhangzókban (Dilley–Shattuck-Hufnagel 1996).

Érzelmi állapotok jelzője. Fónagy–Magdics (1967) többféle érzelmet kifejező beszéd esetén tapasztalt érdes, rekedtes zöngét. Gobl–Ní (2003) formánsszintézissel kimutatta, hogy az irreguláris fonáció megváltoztathatja az észlelt érzelmi töltetet.

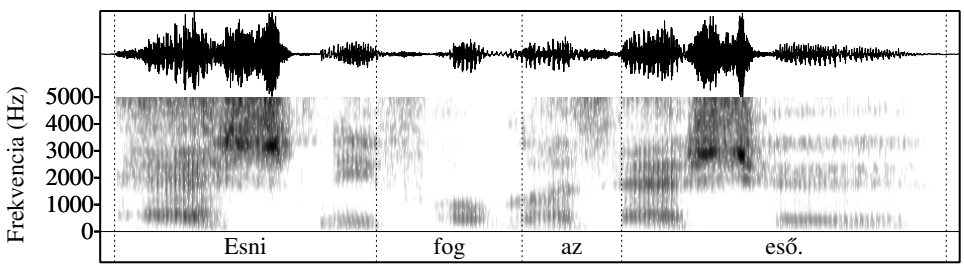
A beszélő egyéni jellegzetessége. A szakirodalomban számos szerző megfigyelte, hogy az irreguláris zönge előfordulásának gyakorisága beszélők között nagymértékben eltérhet. Slifka (2006) a kísérletében a négy adatközlő magánhangzóra végződő bemondásainak végein 0%, 51%, 85% és 85%-ban talált irreguláris zöngeképzést. Redi–Shattuck-Hufnagel (2001) tanulmányukban a 14 amerikai személy közül volt, aki a vizsgált pozíciók 88%-ában képzett irreguláris zöngét és volt, aki csak 13%-ában. Dilley–Shattuck-Hufnagel (1996) öt rádióbemondó szókezdő magánhangzóit vizsgálták, 13% és 44% közötti arányról számoltak be. Henton–Bladon (1987) a 79 brit beszélő közül 10 esetében gyakran tapasztalt irreguláris fonációt, míg másoknál csak ritkán (kvantitatív adatokat nem közöltek). Markó (2005) egy magyar adatközlő spontán beszédében gyakran, a többi háromében csak néhányszor észlelt „nyikorgó zöngét”. Egy másik tanulmány (Bőhm–Ujváry 2008) kísérletében egy szöveget 12 magyar beszélő háromszor olvasott fel. A beszélők között jóval nagyobb különbségeket találtak az irreguláris zönge előfordulási arányában, mint egy beszélő három felolvasása között (a különbségek a mondatok végén fokozottan jelentkeztek).

Az irreguláris zöngeképzés a magyarban is gyakran előfordul, még képzett beszélőknél is. Markó (2009) adatai azt mutatják, hogy magyarban különbségek lehetnek az irreguláris zönge előfordulásában olvasott és spontán beszédmódok között. Beszédtechnológiai szempontból a beszédfelismerő szoftvereket zavarhatja meg, hangtévesztéshez vezethet. Egy színésznő felolvasásos hangfelvételéből mutatunk be egy részletet, amelyben a folymatos felolvasás közben detektált irreguláris zöngeképzés látható (5.62. ábra). Beszédkutatási célokra elérhető egy glottalizáló számítógépes program, amelyik reguláris zöngével képzett beszéd kijelölt részét képes irregulárissá alakítani (`http://www.bohm.hu/glottalizer.html`). A rekedtség az irreguláris zöngeképzés állandósult állapota. A rekedt beszéd hangzása jellegzetes, se nem zengő, se nem suttogó. A zönge periodicitása megszűnik, azonban a gége zöngeállásban működik, csak rendezetlenül (ezért nem alakul ki a suttogás). Általában betegség



5.62. ábra. Irreguláris zöngeképzés a *jön abból* (580 ms) szókapcsolat [ɔ] hangjában

okozza a gége rendellenes működését. A rekedt beszédre jellemző a rezgésamplitúdó rendezetlensége mind az időben, mind az intenzitásban. A frekvenciaszerkezetre nem jellemző az egységes szerkezet a magánhangzókban, a frekvenciakomponensek a gége működésének pillanatnyi történései szerint változnak. Ez detektálható a hangspektrogramon is (5.63. ábra). A rekedt beszéd gépi felismerése valószínűleg



5.63. ábra. Az *Esni fog az eső* (1,7 s) mondat szerkezeti elemei rekedt ejtésből. A rekedt beszédet irreguláris rezgésforma és rendezetlen spektrális kép jellemzi. Ez a hanghullám szerkezetén is érzékelhető (fenn), illetve a hangspektrogramon is (lenn)

nehezebb, mint az egészséges beszédé, hiszen a rekedtség miatt bizonyos mértékig (a rekedtség fokától függően) hiányoznak belőle a normál beszédre jellemző spektrális összetevők. A beszédszintézisnél is lehetnek fokozatok abban a tekintetben, hogy mennyire élethű rekedt hangot akarunk előállítani. A legegyszerűbb megoldás, amely már közel rekedtes hangot szolgáltat, ha az f_0 paramétert 40 Hz-re állítjuk. Ilyenkor már olyan ritkák lesznek a zöngeimpulzusok, hogy a hallgatót a rekedtségre emlékezteti.

6. fejezet
A beszéd szupraszegmentális szerkezete

Olaszy Gábor

A beszéd „hangszerelését” (Szende 1995) a beszélő személy az úgynevezett szupraszegmentális eszközrendszerrel hozza létre. Ez azt jelenti, hogy a hangzást színesítjük, érzékeltetjük a mondandónk jelentését, azaz hangsúlyozással, tempóváltással, szünetekkel, dallammenetekkel, hangszínezetváltásokkal könnyítjük meg az értelmezést. Ezt az eszközrendszert azért nevezik szupraszegmentálisnak, mert elméletileg az alapvető szegmentumokra (hangsorozat) építi rá a beszélő. A beszédképzés során azonban mind a szegmentális, mind a szupraszegmentális elemeket egyazon pillanatban hozzuk létre, ugyanazokkal az eszközökkel. Vizsgálódási szempontból viszont van értelme a külön tárgyalásnak. Szét is választhatjuk a beszéd eme két elemét (például célzott szűréssel), és beszédhangoktól (szegmentális szerkezet) elkülönülten is vizsgálhatjuk, mondjuk a beszéddallamot (Gósy–Terken 1994). A szupraszegmentális eszközrendszert nagyobb nyelvi egységen (mondat, szöveg) alkalmazzuk, tehát más időszerkezeti keretek vonatkoznak rá, mint a szegmentális szerkezet elemeire (beszédhangokra és azok kapcsolódásaira). A szupraszegmentális eszközrendszer – más néven prozódia – elemei a következők: a beszéddallam, a hangsúlyozás, a ritmikai elemek széles tárháza és a hangszínezet esetleges változtatása. Ezek a beszédtulajdonságok tehát nem származtathatók a közlést alkotó beszédhangok sorozatából, azok felettinek tekintjük őket. A prozódia elemei elvont síkon univerzálisak (például minden nyelv használ dallamokat), konkrétan azonban nyelvfüggők, jellemzők a nyelvre. A szupraszegmentumok alkalmazásával kifejezhetünk modalitást, szintaktikai, szemantikai és pragmatikai információt, érzelmeket, tehát nyelvi szintű és egyéni megformálásokat is megkülönböztethetünk. A hangsúly megléte, illetve hiánya egy adott szón megváltoztathatja a mondat értelmét. A hangszínezet megfelelő megválasztásával kifejezhetünk érzéseket (szeretet, utálat). A mondatdallam bizonyos esetekben a kérdéses mondaton túl is hathat, főleg dialógusszituációban előre jelezheti a következő mondat típusát. A hangutánzó művészek többnyire az utánzott személy prozódiai sajátosságait imitálják (Gósy 1999). A beszélő több szupraszegmentális elem együttes alkalmazásával tudja kialakítani a megfelelő je-

lentéstartalmat. Felolvasás esetén a prozódiai tagolást a szöveg központozása (azon belül a szintaktikai szerkezet) némileg meghatározza. A tagolási határok közötti részeket nevezzük prozódiai frázisnak, két hangsúlyozott szó közötti részt pedig hangsúlyfrázisnak. Egy mondaton belül tehát lehet több prozódiai frázis, és azokon belül pedig több hangsúlyfrázis is. Spontán beszéd esetén a szupraszegmentumokat másképpen használjuk, mint felolvasásnál. Beszédtechnológiai szempontból másképpen osztályozunk. A kontrollált prozódiát alkalmazzuk a mai modellekben, azt a változatot, amit gondos beszéd során hozunk létre, többnyire felolvasásból (hírek, mese, rádiójáték). A jövő kutatási iránya a szabadabb prozódia modellezése (spontán vagy szociális prozódia), amelyet a mindennapi életben a beszélgetésekben hozunk létre. Ez utóbbiban a prozódiai elemek tárháza kitágul a hangulati elemekkel. Nevetés, hümmögés, hangulathangok, mint cuppogás, csettintés és még sok más akusztikai elem fejezhet ki a szituációt (Campbell 2005, 2007b). A jövő beszédkutatása fogja csak feltárni azokat a jellegzetességeket, amelyek az ilyen beszédre jellemzők.

Alább összefoglaljuk a felolvasásos és a gondozott beszéd vizsgálataiból kapott általános, mondhatjuk úgy is, hogy invariáns prozódiai eszközök jellemzőit a magyar beszédre. Tisztában vagyunk azzal, hogy a spontán dialógusban megvalósuló beszéd prozódiája sok szempontból eltérhet a megadottaktól, beszédtechnológiai vonatkozásban azonban csak a mért adatokra alapozhatók azok a modellek, amelyek jelenleg segítik akár a beszéd gépi előállítását, akár a beszéd gépi felismerését és egyéb gépi alkalmazásokat. A spontán megnyilatkozásokkal kapcsolatos kutatások fonetikai szinten már folynak. Ebben a könyvben is érintünk egy témát, ami idesorolható (például az érzelmek gépi detektálását a 9.11. fejezetben). A prozódia elemei gondozott beszédre vonatkozóan a következő paraméterekkel jellemezhetők:

A beszéddallamot egyetlen paraméterrel leírhatjuk, az alapfrekvenciával, amelyet az időben változtatunk.

A hangsúlyt a hangra jellemző három fizikai paraméterrel jellemezhetjük, az alapfrekvenciával, az intenzitással és az időszerkezettel. A frekvenciaszerkezettel hozható kapcsolatba az alapfrekvencia kiemelkedése (csúcs). Az esetleg erősebb hangintenzitás is hozzájárulhat a hangsúlyélmény kialakításához. Az időszerkezeti tényezőkhöz tartozhat az alapfrekvencia-csúcs helye a hangsorban, az alapfrekvenciaváltozás belső időszerkezete, a hangok nyújtása és az esetleges szünettartás a hangsúlyozandó szó előtt. A prozódia legbonyolultabb eleme tulajdonképpen a hangsúly, amelyben három fizikai paraméter kombinációja alakítja ki a kívánt hangzást. A hangsúly kutatása a mai beszédtechnológia fontos területe, annak ellenére, hogy azt már a 20. század első felétől folyamatos vizsgálják.

A ritmika jellemzéséhez időszerkezeti tényezőket használunk, ilyenek például a beszédhangok időtartamának változtatása adott beszédszakaszokon (gyorsabb-lassabb artikuláció) és szünetek beiktatása is.

A hangszínezetet a beszédhangok spektrális komponensei határozzák meg. Ezek pedig egyrészről az artikulációval és a mimikával függnek össze, másrészről a ger-

jesztési jel tudatos megváltoztatása is hat a spektrumra (préselt gégehang, levegős zönge, suttogás stb.). A kettő együttes hatása a megszokottól eltérő hangzást eredményezhet, akár szituációs állapotot is jelezhet.

A magyarra (és más nyelvekre) a gondozott beszéd és a felolvasás prozódiai alapelemeinek feltárása már legnagyobbrészt megtörtént (Hirst–diCristo 1998, Varga 1994, Olaszy 1995b). Az egyéni produkció és a nyelvi tartalom közötti összefüggések beszédtechnológiai szempontból is fontos területet képeznek. A beszédprodukcióból származó beszédjelben benne vannak a beszélő személy egyéni megformálásai is, amit a felolvasott szövegtípus is meghatároz. Másképpen olvasnak híreket és másképpen egy novellát vagy egy mesét. Az ilyen típusú kutatások csak az utóbbi évtizedekben élénkültek meg (Elekfi–Wacha 2003, Olaszy 2005). Az egyes szövegtípusok vonatkozásában végzett statisztikai vizsgálatok kimutatták, hogy a prozódiai jellemzők nyelvfüggők, ezen belül pedig szövegtípusfüggőek (Fackrell et al. 2000).

A prozódia – hasonlóan a beszédhanghoz – a pillanatnyi beszédprodukcióhoz tartozik, a beszélő nem tud két egyforma dallamformát produkálni, a létrehozott prozódiát csak nyelvileg lehet ugyanolyannak tekinteni, még akkor is, ha ugyanazt a mondatot többször ejti ugyanaz a beszélő. Az ilyen variabilitást – ami a biológiai rendszer pillanatnyi állapotaiból adódik – nehéz modellezni a beszédtechnológiai felhasználásokban. Ezzel kapcsolatos kísérleteket Németh és Csapó (Németh et al. 2007c) kezdeményezett. A prozódia általános modellezésének fontos szerepe van mind a beszédszintézisben, mind pedig a gépi beszédfelismerésben is. A magyar beszédre már készítettek olyan algoritmusokat, amelyekkel gépi beszédelőállításnál generálni lehet a prozódia mindhárom elemét (Olaszy et al. 2001, Olaszy 2002).

6.1. A beszéddallam

Alapvetően a gondozott, felolvasásos beszéd dallamformáit vizsgáljuk, de röviden érintjük a spontán beszéddel kapcsolatos kutatásokat is. A beszédben az alaphang frekvenciaértékének hosszabb közlési egységre vonatkoztatott változásait hívjuk beszéddallamnak. A hosszabb közlés kategória pontosabban mondatot, illetve a mondaton belüli prozódiai frázist jelent. Ebből következik, hogy a dallamnak valamiféle határjelző szerepe is van. Ugyanakkor a spontán beszédben ezek a kategóriák lazábban értendők. Gósy (2003) a beszéddallam határjelölő szerepét vizsgálta egy kísérletben. A résztvevők a mondatvégek jelölésében elsődleges akusztikai kulcsként a szünet megjelenését és még inkább annak hosszát használták fel, és csak másodlagosan támaszkodtak az alapfrekvencia-változásra.

A beszéddallam fontos része a beszédnek. A beszélő személy a hangszalagok közreműködésével létrehozott zöngehang frekvenciaértékét változtatja meg, többnyire tudatos megformálással. A dallam formáját a nyelv és a beszélő egyéni aka-

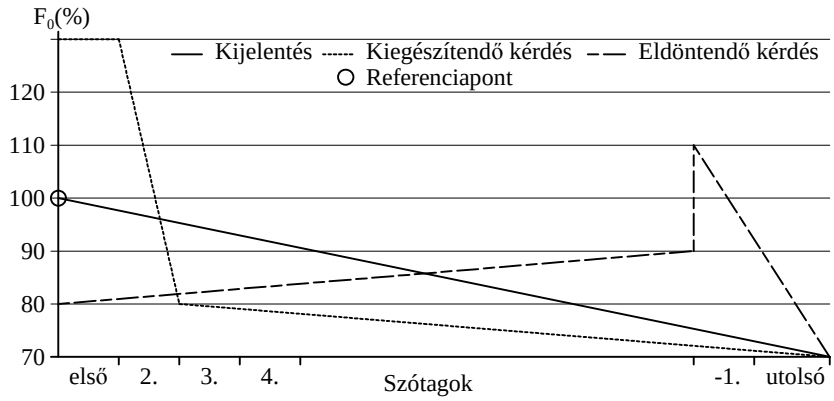
rata határozza meg, a konkrét alapfrekvencia-értékeket pedig leginkább az egyén hangfekvése. A beszédképzéskor a tüdőből jövő levegőnyomás nagysága és az alapfrekvencia értéke egyenes arányban függ össze. Ha nagyobb a nyomás, akkor magasabb az alapfrekvencia is. A magasabb alapfrekvenciájú hullámformarészek tehát nagyobb jelamplitúdóval jönnek létre, mint az alacsonyabbak. Másként megfogalmazva azt mondhatjuk, hogy a magas F_0 érték intenzitásnövekedéssel párosul. Az alapfrekvenciának alapvetően kettős funkciója van a beszédben: dallamot alakít ki és hozzájárulhat a hangsúly megvalósításához is (a mikrointonáció szegmentális szintű változás, nyelvi funkciója nincs). Abban a speciális esetben, mikor a mondat rövid, benne csak egyetlen magánhangzó van (*Ő., Ő?, Én., Én?, Én!*), a hangsúlyozási és a dallamfunkció nehezen választható szét. Az alapfrekvencia elsődleges szerepe a magyarban a beszéddallam létrehozása, a hangsúlyozásban való részvétele csak másodlagos (Gósy 2004b). A felolvasásos, illetve a gondozott beszédben a beszéddallammal főleg a mondatok modalitását (kijelentő, kérdő, óhajtó stb.) fejezzük ki, de spontán beszédben számos más funkciója is lehet, például szituáció érzékeltetése. Az alapfrekvenciához kapcsolódik még két fogalom, a beszélő egyéni hangfekvése, valamint a hangterjedelem. A hangfekvés egy átlagos frekvenciaértékkel hozható kapcsolatba. Azt fejezi ki, hogy a beszélő általánosságban milyen alapfrekvenciával beszél. Ez fiziológiai adottságokra vezethető vissza, így a férfiak hangfekvése átlagosan mélyebbnek tekinthető, mint a nőké, a gyermekek hangfekvése pedig relatíve magas. A hangterjedelem az a frekvenciatartomány, amelyen belül a beszélő az alapfrekvenciáját változtatja beszéd közben. A hangterjedelem változhat a szándék és a témakör függvényében. Mesemondáskor például nagyobb a hangterjedelem, mint egy novella felolvasásakor, vagy egy híroblokk elmondásakor létrejövő beszédben. A hangterjedelem nyelvfüggő is. A magyar nyelv szűkebb hangterjedelmű, mint például az angol (Varga 1994). A beszéddallam teljes vonulata a dallamforma, amely felfogható építőkockák sorozataként. Ha az építőkockákat sematizáljuk, akkor jutunk el az alapszintű dallammenetekhez. Ezek a kategóriák leíró jellegű rendszerezések. Háromféle ilyen elvi építőkocka létezik: emelkedő, ereszkedő és lebegő (szinttartó) dallammenet. A fonetikai szakirodalom még hozzáteszi ezekhez a szökő (rövid idő alatt sokat emelkedik), illetve az eső (rövid idő alatt sokat csökken) dallammeneteket. A mondat teljes dallamformáját tehát dallammenetek sorozataként is elképzelhetjük. Attól függően, hogy az elvi építőkockákon belül az egyes dallamváltozások mennyi idő alatt zajlanak le, illetve hogy milyen frekvencián szinttartóak, végtelen számú ilyen építőkocka van jelen a beszédben. Az egyes nyelvekben az egyes mondatformákra meghatározott dallamformák (dallammenet-kombinációk) a jellemzőek (Hirst–diCristo 1998). A beszéddallam fizikailag csak a zöngés hangokban van jelen, a zöngétlen hangok ejtésekor megszakad. Ennek ellenére a hallgató a dallamformát egységesen és folyamatosan jelen lévő akusztikai jellemzőként fogja fel, és ez alapján azonosítja például a mondat modalitását.

A magyar beszéd dallamformáiról az egyik legrészletesebb fonetikai munka Fónagy–Magdics (1967) összefoglalója, amelyben hallásalapú lejegyzéseiket összegezték. A dallamok vonulatát kottázással adták meg, ezért értelmezésük nehéz. Hasonló munka még Elekfi–Wacha (2003) könyve, amelyben az értelmes beszéd megvalósításához szükséges mondatfonetikai eszközöket ismertetik leíró formában, sok példával. A magyar nyelv dallammeneteinek fonológiai rendszerezését Varga (1994) dolgozta ki. Ez a rendszer értékes elméleti összefoglalás, beszédtechnológiai algoritmizálásra azonban nem alkalmas, mivel olyan kifejező információkra is támaszkodik (feszült figyelem, izgalom stb.), amelyek fizikai háttere még nem tisztázott. A fonológiai jelzések elhelyezését a szövegben ezért csak szakember képes elvégezni, mivel ahhoz egy komplett szintaktikai mondatelemzésnek és értelmezésnek is társulni kell (ilyen algoritmus még nem áll rendelkezésre). A magyar beszédre vonatkozó fizikai alapú dallammenetek létrehozására Olaszy (2001a) dolgozott ki modellt (beszédszintetizátorok vezérlésére), amely független a hangfekvéstől. Készített továbbá egy olyan transzformációs fonetikai szabályrendszert is (nem algoritmust), amellyel fizikailag is megvalósíthatóvá váltak Varga fonológiai szintű intonációs jelzései, amelyeket a szövegbe jegyzett. A szintézissel létrehozott mesterséges beszédmintákkal végzett percepciós tesztek igazolták, hogy a fonológiai szabályrendszer az elvárt dallamélményt adja (Olaszy 2001b). A magyar beszéd dallamában két kulcsfontosságú pontot lehet meghatározni a dallam elméleti jellemzéséhez. Az első a mondat kezdésekor jellemző alapfrekvenciaértéke, a másik a befejezésre jellemző érték. A befejezési érték invariáns jegynek tekinthető. A befejezett közlés dallammenetében minden esetben a legvége van a legalacsonyabb alapfrekvencián (nem spontán beszédben), és ennek Hz-értéke nemigen változik, a beszélőre jellemző (főleg a hangfekvés befolyásolja). A mondat kezdete és vége közötti viszony adja a jellemző elméleti dallamformát. A magyar mondatformák többségében a kiindulási alapfrekvencia magasabban van, mint a befejezési, ezt a jellemző elméleti változást az ereszkedő jelzővel fejezik ki (Gósy 2004b). Az, hogy a két végpont között milyen konkrét alapfrekvenciamozgások zajlanak le, a mondat fajtájától, összetettségétől, illetve a beszélő akaratától függ. A magyar spontán beszédre vonatkozó beszéddallam-kutatások is kezdenek élénkülni. Ezek mindenképpen támaszkodnak a gondozott beszéd vizsgálati tapasztalataira. A dallamnak a hümmögésben játszott szerepét is vizsgálták, és spontán beszéden végzett vizsgálatok eredményei alapján fogalmaztak meg megállapításokat (Markó 2006). Egy, a magyar spontán beszéd dallamrealizációira vonatkozó tanulmányban (Markó 2007) közel kétszáz kérdő megnyilatkozás elemzésének eredményei alapján megállapították, hogy a spontán megvalósuló kérdések sokszínűbbek, mint amilyennek a felolvasott, elicitált vagy eljátszott mintapéldáik alapján a szakirodalom bemutatja őket. A változatosság elsősorban az F_0 modulációjának lehetőségeiben, másrészt pedig a hangköz értékeiben mutatkozik meg. Beke (2008) a spontán beszéd és a felolvasás dallami realizációit vetette össze, egy másik vizsgálatban (Markó 2009) pedig a stigmatizált úgynevezett szökőzár megjelenését vizsgálta ol-

vasásban és spontán beszédben. Váradi (2008) különböző spontánbeszéd-műfajokat vetett össze percepciós szempontból, és azt találta, hogy a narratív és leíró szövegekben eltérő a dallam határjelző szerepének a megvalósulása.

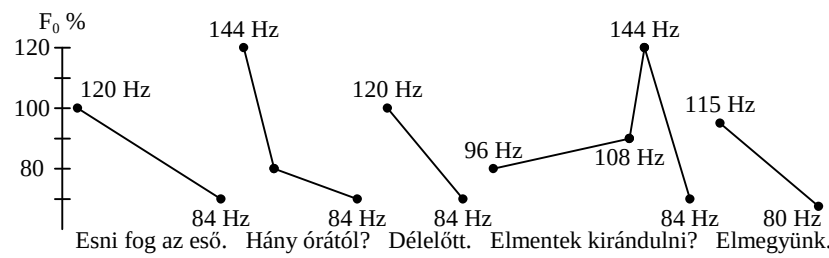
6.1.1. A mondatdallamok kapcsolódási rendszere

Mielőtt a magyar mondatokra jellemző dallamformákkal foglalkoznánk, szólnunk kell egy magasabb szintre vonatkozó összefüggésről, a mondatdallamok kapcsolódási rendszeréről. A mondatdallamok folyamatos beszédben és dialógusban is meghatározott nyelvi szabályok szerint kapcsolódnak egymáshoz. A kijelentés után esetleg kérdés következik, utána válasz egy kijelentés formájában, vagy viszontkérdezés stb. A kommunikáció mondatai tehát sokféle kombinációban követhetik egymást. Mindezekből következik, hogy létezik egy magasabb szintű dallamkapcsolódási rendszer is a beszédben, a mondatok közötti. Ennek jellemzőit egyrészről a nyelvi rendszer, másrészről a kommunikációs forma határozza meg. A beszédprodukcióban az egyes mondatokra jellemző egyedi dallamformákat hozzáigazítjuk a megelőző és a követő mondatokéihoz. Az ilyen illesztést a beszélő személy automatikusan végzi. Kérdés, hogy milyen szabályok alapján végezzük ezt az illesztést. A magyarra vonatkozó dallamkapcsolódások ilyen rendszerét Olaszy (1995a) vizsgálta, párbeszédek dallamformáinak elemzésével. Univerzális modellt alakított ki, amelyben egyetlen referenciapont segítségével meghatározhatók a dallamkapcsolódások és az egyedi dallamformák jellemző pontjai. A referenciapontnak megfelelő Hz-érték a hangfekvés függvényében változhat, így konkrét dallamformák is meghatározhatók. A referenciapont ebben a modellben a kijelentő mondat kezdési pontjára jellemző frekvenciaérték, mivel ez a mondatfajta fordul elő a leggyakrabban a kommunikációban. A modellből kiszámítható (akár a szöveg alapján, például beszédszintézisnél), hogy két mondat kapcsolódási pontján milyen lesz majd az alapfrekvencia várható értéke, hogy a hangzás a beszéddallam folytonosságával szemben támasztott percepciós elvárásoknak eleget tegyen. Vannak modellek, amelyek a mondat zárására jellemző frekvenciát használják referenciának. A magyarra kidolgozott dallamkapcsolódási modellt a 6.1. ábra mutatja. A modell relatív adatokat ad meg a mondatfajták kezdő és befejező pontjára, valamint a kérdések belső szerkezetére és a dallamcsúcsot tartalmazó szótagra. Itt jegyezzük meg, hogy a fejezet további részének ábráiban a referenciapont frekvenciaértéke (a bemondóhoz igazítva) 120 Hz, ezt minden ábrán feltüntetjük. Ezzel azt érjük el, hogy az ábrák dallamgörbéi egymással és a modellel is összevethetők. Az átfogási sáv a referenciaponthoz viszonyítva plusz-mínusz 30%. Látható, hogy a magyarban a mondat végpontja egységesen a legalacsonyabb frekvencia a dallammenetben, a kezdőpont pedig erősen változik a mondat fajtájától függően. Áttételesen azt is kiolvashatjuk az ábrából, hogy a megelőző mon-



6.1. ábra. A kérdések és a kijelentés sematikus dallamformái és kapcsolódásaik rendszere a kijelentéshez viszonyítva a szótagok függvényében. A referenciapont a kijelentés kezdeti F_0 értéke (például 100% = 120 Hz)

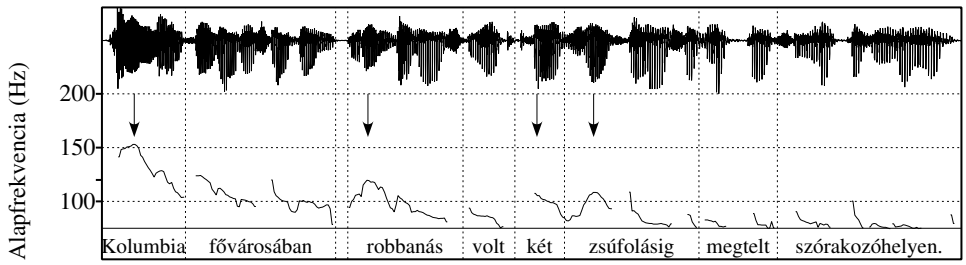
dat tükrében jósolni lehet a következő mondat modalitását, anélkül, hogy az egész mondatot meghallgatnánk. Ez beszédtechnológiai szempontból segítheti például a beszédfelismerést. Megjegyezzük, hogy ebben a modellben más modalitású mondatok viszonyított dallamformája is elhelyezhető (felszólító, óhajtó, más típusú kérdés stb.). Megjegyezzük továbbá, hogy a mondat szintaktikai szerkezete kihat a modell konkrét alkalmazására. A modell használatakor a referenciaponthoz egy Hz-értéket kell rendelni, ettől kezdve a dallam fizikailag is értelmezhető. Ezt példán keresztül szemléltetjük a 6.2. ábrán. A referenciapont feleljen meg 120 Hz-nek (férfi hangfekvés). Legyen a több mondatból álló közlés a következő: *Esni fog az eső. Hány órától? Szerintem délelőtt. Elmentek kirándulni? Elmegyünk.* A jellemző dallamokat és kapcsolódásukat követhetjük végig az ábrán. Ez a modell magyar beszédszintetizátor prozódiai moduljában működik (lásd a 10.3.6.1. fejezetet).



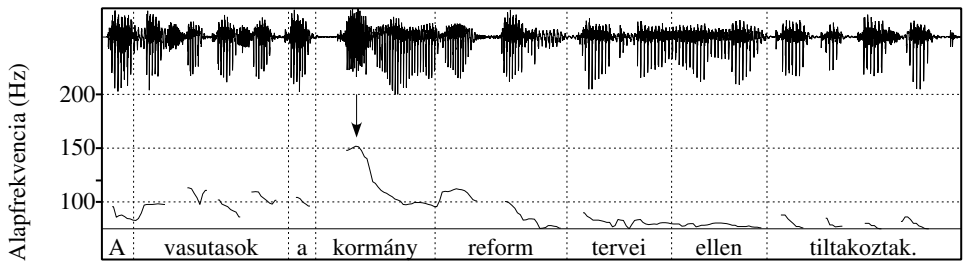
6.2. ábra. A mondatdallamok jellemző kapcsolódása folyamatos beszédben vagy dialógusban

6.1.2. A kijelentés dallamszerkezetei

A magyar kijelentő mondat dallamformája általános megvalósulásban ereszkedőnek tekinthető (Deme 1962, Gósy 2004b). A látszólag egyszerű dallamforma azonban bonyolult szerkezeti formációkat takarhat, amiket egyrészről a mondat szintaktikai szerkezete, hossza, összetettsége, másrészről a beszélő akarata határoz meg. A nyelvi üzenet is befolyásolja a dallam megformálását. A beszélő szinte automatikusan alkalmaz egyfajta beszédstratégiát a tartalom függvényében. Másként olvassuk a híreket, másként egy novellát, egy mesét, egy hirdetést, illetve megint más beszédstratégiát alkalmazunk, ha szabadon, spontán fogalmazunk. Ezért van az, hogy az elemzések során nehéz invariáns jegyeket meghatározni a kijelentés dallamformájának komplex jellemzésére. Az elmúlt évtizedek elemzéseinek és szintézissel végzett kísérleteinek köszönhetően kissé közelebb jutottunk ennek a problémának a feltárásához. Ezek alapján foglaljuk össze a legfontosabb invariáns jegyeket erre a mondatfajtára. A kijelentés teljes dallamformájában elméletileg három stilizált, vonalas közelítéses dallammenetet tartanak számon, ezek az ereszkedő, az emelkedő és a szinttartó. A dallammenet elhelyezkedése, váltakozása a mondaton belül attól függ, hogy egyszerű, összetett vagy többszörösen összetett mondatról van-e szó. Az ereszkedő, emelkedő és szinttartó dallammenetek hatóköre mondatra, mondatrészre és szavakra egyaránt vonatkoztatható. A szinttartó meghatározás tágan értendő, tehát közel azonos frekvenciaértéket jelent, belsejében lehet némi hullámzás. A végleges dallamforma kialakításába beleszól a hangsúlyozás is (ezt külön fejezetben tárgyaljuk). A hangsúlyos és hangsúlytalan elemek váltakozása meghatározza az alapvető dallamformát is. A hangsúlytalan elemeket (névelők, kötőszók stb.) jellemzően szinttartó dallammenettel ejtjük, a hangsúlyozott részeket pedig ereszkedővel (a hangsúly okozta alapfrekvencia-változást ebbe nem értjük bele). Enyhén emelkedő a dallam például a felolvasásban a vesszők előtt. Előfordulhat az erősen hangsúlyozott szavak előtt is ilyen emelkedés, mintegy előkészítve a hangsúlyozást. A kijelentő mondat teljes dallammenetének frekvenciaátfogása átlagosan 30%-nyi. Ez annyit jelent, hogy a mondatban ennyit esik az alapfrekvencia. Az esés nem lineáris, nagyobb a mondat elején, kisebb a végén. Ez természetesen függ a beszélőtől és a tartalomtól. A következőkben példákban érzékeltetjük a kijelentés dallamformájának változatosságát. Egy hírolvasásból származó mondatot mutatunk be a 6.3. ábrán, férfi ejtésben. Ha ennek a mondatnak a jellemző dallamformáját szeretnénk törtvonalas közelítéssel megadni, akkor azt mondhatjuk, hogy egyetlen ereszkedő dallamformát kellene alkalmazni, amely 100 Hz-től 70 Hz-ig változik. Erre szuperponálódnak a hangsúlyozásból keletkező F_0 csúcsok. Általában a mondat tartalma és összetettsége befolyásolhatja a dallamkép általános vonulatát. A legmagasabb alapfrekvencia-érték lehet a mondat belsejében is, ellentétben a korábbi sematikus ábrázolással. Erre vonatkozó példát láthatunk a 6.4. ábrán szintén hírfelolvasásból. A fentiekből láthatjuk, hogy egy adott kijelentés dallamformájának közelítő leírásához sok esetben szintaktikai

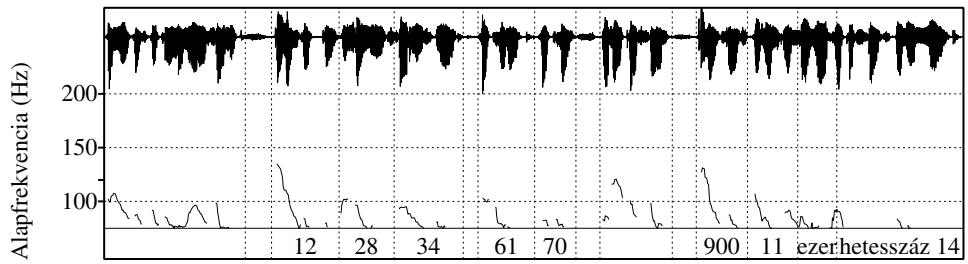


6.3. ábra. Ereszkedő dallammenetek teszik ki a kijelentés (hossza 4,4 s) nagy részét. Az ereszkedő dallamformák frekvenciaátfogása a mondat elején a legnagyobb, majd fokozatosan csökken. A hangsúlyokat az F_0 kiemelkedései mutatják (nyilak)



6.4. ábra. A mondatfókusz dallammenete a kijelentő mondatban. A teljes hossz 3,1 s

elemzésre, illetve a mondanivaló értelmezésére is szükség van. Speciális kijelentésnek számítanak a felsorolást tartalmazó mondatok. Itt a felsorolás elemeit ejthetjük enyhén emelkedő, szinttartó, illetve enyhén ereszkedő dallammal is. A 6.5. ábrán a lottószámokat és a dzsókerszámot mondja be egy rádióbemondó (az ábrán csak a számoknak megfelelő hullámformarészletekhez adtuk meg a szöveges magyarázatot). Ebből a példamondatból azt a következtetést vonhatjuk le, hogy a felsorolásokban a bemondó, egyéni döntés szerint bizonyos csoportosításokat végez. Ez látható a dallammenetek alakulásán és a szüneteken is. A felsorolások prozódiai modellezésében ezt ki lehet használni. A bemutatott példákból is látszik, hogy egyrészről a kijelentő mondatra megadott elvi dallamforma távol áll a ténylegestől, másrészről, hogy a hangsúlyozás is lényeges alapfrekvencia-mozgással jár. Ezért azt kell mondanunk, hogy a kijelentő mondatnak van a legösszetettebb alapfrekvencia-szerkezete a mondattípusok között.



6.5. ábra. Példa a felsorolás kiejtési stratégiájára (az alapfrekvencia és a szünetek használata). *A heti nyerőszámok a következők: 12, 28, 34, 61, 70. A dzsókerszám: 911 714.* A teljes hossz: 9,9 s.

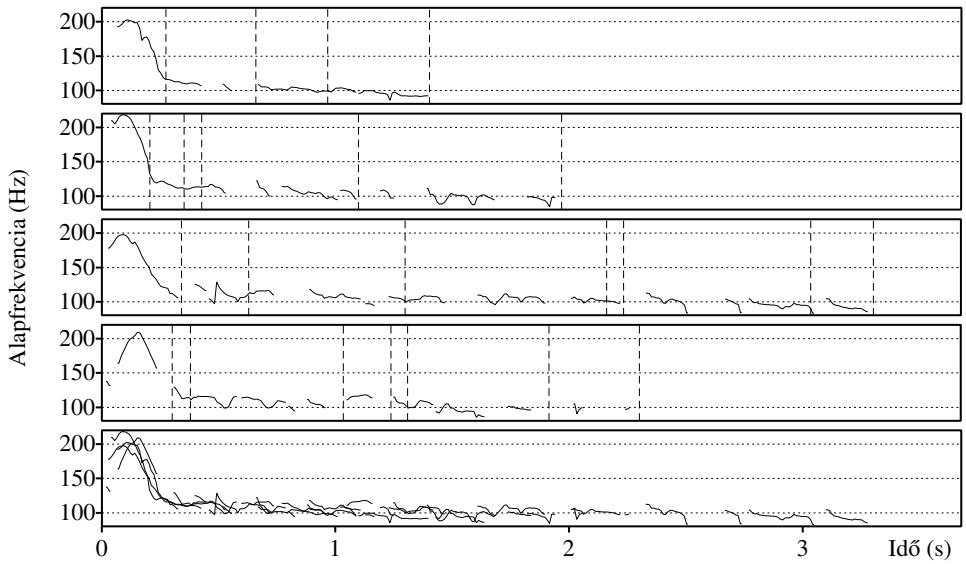
6.1.3. A kérdésformák dallamai

A beszédkommunikációban a kérdésnek kiemelt szerepe van, ezért vizsgálatával is sokan foglalkoztak már a 20. század első felétől kezdve (Csűri 1919, Varga 1994, Olaszy–Koutny 2001, Olaszy 2002, Hegedűs 1930, Deme 1962). A kérdésnek olyan a dallama, amelyik kifejezi, hogy választ várunk (Fónagy–Magdics 1967). Percepciós kísérletekben kimutatták, hogy sok esetben elégséges csupán a dallamvonulat is arra, hogy a kérdést megkülönböztessük a kijelentéstől (Gósy–Terken 1994). A magyar kérdés dallamformájának általános ismertetőjegye a környezetből kiemelkedő dallamcsúcs, amely szótagszinten valósul meg. Dallamcsúcson jelen esetben azt értjük, hogy az alapfrekvencia jelentősen megnövekszik és magas Hz-értéket ér el a szótag magánhangzójában. A dallamcsúcs helye a mondaton belül, valamint a csúcsot követő alapfrekvencia-változás lefolyása a kérdés típusára jellemző.

6.1.3.1. A kiegészítendő kérdés

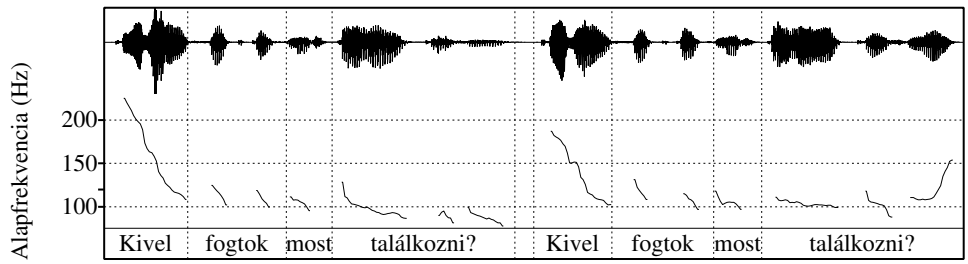
A kiegészítendő kérdés általános dallamgörbéjét már Hegedűs (1930) leírta. Deme (1962, 504. o.) a következőképpen foglalja össze e kérdésfajta dallamát: „...a kiegészítendő kérdéseknek sajátos, csak rájuk jellemző hanglejtésformájuk nincsen. Kérdő jellegüket egyedül a kérdő névmás mutatja. A rájuk jellemző dallamforma a kijelentő mondatra megállapított elül eső.” Ez a megállapítás általánosságban helyes. Fonetikai vizsgálatok szerint azonban két jellemző ponton mégis megkülönböztethetjük a kiegészítendő kérdés dallamformáját a kijelentésétől (Olaszy 2002). Az egyik, hogy a kérdőszó első szótagján kialakult dallamcsúcs magasabb frekvenciájú, mint a kijelentő mondatra jellemző kezdési pont. Minél magasabb a csúcs, annál kifejezőbb a kérdés. Ezt percepciós tesztek eredményei is megerősítették (Olaszy 2001a). A másik, hogy az ilyen kérdésben általában nincsenek hangsúlyok, csak egyetlen dallamcsúcs van, a kérdőszón (más a helyzet a beágyazott kiegészítendő kérdésnél). Ez azt jelenti, hogy a kiegészítendő kérdés dallamformáját függetlení-

teni lehet a szöveg tartalmától (amit a kijelentésnél nem lehetett megtenni). Ezt a tényt például beszédszintézisnél ki lehet használni (Olaszy 2006a). Négy különböző hosszúságú kiegészítendő kérdőmondatot mutatunk be a 6.6. ábrán. A dallamformák hasonlítanak egymásra, függetlenül a kérdések tartalmától és a hosszától. A kérdő-



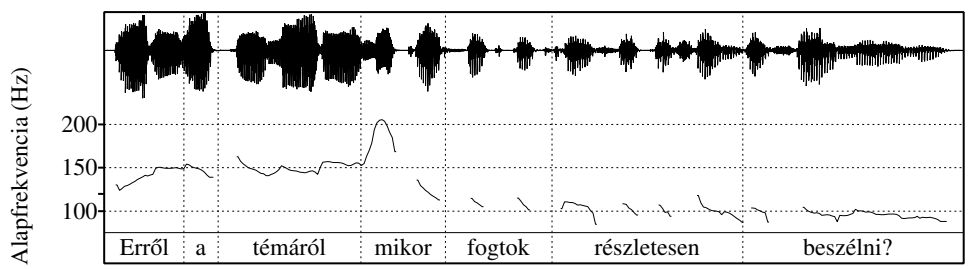
6.6. ábra. A kiegészítendő kérdés dallamformája függetleníthető a tartalmától. A függőleges vonalak a szóhatárokat jelölik

szó első magánhangzójának elején már mindegyik mondatban magas az alaphang (ez az F_0 tér legmagasabb pontja), majd folyamatosan csökken a második szótag végéig. Az F_0 változás teréből ez a két szótagnyi rész, vagyis a kérdés magja foglalja le a legnagyobb részt. A mag utáni ereszkedő részben az F_0 változás kicsi, a kérdés végpontja alacsony frekvencián van (gyakorlatilag ugyanazon, mint a kijelentő mondat vége, ez az F_0 tér legalacsonyabb pontja). Az ilyen kérdéseknél tehát csak a kérdés magjának szótagszerkezetét kell ismerni, hogy a helyes dallamformát például mesterségesen rá tudjuk ültetni. A kiegészítendő kérdésnek a beszélt nyelvben létezik egy ejtési variánsa is (Deme 1962, 512. o.). Ebben a beszélő az utolsó szótagban felkapja az alaphangot (***Kivel*** *fogtok most találkoz****ni****?*). Ennek a variánsnak a létjogosultságát Gósy–Terken (1994) percepciós vizsgálattal is bizonyította, az ilyen dallamformát kérdésként azonosítjuk. Az utolsó szótagban az alapfrekvencia mintegy 10%-nyit emelkedik folyamatosan (6.7. ábra). Ez az emelkedő rész hat a megelőző rész dallammenetére, azt megemeli, vagyis az egyébként ereszkedő rész inkább szinttartóvá válik, hogy előkészítse a dallamvégi felugrást. Ez a kérdésforma a köznapi beszédben nagyon terjed.



6.7. ábra. A kiegészítendő kérdés dallamformája (bal 1,8 s) és annak variánsa, amikor az utolsó szótagban emelkedik az alapfrekvencia (jobb 1,9 s)

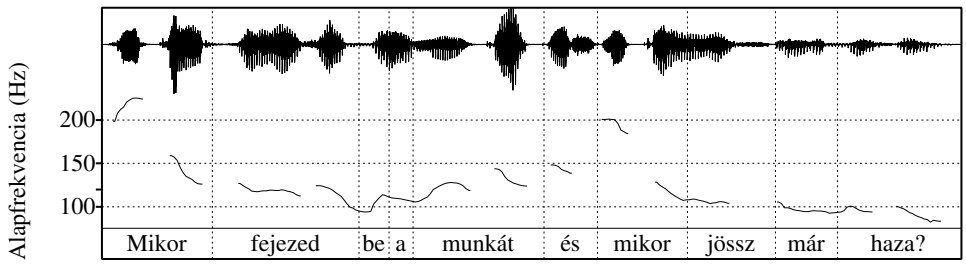
Kiegészítendő kérdés elődallammal. Ha a kiegészítendő kérdést egy bevezető rész előzi meg, akkor annak a mondatrésznek elődallama lesz (*Erről a témáról mikor fogtok részletesen beszélni?*). A példamondat dallamgörbéjét a 6.8. ábra mutatja. Az elődallam ebben a példában alacsony értékről indul és kissé emelkedő jellegű. A dallamcsúcs kialakulása az elődallam végétől kezdődik. Az elődallam más formával is előfordulhat, például magasabbról induló és inkább szinttartó (Varga 1994). Ezt a beszélő szándéka határozza meg.



6.8. ábra. A kiegészítendő kérdés elődallammal. A példamondat hossza: 3,3 s

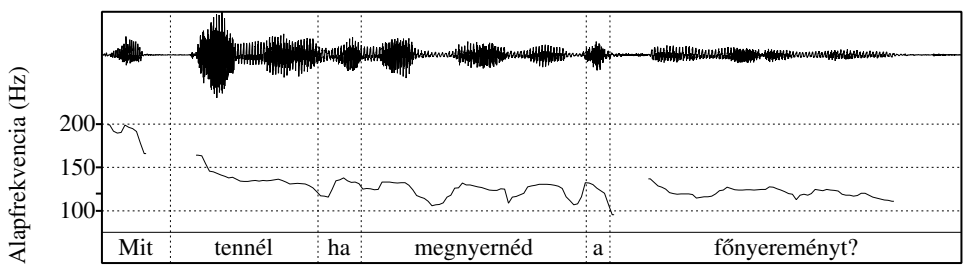
Összetett kiegészítendő kérdés. Az összetett kiegészítendő kérdés két vagy több kérdésrészt tartalmaz. Ezek dallamai egyedi kérdésként valósulnak meg, a kérdések dallammenete formailag megismétlődik, illeszkedve az F_0 térhez (6.9. ábra). Ennek megfelelően az első kérdésben az ereszkedő rész végpontja nem fogja elérni azt a mélységet, ami az ilyen kérdések legvégére jellemző (ezzel jelzi a beszélő, hogy nincs a közlés teljesen befejezve). A második kérdésrész ugyanolyan F_0 szerkezettel rendelkezik, mint az első, csak az értékek lejjebb csúsznak, hogy a kérdés legvége elérje az F_0 tér legalsó pontját (lásd a 6.7. ábra). A helyzet hasonló, ha a kérdés után még további tagmondat következik, de az nem kérdés. Például:

Mit tennél, ha megnyernéd a főnyereményt?



6.9. ábra. Az összetett kiegészítendő kérdés dallamformája. A két kérdés egyedi dallamformája kis szintkülönbséggel megismétlődik. A példamondat hossza: 2,9 s

Itt két dallamforma tölti ki az F_0 teret. A kérdés teljes formája van jelen az első tagmondatban, és a folyamatosan gyengén eső rész folytatódik a másodikban (6.10. ábra). A kiegészítendő kérdések egyik speciális szituációfüggő formája,



6.10. ábra. A kiegészítendő kérdés és a hozzá kapcsolódó mondatrész dallamformája. A teljes példamondat hossza: 2,2 s

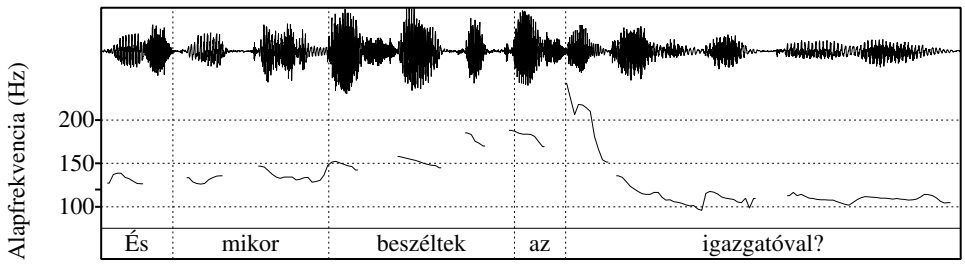
amikor nem a kérdőszó kapja meg a hangsúlyt. Ez olyan mondatokban fordul elő, amelyekben a mondat más részére kérdezünk.

*És mikor beszéltek az **ig**azgatóval? Mikor fog a cikk **meg**jelenni?*

A példákban a mondat első része (beleértve a kérdőszót is) előtdallamként valósul meg, a kérdésre jellemző dallamforma csupán egy szóra (annak is az első szótagjára) szűkül le, amelyre kérdezünk (mint ha egy kijelentő mondatban ugyanezt a szót hangsúlyosan ejtenénk). Az ilyen mondatoknál már szerepet játszik a jelentés, a beszélő dönti el, hogy melyik szóra vonatkoztatja a kérdést (6.11. ábra).

6.1.3.2. Eldöntendő kérdések

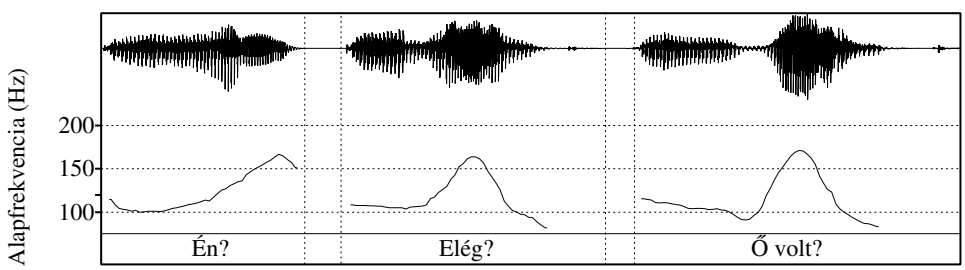
Ennek a kérdésnek sajátos dallamformája van, Deme (1962) emelkedő-esőnek, vagy inkább végén esőnek nevezi. „E formát az jellemzi, hogy utolsó előtti és utolsó



6.11. ábra. Speciális kérdésforma dallama. A példamondat hossza: 2,1 s

szótagja között nagy esés van, a magasság itt a felsőfokról az alsóra esik le"(i. m. 505). Ez a dallamforma jellegzetesen csak az eldöntendő kérdésekre jellemző, azok közül is csak a három, vagy több szótagúakra. A dallamforma tehát különbözik a szótagszám függvényében. Gósy–Terken (1994) szerint ezt a mondatfajtát az emelkedő dallamrész különbözteti meg az egyéb mondatfajtáktól. Ez a megfogalmazás közelebb áll a tényekhez, hiszen emelkedő dallamrész minden eldöntendő kérdésben van. A következőkben pontosabb jellemzését adjuk ennek a kérdésfajtának.

Egy és két szótagú eldöntendő kérdés. Az egy szótagú esetre nem vonatkozik az előbbi általánosnak leírt dallamforma, egyedi dallam jellemzi a kérdést. Fokozatosan, egyre meredekebben emelkedik a dallam az (*Én? Ő?*) kérdéseknél, egészen a közlés végéig (nincs eső rész). A frekvenciaátfogás nagy a szótag eleje és vége között (6.12. ábra bal oldala). Az ilyen egy szótagú kérdésekben a magánhangzó jelentősen megnyúlhat, ami azt biztosítja, hogy a kívánt dallamforma maradéktalanul megvalósuljon.



6.12. ábra. Az egy és két szótagú eldöntendő kérdések néhány változatának dallamformái. A mondatok hossza: 545 ms, 667 ms és 885 ms

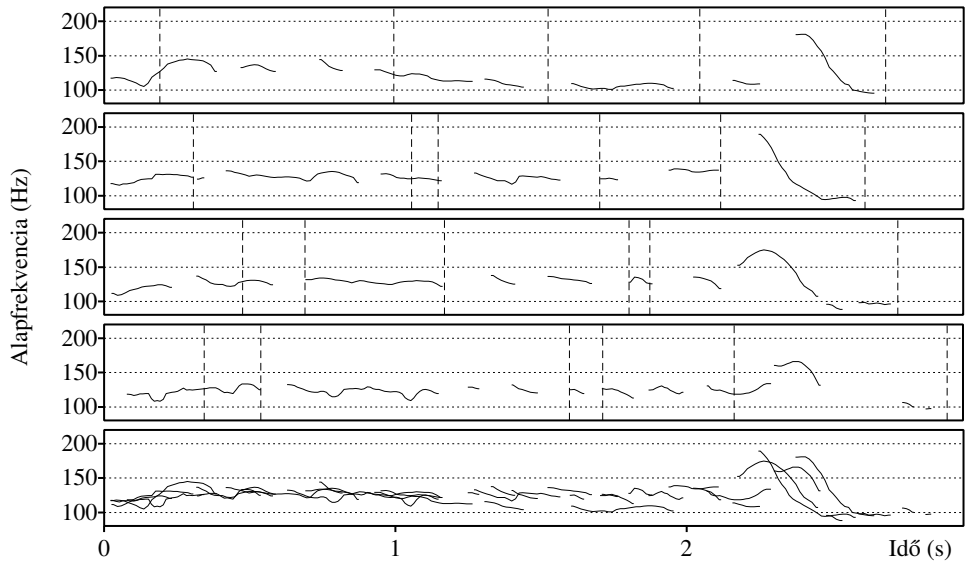
A két szótagú eldöntendő kérdés dallamformája emelkedő-eső karakterű, az emelkedés is és az esés is a kérdés utolsó magánhangzójában jön létre (6.12. ábra). A dallamforma legfontosabb része a hirtelen emelkedés és a meredek csökkenés. Ez bizto-

sítja a megfelelő hangzást. Ez a dallamforma független a mondat szószámától (*Elég? Holnap?, Ő volt?, Kell só?, Ott lesz?, Én is?*). Ugyanakkor az egy szótagú eldöntendő kérdésre korábban megállapított emelkedő dallamforma valósul meg a következő két egyszótagú szóból álló mondatok második szavában: *Ez ő? Már volt?* Egyértelmű szabály tehát nem adható az ilyen rövid kérdéseknél a dallamforma jellegére, a jelentés adhat támpontot.

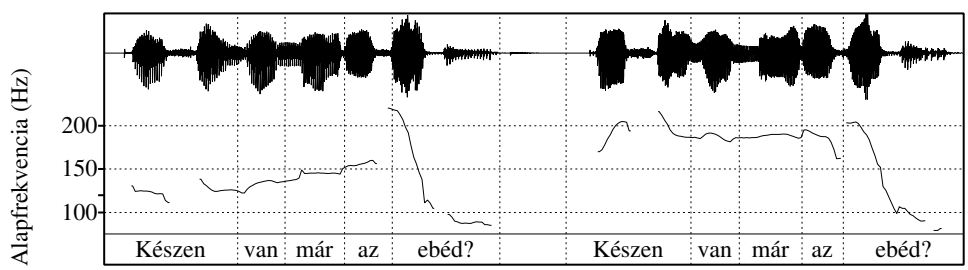
Amennyiben az egy, illetve két szótagú kérdést egy előkészítő rész előzi meg, akkor erre az előzdallamra az enyhén eső jelleg lesz a jellemző, mintegy előkészítendő a nagy frekvenciaváltozási kontrasztot.

Például: *Ennyi már jó? Ennyi már elég?*

Három és több szótagú eldöntendő kérdés. A három és több szótagú eldöntendő kérdés dallamformája a korábbi jellemzések tükrében a következő képet mutatja. Viszonylag mélyről indul és enyhén emelkedő az utolsó előtti szótagig, majd az ezt követő két szótagnyi területen emelkedő-eső forma jellemzi. Az emelkedő-eső rész pontos szerkezete a következő: az F_0 az utolsó előtti szótag magánhangzójának elejére maximumra emelkedik (szökik), majd még ebben a magánhangzóban mélyre csökken. Az utolsó szótagban az F_0 további enyhe csökkenést mutat. A csúcs helyét a szótagszerkezet határozza meg, nem függ a mondat szószerkezetétől. Így előfordulhat, hogy például névelőre esik: *Elhoztad **a** sót?* Az ilyen kérdésben szintén nincsenek hangsúlyok, csak egyetlen dallamcsúcs van, az utolsó előtti szótagon (más a helyzet a beágyazott eldöntendő kérdésnél). Ez azt jelenti, hogy az eldöntendő kérdés dallamformáját függetleníteni lehet a szöveg tartalmától. A (6.13. ábrán) látható, hogy négy különböző tartalmú eldöntendő kérdés dallammenete ugyanazon személy ejtésében mennyire hasonló, ha egymásra vetítjük őket (az ábra legalsó része). Vizsgálataink szerint a kérdés eleje, a mélyről induló, enyhén emelkedő rész már önmagában hordozza a kérdést, ami azt jelenti, hogy nem kell megvárnunk a mondat végét, már előbb is tudjuk, hogy kérdés lesz. Ennek igazolására egy kísérletet mutatunk be. Egy kísérletben 12 preparált mondatot készítettünk, amelyekben a kérdés végén lévő emelkedő-eső részt (az utolsó két szótagot) levágtuk. Percepciós tesztet végeztünk (Olaszy 2001a), amelyben dialógusszituációt imitáltunk, azaz 2-2 mondatos egységeket (kijelentő mondat és utána közvetlenül a megcsonkított kérdés) hallgattattunk meg 4 személlyel (férfiak, életkoruk 25–40 év). A kísérleti személyeknek arra kellett válaszolni, hogy a kijelentő mondat után milyen modalitásúnak gondolják a befejezetlen mondatot: felszólító, kérdő, felkiáltó, kijelentő. A válaszok 83%-ában a kérdést jelölték meg. Az eldöntendő kérdés első felében az alaphang magasabb fekvésben is előfordulhat. Deme (1962, 512. o.) szerint: „... a kérdés elején egyhangúan menő, vagy lassan emelkedő rész magassága utal az érzelemre. Az alacsonyabb kezdés közömbösebb, a magasabb indulatosabb". Mindkét változatra mutat példát a 6.14. ábra.



6.13. ábra. Négy különböző tartalmú eldöntendő kérdés dallammenete egyenként (szóhatárokkal jelölve), majd egymásra rajzolva. Az eldöntendő kérdés dallamformája függetleníthető a tartalmától

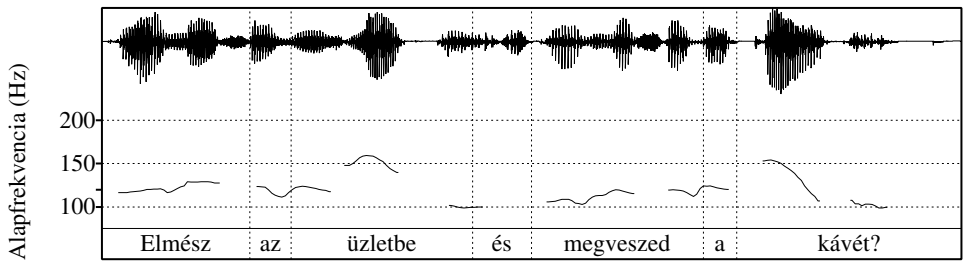


6.14. ábra. Az eldöntendő kérdés közömbösebb (bal 1,6 s) és indulatosabb dallamformája (jobb 1,5 s)

Amennyiben az eldöntendő kérdés a mondatba ágyazva, annak második felében van, akkor előd allam előzi meg (*Tegnap délután elmentél lovagolni?*). Ebben az esetben az előd allam magasabbról indul, és eső jellegű. Az enyhe esés, mintegy előkészíti a mélyről induló, enyhén emelkedő eldöntendő kérdést. Ez az előd allam csak eső lehet, mivel csak ezzel a formával tudjuk elválasztani a kérdés résztől.

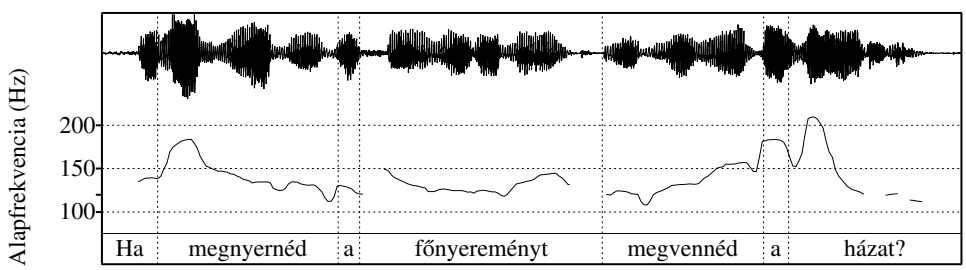
Összetett eldöntendő kérdések. Ha az összetett eldöntendő kérdés két vagy több kérdésrészt tartalmaz, akkor a kérdések dallammenete formailag megismétlődik (*Elmész az üzletbe, és megveszed a kávét?*). Ez azonban tényleg csak formai, mivel a két kérdésnek az egy kérdésre eső F_0 térben kell megvalósulnia (ez a mondat kezdő-

és végpontjára vonatkozik). Az F_0 mozgástér megosztása itt leginkább a kérdés befejező szótagjában lévő alacsony F_0 értékre hat. Ez a pont az első kérdésben magasabban lesz, mint a másodikban. A magasabb értékkel jelzi a beszélő, hogy a közlés nincs befejezve az első kérdésrész elhangzása után (6.15. ábra). Ha az összetett mondatban a kérdést nem kérdő tagmondat előzi meg, a helyzet hasonló. Például: *Megnézed azt a filmet, amiről a múlt héten beszéltél?*



6.15. ábra. Az összetett kiegészítendő kérdés teljes dallamformája. A példamondat hossza: 2,5 s

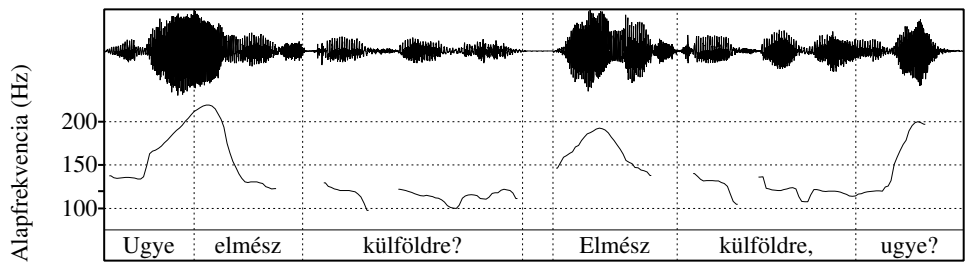
Itt a kérdés tulajdonképpen a főmondatban fogalmazódik meg (*Megnézed*). Az ilyen mondatokban az összetett mondat egészében jön létre ugyanaz a dallamforma, ami az egyszerű kiegészítendő kérdésre jellemző. Az első tagmondat mélyről indul és a dallam enyhén emelkedni fog. A második mondatrészben ez folytatódik és az eldöntendő kérdésre megadott teljes dallamforma itt valósul meg. Ha a kérdést megelőző mondatrész topik jellegű (*Ha megnyernéd a főnyereményt, megvennéd a házat?*), akkor ebben a részben egy magasabbról induló és fokozatosan eső dallamforma valósul meg, ami mintegy előkészíti az eldöntendő kérdés indítását (6.16. ábra).



6.16. ábra. Az összetett kiegészítendő kérdés teljes dallamformája. A példamondat hossza: 2,8 s

Morfémával jelzett eldöntendő kérdés A morfológiai eszközökkel kifejezett kérdésben nem a dallamforma, hanem az adott morféma jelzi, hogy kérdésről van

szó. Amennyiben a kérdés jelzése az *-e* morfémával történik, a dallam hasonló lesz a kijelentéséhez (*Elkészíted-e holnapra a cikket?)*. Amennyiben az *ugye* szóval kezdjük a kérdést (*Ugy**e el**mész külföldre?)*, akkor ennek a szónak a fokozatosan emelkedő dallammenete határozza meg a kérdést. Amennyiben az *ugye* szót mint ellenőrző kérdést használjuk (*Elmész külföldre **ugye**?)*, akkor a kérdés első része a kijelentő mondatra jellemző dallamszerkezetet kapja, az *ugye* szó pedig a két szótagú ellenőrző kérdésre leírtakat (6.17. ábra).



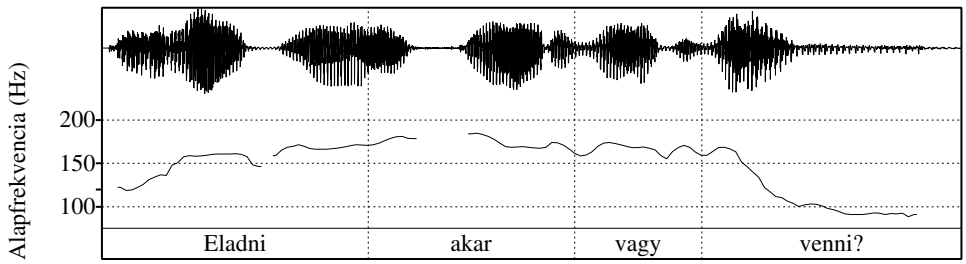
6.17. ábra. Az *ugye* szó szerepe a kérdésekben. A mondatok hossza: 1,5 s és 1,4 s

6.1.3.3. Ellenőrző kérdés

Az ellenőrző kérdést azért tesszük fel, mert meg akarunk erősíteni egy már ismert információt (*Mikor indul a repülő? Három órakor. **Mikor?** Háromkor.)* Az ilyen ellenőrző kérdést mindig az eldöntendő kérdésre jellemző dallamszerkezettel valósítjuk meg (a kérdőszavak esetében tehát kétféle dallamforma is előfordulhat, eső és emelkedő-eső).

6.1.3.4. Választó kérdések

A választó (alternatív) kérdés két részből áll, melyek a *vagy* szóval vannak elválasztva egymástól (*Eladni akar vagy venni?).* Az ilyen kérdés dallamformája teljesen eltér az eddigi kérdésekétől, két részből tevődik össze és ezek dallamszerkezetét alapvetően szótagszintű szabályok határozzák meg (6.18. ábra). A kérdés első felében az első két szótagnak van kitüntetett szerepe. Az elsőben az F_0 alacsonyabb értékű, a másodikban emelkedik, majd szinte szinttartóvá válik a dallam egészen a választóvonalig (*vagy*). A kérdés második felében is az első két szótag a legfontosabb, itt az első képviseli a kérdés csúcspontját, majd a másodikban hirtelen csökkenés következik. A kérdés további részére enyhén csökkenő dallam a jellemző, függetlenül a mondat további hosszától. A befejezéskor az alapfrekvencia az indulási értékhez viszonyítva

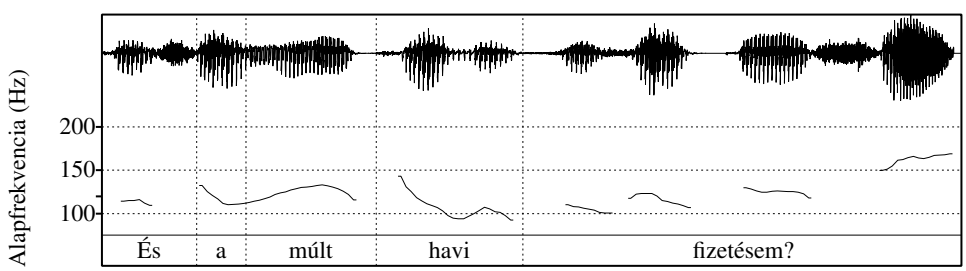


6.18. ábra. A választó kérdés dallamszerkezete. A példamondat hossza: 1,5 s

lényegesen alacsonyabb lesz (a kijelentő mondat befejező értékéhez közeli). A fenti dallamforma csak olyan választó kérdésekben valósul meg, amelyekben mind az első rész, mind pedig a második legalább három szótagnyi hosszúságú. Ellenkező esetben a dallamforma változik, alapvetően az első emelkedő rész, a csúcs és az utána következő meredeken eső rész marad meg (***Én*** *vagy* ***ő****?*).

6.1.3.5. Befejezetlen kérdések

A befejezetlen kérdések magukban hordozzák, hogy a gondolat még folytatódik, ezért dallamgörbéjük mindig magas alapfrekvenciaértéken fejeződik be. Ezek a kérdések gyakran az *És* szóval kezdődnek. Dallamformájukra Deme (1962) például az enyhén emelkedő jelzőt alkalmazza. A dallammenet pontos szerkezete szótagszinten írható le. Az emelkedés jellemzően az utolsó három szótagban fokozatosan történik meg (*És a fiz****etésem****?*), az utolsóban a legmeredekebben (6.19. ábra). Az alapfrekven-



6.19. ábra. A befejezetlen kérdés dallamformája. A példamondat hossza: 1,6 s

cia a kérdés végén a legmagasabb. Ha nincs meg a kívánt szótagszám, akkor balról jobbra csonkul a fenti dallamforma. Egy szótagú esetben például az eldöntendő kérdésre jellemző meredek emelkedés valósul meg (*És ő?*), akár egy hangon belül is.

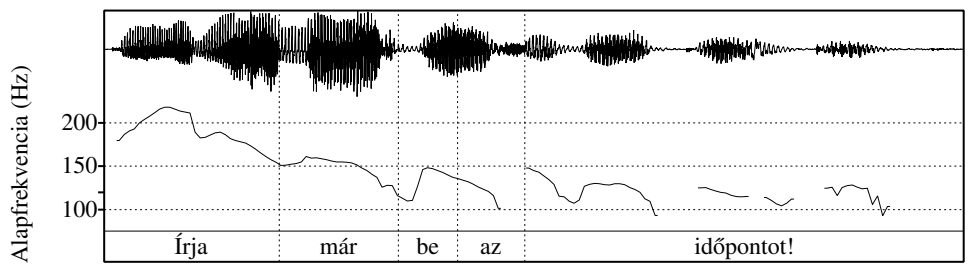
Amennyiben hosszabb előrész előzi meg az utolsó szót, akkor ennek az elődallama alacsony lebegő jellegű (6.19. ábra).

6.1.4. Más modalitások dallamformái

A beszéddallam a beszélő tudatos döntésének az eredménye. Ezért sokféle dallamformával találkozhatunk az elemzések során. A kijelentés és a kérdés mellett bemutatunk néhány más, jellemző dallammenetet is.

6.1.4.1. A felszólítás dallama

A felszólításban (6.20. ábra) az utasítás és a türelmetlenség fejeződik ki (*Írja már be az időpontot!*), már a mondat elején. Az általános dallamot leíró szerkezeti egységek



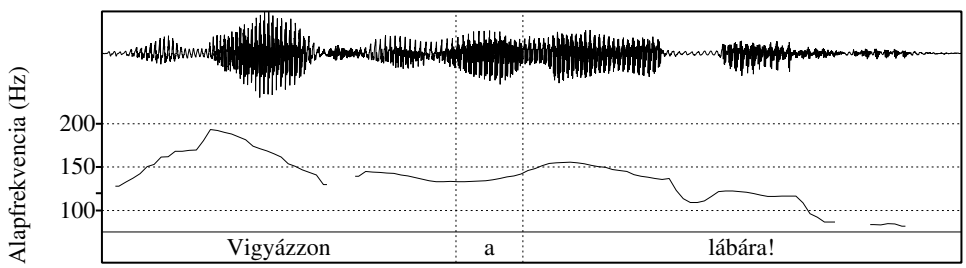
6.20. ábra. A felszólítás dallamformája. A példamondat hossza: 1,7 s

a következők. Az alaphang indítása a kijelentéshez viszonyítva magasabb. Ezután az első szótagban meredek esés figyelhető meg, és ez az esés adja a felszólítási kifejezési forma lényegét. A második és harmadik szótagban további esés valósul meg, ez azonban sokkal enyhébb, mint a korábbiak. Ezután az alaphang fokozatosan csökken, de a mondat végén sem éri el a kijelentés befejezésére jellemző frekvenciaértéket. A felszólításban tehát három különböző meredekségű, csökkenő dallammenet mutatható ki. A legmeredekebb a legrövidebb, a legkevésbé meredek a leghosszabb (több szavas mondatnál). Látható tehát, hogy mind értékben, mind pedig struktúrában ismét lényeges eltérés van a kijelentő mondat intonációs tendenciagörbéjéhez képest. Az első szótagi hangsúlyozási szabály megvalósulhat ebben a közlési formában a beszélő szándékától, valamint a közlés hosszától, tartalmától függően. A felszólítás eme alapformájának van egy variánsa is, hasonlóan, mint a kiegészítendő kérdésnél. A dallamforma itt is hasonlóan alakul, az utolsó szótagban a beszélő egy kismértékű emelkedő dallamrésszel zárja a mondatot. Az ilyen megformálásban ez

az emelés visszahat a dallamgörbe korábbi alakulására, azaz nem süllyed annyira, hanem inkább szinttartó lesz. Amennyiben a felszólító mondat udvariasabb formában hangzik el, akkor több hangsúlyos szó is lehet benne (*Tessék helyet foglalni!).* A felszólításban az indulati fokozatok inkább az intenzitásgörbében jelentkeznek, a dallammenet indulattól függetlenül ugyanolyan struktúrájú. Az intenzitásgörbe indulási pontja a felszólításokban mintegy 5 dB-lel magasabb, mint a kijelentés indítási intenzitása, de ez az érték – érzelemtől függően – akár 15 dB-lel is magasabb lehet.

6.1.4.2. A figyelmeztetés dallama

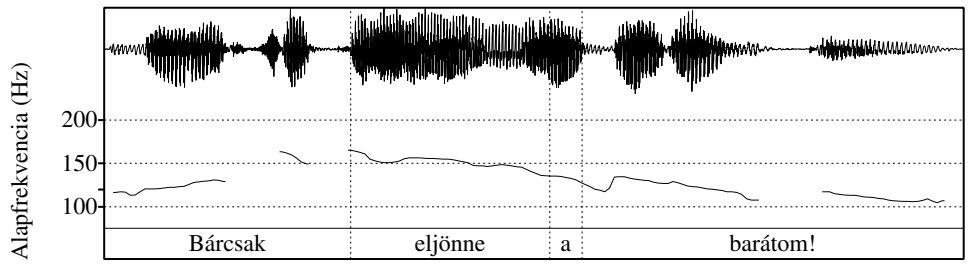
A figyelmeztetés dallama két ponton tér el a kijelentő mondatétól (6.21. ábra). Az első, hogy a figyelmeztetésben az alaphang magasabb értékről indul, és ezután hirtelen csökken, a másik, hogy magasabb alapfrekvenciaértéken fejeződik be, mint a kijelentés.



6.21. ábra. A felszólítás dallamformája. A példamondat hossza: 1,2 s

6.1.4.3. Az óhajtás dallamformája

Az óhajtást kifejező felszólító mondatokra Deme (1962) meghatározása szerint az enyhén ereszkedő dallamforma jellemző. Olaszy (2001a) emelkedő-eső dallamformát mért az ilyen mondatokban. A *Bárcsak eljönne a barátom!* mondat dallamgörbéjét az a 6.22. ábra mutatja. Az alaphang átfogási sávja a mintamondatban 105 Hztől 170 Hz-ig terjed. Ez meglehetősen széles tartomány. Az átfogási sáv szélessége attól függ, hogy mennyi érzelem van az óhajtásban. A dallamvonal kissé mélyebbről indul, mint a kijelentésnél, majd az első szótagban erősen emelkedik. Az alapfrekvencia emelkedése a második szótag elején éri el a maximumot. A harmadik szótag elejétől válik ereszkedővé a dallam, és az indulási értéknél kissé alacsonyabban fejeződik be a mondat végén. A mondat óhajtó jellegét az elején lévő emelkedő dallamforma alakítja ki két szótagnyi részen. Percepciós kísérletek szerint az



6.22. ábra. A *Bárcsak eljönne a barátom!* mondat (1,8 s) dallamgörbéje

óhajtó mondat dallamformája akkor fejezi ki az óhajtás tényét, ha a mondat elején az első szótagban emelkedik a dallammenet (Olaszy 2001a). Amennyiben az óhajtást kifejező rész két szótagnál rövidebb (*Bár eljönne!)*, a csúcsképződés szűkítve, az első szótagban jön létre, tehát a magánhangzón belül zajlik le az alapfrekvencia emelkedése és csökkenése is. Ez nem annyira intenzív, mint a több szótagú esetben. Ilyen esetben az első magánhangzó lényegesen megnyúlhat, hogy legyen idő az alapfrekvencia-változtatására.

6.2. A hangsúlyozás

A hangsúlyozás vizsgálata a beszédkutatás egyik legnehezebb területe. Elméleti szempontból a nyelvészeti szakirodalom ad iránymutatást (Laziczius 1944, Kálmán–Nádasdy 1994, Gósy 2004b), gyakorlati megoldásokról a beszédszintézissel foglalkozó munkákban találunk példákat (Hunyadi 1995a, Olaszy 1989a, 2001a). Legújabban a gépi beszédfelismerésben is helyet kap a hangsúlyozás vizsgálata, hatékonyabbá tehető a felismerés pontossága (Szaszák 2008). Az egyes nyelvek hangsúlyozási rendszere eltérhet, vannak kötött (magyar, francia) és szabad (angol, orosz) hangsúlyozásúak. A kötött hangsúlyozás esetében a hangsúlynak kiemelő, határjelző szerepe van, döntően a közlés értelmezéséhez nyújt segítséget, lehet szó- és mondathangsúly. A magyarban az elmélet szerint a hangsúlyozott szó első szótagján van a hangsúly, ezért a hangsúly jelenléte egy új szó kezdetét is jelenti. Ezt ki is használják a beszédtechnológiában (lásd a 9.10. fejezetben). A szabad hangsúlyozású nyelvekben más a helyzet. Itt a hangsúly helyének változtatásával a szó jelentését (és hangalakját) módosíthatjuk. Ilyen az angol (*rec**o**rd* mint ige: felvételt készíteni, illetve a *r**e**cord* mint főnév: a felvétel). A hangsúlyozás és a mondat szerkezete között is van összefüggés. Ezt fejezi ki a mondathangsúly (fókusz), amely a hierarchia tetején áll, tehát ez a legerősebb. Nyelvi szempontból azt mondhatjuk, hogy a hangsúlyozáskor nyomatéknövelést használunk. Fiziológiailag ez azt jelenti, hogy a beszédképzéshez használt izomműködéseket hirtelen fokozzuk. Fizikai szempontból a hangsúlyozás-

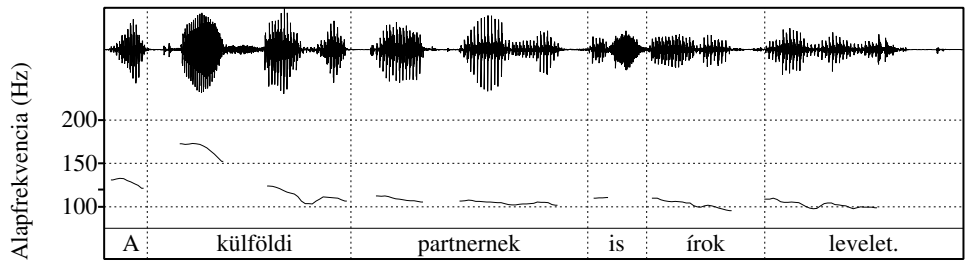
kor a hangintenzitás és az alapfrekvencia képezhet helyi kiemelkedést, valamint a hangidőtartam növekedhet. A kiemelkedés azt jelenti, hogy a környezeti értékekhez képest nagyobb az adott paraméter értéke. A lényeg a kontraszt létrehozása. Ezt akár a hangsúlyozás előtt tartott szünettel is lehet fokozni. Az, hogy melyik paraméter milyen mértékben vesz részt a hangsúlyélmény kialakításában, a nyelvtől, a beszélőtől és a közlés tartalmától egyaránt függ. Fónagy (1958) vizsgálatai szerint a szóhangsúlynak lehet dallamvetülete, de ez nem kötelező. Gósy (2004b) szerint a hangintenzitás megnövelése jellemzi döntően a hangsúlyozást. Hunyadi (1995a) azt vallja, hogy az intenzitás és az alapfrekvencia komplex relációjával lehet a hangsúlyt leírni. Ennek a relációnak a kifejezésére alkotott meg egy új eljárást, amelyben a PET (Pitch and Energy over Time) értéke mutatja meg, hogy a hangsúly létrehozásában az alapfrekvencia növekedése, vagy az intenzitás kap-e nagyobb hangsúlyt. Hunyadi módszere fizikai mérésen alapul, és a hangsúlyok kimutathatók a mondaton belül. Ha fiziológiai oldalról közelítjük a kérdést, akkor azt kell mondanunk, hogy a hangsúlyozásnál az alapfrekvencia- és a hangintenzitás-növekedés együtt járhat, ugyanis a levegőnyomás (nyomaték) növelésével a hangszalagok feszítettsége erősebb lesz, tehát növekszik a mozgásuk rezgésszáma is. Az más kérdés, hogy a hangsúlyozás megvalósulását melyik fizikai paraméter mérésével próbálják kimutatni.

Egy alapfrekvencia-változást figyelembe vevő modellben (Olaszy 2001a) a következő kérdésekből indult ki a kutatás:

- a hangsúlyos szótagon belül van-e alapfrekvencia-kiemelkedés?
- ha van csúcs, akkor hol (az első magánhangzó elején, végén vagy máshol)?
- mennyi az F_0 kiemelkedése?
- hol kezdődik az F_0 csökkenése?
- hol fejeződik be az F_0 csökkenése?
- mennyire csökken a frekvencia a maximum után?
- hogyan függ a hangsúly megvalósulása a szó mondatban elfoglalt pozíciójától?

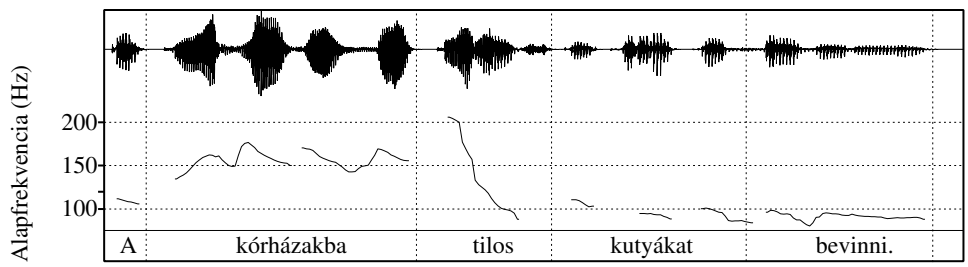
A vizsgálatokban 140 természetes ejtésű mondat szavainak első szótagjában végeztek mérést. A következő öt osztályozási formát eredményezték a vizsgálatok.

a) A kiemelten hangsúlyos esetekben az alapfrekvenciaváltozás formája a legjellegzetesebb és egységes képet mutat, minden esetben ugyanazt a formát találtuk. Ilyen dallammenet valósul meg például a mondathangsúlynál. Az alapfrekvencia a hangsúlyos szótag magánhangzójának a kezdetén már maximumon van, onnan fokozatosan csökken a magánhangzó végéig. A kiemelkedés a hangsúlytalan részhez viszonyítva az 50%-ot is elérheti. Ez tehát egy határozott eső alapfrekvencia-változás. Erre a formára két példát mutatunk be. *A **kül**földi partnernek is írok levelet.* mondatban a hangsúlyos szótagban megvalósul az alaphang meredek esése (6.23. ábra), mintegy 30 Hz. Ez a tendencia ugyanilyen meredekséggel folytatódik a második szótagban is. Összességében tehát a két szótagban 80 Hz-nyi esés jön létre. *A kórházakba **til**os kutyákat bevinni!* mondatban a mondathangsúly a *tilos* szón van. A kiemelt



6.23. ábra. A mondathangúlyt hordozó szóban meredek alapfrekvencia-esés jön létre, a mondat többi részében a csökkenés minimális (a bemondó jellemző hangfekvése 120 Hz). A példamondat hossza: 2,3 s

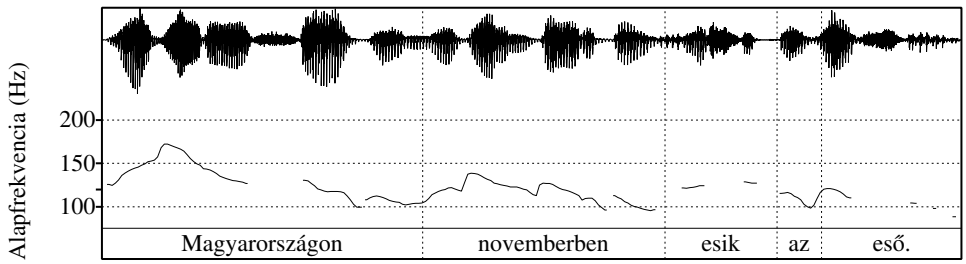
szó miatt a beszélő nem hangsúlyozza az előző szót, mintegy fokozva a kontrasztot a két szó között. A *tilos* szó első szótagjában viszont a fent leírt – kiemelt hangsúlyra utaló – egy hangon belüli határozott eső alaphangmagasság-változás látható (6.24. ábra), amely folytatódik a második szótagban is. Itt összességében 100 Hz-nyi alapfrekvencia-esés valósul meg a két szótagon.



6.24. ábra. A kiemelten hangsúlyos szó hatással van a megelőző hangsúlyos szóra, amiben az alapfrekvencia magas értékű lesz. A kontraszt csak a *tilos* szóban jön létre. A példamondat hossza: 2,6 s

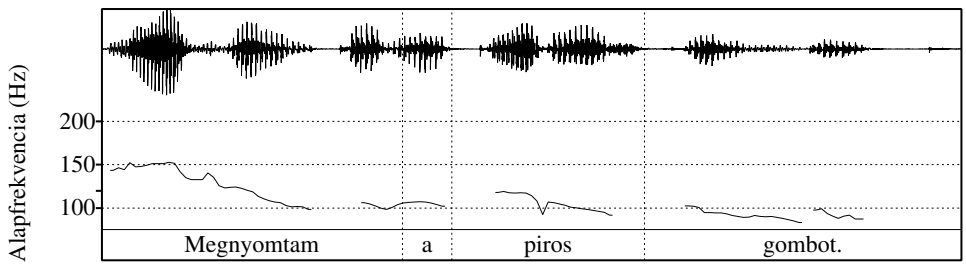
b) Ha a szót kevésbé hangsúlyozzuk, akkor az alapfrekvencia-csúcs gyakran a második szótagra, annak magánhangzójának elejére tolódik el, annak ellenére, hogy a szóhangsúlyt semmiképpen nem a második szótagon lévőnek halljuk. Ezt a percepcióban nem érzékeljük. Ez a hangsúlyozási forma többnyire a három szótagnál hosszabb szavakban fordul elő. Erre a második formára mutat példát a *Magyarországon novemberben esik az eső.* mondat alapfrekvencia görbéje (6.25. ábra). Az első két szótagon az eltolódott csúcs világosan látható. Az első magánhangzóban még csak emelkedik az alapfrekvencia, majd a második elején éri el a maximumot. Ezután e hang végéig csak néhány Hz-et esik, utána az esés folytatódik a harmadik szótagban. A kontraszt tehát három szótagnyi részen valósul meg. Ugyanez a tendencia látható a 6.25. ábra második szavának első két szótagjában is.

c) Amennyiben a szó három, illetve két szótagú, az eltolt csúcs megvalósulására nincs meg a szükséges szótagszám, ezért a csúcs az első magánhangzón lesz (leg-



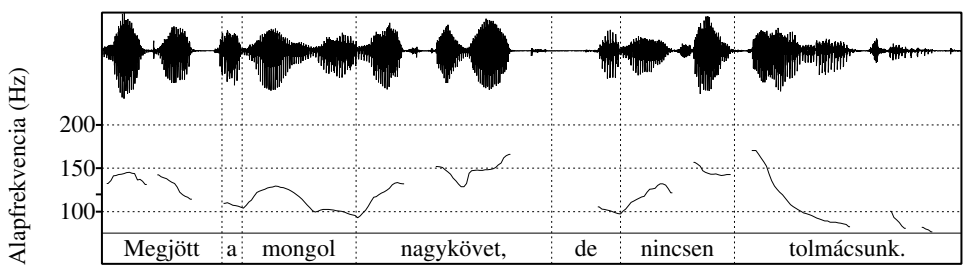
6.25. ábra. Az első szóban az alapfrekvencia csúcsa a második szótagra esik. A példamondat hossza: 2,6 s

inkább a végén), a csökkenés pedig a második magánhangzó végéig lezajlik (6.26. ábra).



6.26. ábra. Az első szóban az alapfrekvencia csúcsa az első szótagra esik. A példamondat hossza: 1,5 s

d) A hangsúlyra jellemző alapfrekvencia-csúcs el is maradhat, ha a hangsúlyos szó rövid (maximum három szótag) és utána például vessző, határjelzés van. Ha a vessző hatására az alaphangot megemeljük a szó végére, akkor ez elnyomja az első szótagban elvben megvalósuló alaphang-emelkedést (itt a vessző intonációkontrasztja fontosabb, mint a hangsúlyozásé), a hangsúlyt ilyenkor valószínűsíthetően a hangerő megnövelése alakítja ki (6.27. ábra). Az ábrán látható mondat minden szava



6.27. ábra. A *nagykövet* szó hangsúlyát nem az alapfrekvencia növekedése jellemzi. A példamondat hossza: 3,2 s

hangsúlyosnak hallatszik, a *nagykövet* szó első szótagjában az alapfrekvencia mégis a legalacsonyabb. Ebben a szóban a hangsúlyos szótag tehát az őt megelőző szó végén lévő alacsony alapfrekvenciaértékhez csatlakozik, innen folytatódik a dallam emelkedése. A szót mégsem érezzük hangsúlytalannak. Az ilyen esetekben a mondatszintű dallamkényszer (a vessző hatására képzett emelkedő alapfrekvencia) uralkodik, és nem engedi, hogy a szóhangsúlyt az alapfrekvencia-emelkedéssel fejezzük ki. A hangsúlyt jelző fizikai értékek nagysága függ a hangsúlyozott szó helyétől is a mondaton belül. A mondat végéhez közelebb álló hangsúlyozott szavakban kisebbek az alapfrekvencia- és intenzitásváltozások, mint a mondat elején vagy fókuszpozícióban. Hírolvasási vizsgálatból mutatunk be egy férfi ejtésű mondatot a mért adatokkal (6.1. táblázat).

(1)Kiadható (2)Magyarországnak (3)Kulcsár Attila, döntött (4)nem jogerősen a (5)bécsi (6)legfelsőbb (7)tartományi (8)bíróság. A táblázatban a hangsúlyozott szó első és második szótagjának magánhangzójában mért átlagokat tüntetjük fel a mondat 8 pontján mért adatokból. A fenti vizsgálatok eredményeiből látható, hogy a

6.1. táblázat. Alapfrekvencia-átlagok szótagokban a hangsúlyos szavakban

Sorszám	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8
Szótag	ki	ad	Ma	gyar	kul	csá	nem	jo	bé	csi	leg	fel	tar	to	bí	ró
F_0 (Hz)	175	113	107	91	110	80	111	93	115	95	96	82	91	82	76	68
I (dB)	32	27	28	23	28	20	25	22	24	24	24	23	26	23	19	18

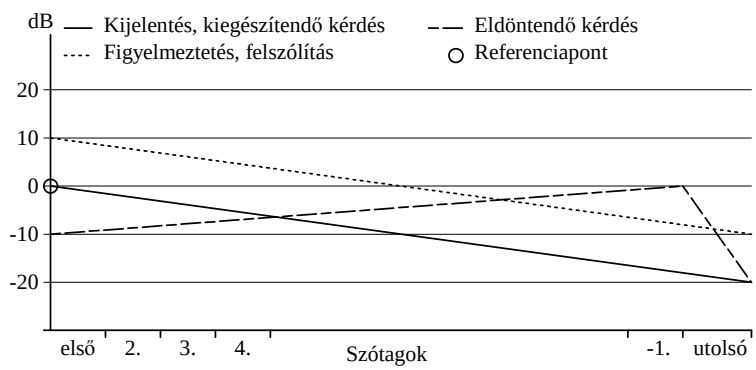
hangsúlyozás percepciós élménye mögött számos fizikai megvalósulási lehetőség van. További vizsgálatok szükségesek a témakör teljes feltárásához. Végül megjegyezzük, hogy a magyarban is előfordulnak olyan esetek, amikor a kötött hangsúlyozás szabályát megsérti a beszélő. Ezek azonban olyan szituációkban fordulnak elő, amikor a beszélő ezzel a hangsúlyáthelyezéssel segíti a mondandója értelmezését (*Pécsett, nem Pécsen*. *Naná, hogy elmegyek!*). A spontán beszédben a hangsúlyozási formák akár lényegesen is eltérhetnek az előbb tárgyaltaktól.

6.3. Hangintenzitás mondatkeretben

A beszéd szupraszegmentális szerkezetében az intenzitásnak a mondatra, illetve a szavakra vonatkoztatott vetületét tárgyaljuk. Nem foglalkozunk az egyénre jellemző átlagos hangerősséggel, csak a mondaton belüli viszonyok alakulását és a megvalósított dinamikatartományt vizsgáljuk. A mondatok modalitásának megkülönböztetésében az intenzitás alakulásának is szerepe van (az alapfrekvencia és az intenzitás szoros összefüggésére gondolunk). Egy beszédtechnológiai példával élünk ennek érzékeltetésére: a kérdés, hogy mit kell tenni, hogy egy kijelentő mondatból a hangtest megváltoztatása nélkül eldöntendő kérdést csináljunk?

A példamondat legyen a következő: *Balassagyarmatnál a várakozási idő 3 óra.* Az általános válasz (ahogy azt sok szakkönyv is sugallja), hogy a dallammenetet meg kell változtatni (például a 6.1. ábra szerint). Amennyiben ezt jelfeldolgozási módszerrel korrektül végrehajtjuk, nem kapunk jól hangzó kérdő mondatot. Miért? Mert a kijelentő mondat végén a hangintenzitás már kicsi, és ennek megemelése nélkül nem fog érvényesülni az eldöntendő kérdésben fontos szerepet betöltő, utolsó előtti szótagra helyezett magas alapfrekvencia-csúcs, amitől a mondat kérdéssé válna. A választ az előbbi kérdésre tehát ki kell egészíteni azzal is, hogy a megfelelő intenzitáskorrekciót is el kell végezni. Ebből következik, hogy a mondatok modalitása és intenzitásszerkezete között is szoros összefüggés van (Olaszy et al. 2000a). A mondat intenzitásszerkezetének részletezett kutatását a mesterséges beszéd-előállítás fejlődése kényszerítette ki. Ha nem megfelelő intenizitásszinteket alkalmazunk a beszédszintézis során, akkor a beszéd huppogóvá, hullámzóvá válik. A beszéd stílusa is hatással lehet a hangintenzitás lefolyására. Más stratégiát alkalmaz a beszélő, ha szépirodalmat olvas, ha híreket mond, ha mesél, illetve reklámot hangosít meg (Olaszy 2005). Ezeket a témaköröket vizsgáljuk a következőkben.

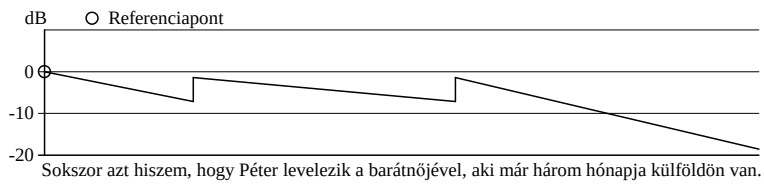
Mondattípusok. A különböző modalitásokra hasonló összefüggéseket lehet megállapítani a mondatintenzitás lefolyását illetően is, mint amilyeneket az alapfrekvencia tekintetében tettünk. A referenciapont itt is a semleges ejtésű kijelentő mondat kezdeti intenzitásszintje. Ehhez viszonyítjuk a mondaton belüli változásokat, valamint a mondat végi szintet (6.28. ábra). A változásokat dB-ben adjuk meg. A mondaton belüli átfogási sáv általában 20 dB (a mondatban található legmagasabb



6.28. ábra. A mondattípusokra jellemző általános intenzitásszerkezetek és kapcsolódásaik rendszere a kijelentés függvényében

és legalacsonyabb szint között). Az intenzitásváltozás jellemzésére itt is a három formát alkalmazzuk: emelkedő, ereszkedő, szinttartó. Az intenzitás tekintetében nem jellemző a hirtelen változás, így az alapfrekvenciában alkalmazott szökő,

illetve eső kategóriákat intenzitásszinten nem használjuk. A korábbi megállapítások a dallammal és a hangintenzitással kapcsolatosan azt sugallják, hogy ott biztosan nagy az intenzitás a mondatban, ahol a magas az alapfrekvencia. A felszólítás, figyelmeztetés indulási értéke magasabb, mint a kijelentésé, ez egyrészt a közlési célból adódik (nagyobb hangerő), másrészt a megvalósításból, hogy valóban magasabb alapfrekvenciával indítjuk, mint egy kijelentést. A kijelentés és a kiegészítendő kérdés intenzitás szempontjából ugyanazt a struktúrát mutatja. Az eldöntendő kérdésben az utolsó előtti szótagon van az intenzitásmaximum. A kis türelmetlenséget kifejező eldöntendő kérdésnél az intenzitás induló értékében van változás, a mondat már a kezdeteknél a referenciaértékhez közeli szinttel indul, és ezen lesz egészen az utolsó előtti szótagig (érthető, hiszen az ilyen kérdések dallama is magasabb értékről indul, ahogy azt a korábbi ábrák is mutatják). Amennyiben több prozódiai egység van a mondatban, akkor az emelkedő, ereszkedő és szinttartó alakzatok formái kombinálódnak. Így a mondat belsejében lebegő forma is előfordulhat. A 6.29. ábra



6.29. ábra. Egy összetett kijelentő mondat sematikus intenzitásszerkezete ereszkedő elemekkel jellemezve

szerinti intenzitásszerkezet nem a végleges forma, ez csak az alapot képezi. A szavak szintjén is értelmezni kell az intenzitás változását. A hangsúlytalan elemeken az intenzitás is jellemzően csökken. Ilyen szerkezetek lehetnek például a kötőszók.

Beszédstílusokra jellemző intenzitásviszonyok. A hírolvasásban a prozódiai egységekre enyhén eső intenzitásváltozás a jellemző. A hangerő a mondat végére – a kiindulási értékhez képest – mintegy 10 dB-t esik. A hírolvasók tehát nem alkalmaznak nagy dinamikatartományt. Ugyanez mondható el a prózafelolvasásról. A mesemondásban nagyobb a dinamika, mint a hírolvasásban. Az intenzitás 20 dB-nyit is változhat. A meséléskor a prozódiai egységek végére jellemző lehet a hangintenzitás emelése (ez általában alapfrekvencia-emelkedéssel is együtt jár). Ilyenkor az intenzitás akár 8–10 dB-lel is megemelkedhet az egység utolsó magánhangzóján. A reklámok vizsgálata változatos képet mutat, függ a reklám céljától, tartalmától. Általánosságban a változatos hangerő alkalmazása a jellemző, az intenzitásértékek az erősen hangsúlyozott részeken akár 20 dB-es átfogási sávban is mozoghatnak. A gyógyszerek reklámozásához tartozó egységes tájékoztató – általában elhadart – mondatában viszont nem találtunk lényeges intenzitásváltozást, a bemondó szinte ugyanazon a hangerőn mondta végig minden vizsgált mintában

a mondatot. Ebben a mondatban 5–7 dB-re csökkent a dinamikatartomány. A fentiekből azt a következtetést vonhatjuk le, hogy az érzelmek kifejezésekor nagyobb dinamikatartománnyal beszélünk, mint a semleges beszédben (Olaszy 2005). A beszéd teljes dinamikatartománya a suttogástól a dühös kiabálásig elérheti a 60 dB-es átfogást is (Laziczius 1944).

6.4. Időszerkezeti tényezők

A beszéd alkotóelemei közül talán az időszerkezetre vonatkozóan született a legtöbb kutatási eredmény, mivel ez egyszerű eszközökkel is kutatható volt már több mint száz éve is. Ebben a fejezetben a legújabb kutatási eredményeket ismertetjük. A beszéd időszerkezetének hatóköre szupraszegmentális szempontból a szöveg szintjétől a szótagon át a beszédhang szintjéig terjed. Ez azt jelenti, hogy az esetleges elemzésekhez, modellezéshez a beszédjelben meg kell jelölni a mondathatárokat, azokon belül pedig a prozódiaiegység-, a hangsúlyfrázis-, a szókapcsolati, a szó- és a szótaghatárokat. Az időtengely felosztása mindig attól függ, hogy mit szeretnénk vizsgálni, modellezni (hangidőtartam-változást, a hangsúly belső szerkezetét, szóhosszúságokat, dallamformákat stb.). A beszédtechnológia fejezetben erre több példát is fogunk találni. Meg kell továbbá jegyezni, hogy a beszéd hullámformáján mért fizikai időértékek és a percepciós élmény között nincs egyértelmű megfelelés (Gósy 1991). Ez azt jelenti, hogy például egy fizikailag mérhető szünetet a hallgató nem észlel szünetnek (egy meghallgatásos kísérletben például nem jelöl szünetet az adott helyen a szövegben).

6.4.1. Artikulációs sebesség

Ez a fogalom azt fejezi ki, hogy milyen gyors az artikuláció, időegység alatt hány beszédhangot ejtünk. Értékét többnyire hang/s mértékegységben adják meg. Mivel kifejezetten az artikulációra fordított időre vonatkozik, mérése során a beszédben tartott szüneteket nem veszik figyelembe. A beszéd eme temporális mutatóját nagyobb beszédegységre vonatkoztatva (sok mondat) szokták mérni és megadni, hogy viszonylag pontosan fejezze ki az artikulációs aktivitást. Az artikulációs sebesség függ a személytől, a korától, valamint a beszédformától, a témakörtől. Például különböző eseménydinamikával rendelkező szövegek (történetmondás) összevetéséből kiderült, hogy a cselekményközpontú szövegek artikulációs tempója gyorsabb az értékelésközpontúakénál (Andó 2002). A gyermekek és az idősek artikulációs sebessége lassabb, a fiatalok és középkorúaké gyorsabb. A magyar beszéd átlagos artikulációs sebessége 12–14 hang/s között mozog. Az eddigi kutatásokat reprezentáló adatok

a következők: Fónagy–Magdics (1960) 12,67 hang/s; Gósy (1991) 14; Gocsál (2000) 13,6; Kovács (2000) 12,34; Olaszy (2006a) 13. A spontán beszédben kissé növekszik ez az adat. Markó (2005) négy beszélő spontán narratívájában a következő átlagos artikulációs sebességeket mérte: 14,11; 14,12; 15,61; 12,45 hang/s. A lassú beszédben 10 alatti artikulációs sebesség is mérhető, a gyorsban 15 feletti is, a nagyon gyors beszédben 20–25-ös értékek is előfordulhatnak. Az artikulációs sebességnek határt szab az artikulációs mozgások végrehajtásához szükséges idő. Ezért a 30-as érték tekinthető a maximumnak, felette már torzulnak a hangok. Ez azt jelenti, hogy egy beszédhangra ilyenkor átlagosan 33 ms idő jut, itt már jelentős torzulások léphetnek fel. A hangidőtartamok mérésénél mindig meg kell adni, hogy az eredmények milyen artikulációs sebességre vonatkoznak. A beszédkutatásban egyes kísérletekhez szavak listáját olvastatják fel, az ebből származtatott beszédjelet használják a mérésekhez, más vizsgálatokhoz mondatokat, egybefüggő szövegeket olvastatnak fel. Az ilyen beszédprodukciók áltagos artikulációs sebességadatai jelentősen különböznek (vö. 2.3. fejezet). Az egyén artikulációs sebességét nagyrészben a beszédképzési automatizmusok határozzák meg. Ezért a beszélő személy nemigen tudja a saját artikulációs sebességét tudatosan megváltoztatni, néhány mondat után visszaáll a rá jellemző értékre. Egy kísérletben a képzett beszélők és a köznapi adatközlők tudatos artikulációs sebesség változtatási képessége 3:1 arányú volt a képzett beszélők javára (Laczkó 1993). Tanulással, képzéssel tehát elsajátíthatjuk azt a képességet, hogy változtatni tudunk tudatosan az artikulációs sebességünkön (színészi beszéd). A tempó tudatos változtatásának kérdéseivel foglalkozik Bóna (2006) tanulmánya, és egész monográfia tárgyalja a gyors beszéd produkciós és percepciós sajátosságait (Bóna 2009).

6.4.2. Beszédtempó

A beszéd hangzó részek és szünetek folyamatos váltakozása. A mondatok között és a mondatokon belül is tartunk szüneteket. Ezért, ha a teljes beszédprodukciót szeretnénk valamilyen sebességi adattal jellemezni, akkor a teljes körű jelre kell vonatkoztatnunk az időegység alatt elhangzó beszédhangok számát. A beszédtempó mértékegysége szintén a hang/s. Ez mindig kisebb, mint az artikulációs sebesség értéke. A kettő közötti különbség kifejezi az egyén beszédstílusát. Különböző szövegtípusok felolvasásából származó adatokat mutatunk be a 6.2. táblázatban. Látható, hogy mennyire különböznek a beszéd sebességi jellemzői az egyes szövegtípusok felolvasásánál. A reklám esetében például a beszélő lényegesen kevesebb szünettartással él, mint a többi esetben. Ez az egyik lényeges jellemzője ennek a felolvasási stílusnak. A beszédtempót meg szokták adni szótag/perc, illetve szó/perc mértékegységgel is.

6.2. táblázat. A beszéd sebességi mutatói különböző szövegtípusok felolvasásánál Olaszy (2005) alapján

Mérés	Hírek	Mese	Novella	Reklám
Beszédtempó hang/s	11,2	9,3	10,4	14,5
Artikulációs sebesség átlaga hang/s	14,0	11,8	13,2	15,4

6.4.3. Szünetek

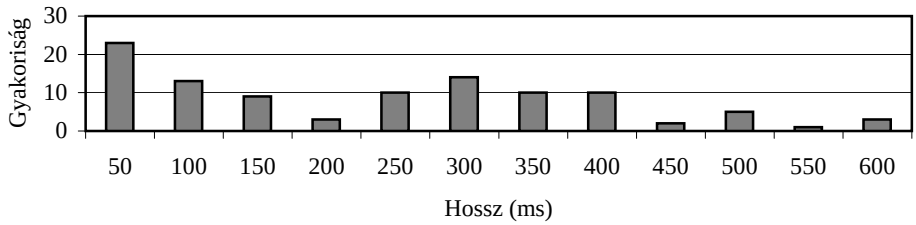
A szünetnek több funkciója van a beszédben. Szende (1976) szerint egyrészről a lélegzetvételhez biztosít időt, másrészről értelmi tagolási funkciót tölt be. Egy különleges tagolási mód a tudatos szünettartás (junktúra) annak érdekében, hogy ugyanazt a szöveget többfélképpen szakítsuk meg, ezzel az értelmezése megváltozzon. Jó példa erre az esztergomi érsek Bánk bánhoz írt levelében fellelhető, klasszikussá vált mondat: *A királynőt megölni nem kell félnetek jó lesz ha mindnyájan beleegyeztek én nem ellenzem.* A legújabb szakirodalom számos más szünetfajtát is megnevez, például gondolkodási szünet, hatásszünet, hezitálás, hangnyújtás, megakadási jelenségek (Gósy 2000a). A szünetek fizikai volta és az észlelés közötti kapcsolatrendszer is fontos kutatási terület (Menyhárt 1998). A beszédtechnológiában a szünet kezelése az egyik legproblematikusabb terület. A beszédszintézisnél a jó szünetmodell természetesebbé teszi a szintetizált beszédet, a gépi beszédfelismerésnél pedig a csend-beszéd detektálás, illetve a jelkimaradás szünetként való felismerése a problémás. Jelen fejezetben csak a néma szünetekről beszélünk.

A beszédszünetek száma és hossza több tényezőtől függ. Elsősorban a beszélő személy veleszületett sajátosságai határozzák meg, aztán a téma, a beszédforma (felolvasás vagy spontán), a beszédhelyzet (előadás, társalgás). A spontán beszédben például a szünettartás gyakoribb, mint a felolvasásban (Gósy 2004b). Első alapvető kérdéskör a szünettel kapcsolatosan a jelkimaradás hossza. A minimális hosszra a legújabb kutatások a 30 ms-ot jelölik meg, a leghosszabb szünet pedig akár több másodperces is lehet. A legrövidebb szünetnek észlelt jelkimaradás ezek szerint kisebb, mint a beszédhangok átlagos hossza (75 ms). Az átlagos szünethossz a magyarban néhány száz milliszekundum. Egy hírolvasásokat elemző kísérletben hat rádióbemondó felolvasott hírblokkjának beszédében mérhető szüneteket elemezték (Olaszy 2006a). Az összesítést a 6.3. táblázatban mutatjuk meg. Az adatok azt mutatják, hogy a hírolvasásban a teljes blokkra vonatkoztatva 20% körüli az az idő, amelyik szünetként realizálódik. A mondaton belüli szünetek átlagából az látszik,

6.3. táblázat. Szünetek hossza hírolvasásban 4 férfi (F) és 2 női (N) bemondó beszédéből

Mérés	F-1	F-2	F-3	F-4	N-1	N-2
Mondatok száma (db)	14	28	12	30	10	13
Szünetek száma a mondaton belül (db)	8	29	16	27	16	8
Szünet átlaga a mondatban (ms)	191	289	264	163	373	283
Szünet átlaga a mondatok között (ms)	1450	1300	1400	1500	1444	1300
Szünet % a hírblokkra vetítve	20	22	20	22	17	16,8

hogy személyfüggő a szünettartási stratégia a hírolvasók között. A mondatok közötti szünetek átlagában már nincs nagy eltérés. A szünetek összevetése szerint a mondatok között lényegesen nagyobb szüneteket tartanak a bemondók, mint a mondaton belül. A szünethosszak eloszlására készített diagram (6.30. ábra) megmutatja, hogy milyen jellemző szünetek fordultak elő a mondatok belsejében a hat bemondó produkciójában. A grafikonból látható, hogy a leginkább jellemző szünethosszak a



6.30. ábra. A vizsgált hat bemondó hírolvasásából mért 103 mondat belseji szünet időtartameloszlása 50 ms-os idősávokra átlagolva

50–100 ms, valamint a 250–400 ms körüli értékek. A szünetekről bővebben Gósy (2004b) munkájában találhatók adatok.

6.4.4. Ritmus

A folyamatos beszédben az artikulációs sebesség rendszerszerű változása, valamint a szünetek hossza és gyakorisága együttesen ad egyfajta ritmikai élményt. A ritmus kialakításában lényeges szerepe van a hangsúlyozásnak is. Egységes nyelvészeti felfogás, hogy a ritmus az időzítéssel kapcsolatos, és nyelvi alapeleme a szótag (Gósy 2004b). A ritmus adja meg a beszéd természetességét, hiánya (például beszédszintézisben) monotonná, nehezen értelmezhetővé teszi a beszédet. A ritmus modellezése nehéz. Fonológiai vonatkozású elemzést Varga (1993) végzett a magyar beszédre. Szavakra és szókapcsolatokra vonatkoztatott fonológiai rendszerezéssel próbálta meg jellemezni a ritmikai változásokat. Fonetika vizsgálatokban Gósy (2000a) szótagok vizsgálatával próbált meg kimutatni valamilyen ritmusosan ismétlődő szabályszerűséget. A szótagok időtartamai azonban alig mutattak rendszerezhető ritmikai változást. Olaszy (2001a) a szavak szintjén állapított meg úgynevezett szóritmust, ami a szó hosszúságától és belső hangtartalmától függ. Modelljében a szavakra jellemző úgynevezett hangidőtérképeket állapított meg, amelyek segítségével bármely szó belső ritmusa (időszerkezete) algoritmussal meghatározható. A modellt nagy mennyiségű szóanyagra (1,5 millió szóalak) alkalmazta és az elkészített szóadatbázist közzétette az interneten (`http://fonetika.nytud.hu/hitint`). Egy másik mérésben (Olaszy 2006a) nagyobb beszédegységeken végzett mérések során

kimutatta, hogy a beszélők széles sávban és kváziperiodikusan változtatják az artikulációs sebességüket. A hírolvasások elemzéséből kapott adatok azt mutatják, hogy a bemondók a szerkezethatár előtt az artikulációs tempójukat lassítják, vagy szinten tartják, a szöveg tartalmától függően, a lényeges információt hordozó beszédszakaszokban a sebességet lassítják, ezzel kiemelik a részt, majd gyorsítják (de előfordult fordítva is, a lényeg a váltakozás kialakítása). A hírolvasások vizsgálatából mutatunk be két mondatot, amelyben az összetartozónak vélhető szövegegységekre mért artikulációs sebesség tudatos változtatása látható (az artikulációs sebességet zárójelbe tett számadat jelzi hang/s-ban).

Az osztrák hatóság (14) *indokoltnak látta* (15) *a sikkasztás és a* (10) *hamisított okmányok* (14) *felhasználása miatti* (13) *magyar kiadatási kérelmet* (12).

Az Országgyűlés (14) *nem változtatott* (16) *a tavaly elfogadott* (15) *személyi jövedelemadó-táblán* (13).

A ritmus lényeges eleme a szünetek elhelyezkedése és hossza, valamint az egymást követő szünetek hosszértéke is. A korábban említett híranyagra alapozva végeztek percepciós kísérletet (Olaszy 2006a). A cél annak vizsgálata volt, hogy a mondat belsejében tartott szünet hossza mennyire van hatással a ritmikai élményre a hallgatóban. Van-e ideális szünethossz, ami a hallgatónak fontos a hír percepciós feldolgozásához? Ez a szünethossz egyezik-e a bemondó által produkált szünethosszal? A hipotézis az volt, hogy bizonyos határok között a hallgató nem érzékeny a szünethossz megváltoztatására. A percepciós kísérlethez 5 mondatot választottak ki, egyetlen hírolvasó felolvasásából. Szempont volt, hogy a mondat két egységet tartalmazzon, amelyeket levegővételi szünettel választ el egymástól a bemondó (például: *Embereken is kipróbálják azt az oltóanyagot ; (335 ms) amelyet az Ebola-vírus ellen fejlesztettek ki az Egyesült Államokban.*). Ezen alapmondatokban a szünethosszakat mesterségesen megváltoztatták (csak a szünet hosszát), így preparált mondatok is rendelkezésre álltak. Minden alapmondatból további négy szünethosszúságú mondatot készítettek úgy, hogy az eredeti szünetet 25%-osra, 50%-osra, 150%-osa és 200%-osra változtatták meg. Így ötfajta mondat állt rendelkezésre minden mondatcsoportban, a teszt tehát 25 mondatból állt. Ezeket véletlenszerű sorrendbe rendezték, majd meghallgattatták 8 férfival és 7 nővel (életkoruk 22–45 év).

A feladat meghatározása a következő volt:

A mondat meghallgatása után döntse el, hogy a hallott mondatot ritmikai szempontból melyik kategóriába sorolja: 5 = nagyon jó, 4 = jó, 3 = közepes, 2 = nem jó. A 3-as és a 2-es ítélet mellé értékelést is kértünk. A teszt eredményeit a 375 válaszra vonatkozóan a 6.4. táblázat tartalmazza. Az eredmények szerint az eredeti szünetek lerövidítése egyáltalán nem zavarta a hallgatókat. A szünet nyújtásánál csupán a 200%-os értékre nyújtott szüneteknél tettek számottevő kifogásokat. Az eredményekből azt a fontos következtetést vontuk le, hogy a hírolvasásnál a levegővételi szünet hossza, ha azt az eredeti hossz 25–150%-os értékén belül bármilyen értékre változtatjuk, nem befolyásolja lényegesen a beszéd ritmikai megítélését. A kísérlet

6.4. táblázat. Az ítéletek száma az osztályzatok függvényében a szünethosszak szerint rendezve

Mérés	Csoport					Összes	%
Szünethossz-osztályzat	25%	50%	100%	150%	200%		
5	76	83	89	61		309	82,4
4				37		37	9,8
3				4	17	21	5,7
2					8	8	2,1

eredményei azt valószínűsítik, hogy a beszédritmus élményének kialakításában az artikulációs sebesség változtatása és a hangsúlyozás nagyobb szerepet kap, mint a szünethosszak. A témakör teljes feltárásához még további kutatások szükségesek.

6.5. A hangszínezet

Minden beszélőnek van egyéni hangszínezete (ez alapján ismerjük fel ismerőseink hangját akár egyetlen szóból is). Az egyéni hang jellegzetességeit két alaptényező alakítja ki. Egyrészről a zöngehang formája (időfüggvénye), amely a zöngét felépítő frekvenciakomponenseket határozza meg (gazdag felharmonikusokban vagy esetleg szegény), másrészről az egyén artikulációs csatornájának méreti tulajdonságai továbbá a személyre szabott artikulációs mozgások. A normál beszéd hangszínezete (ahogy a hétköznapokban beszélünk), tehát alapvetően a szegmentális szerkezethez tartozik. Már évtizedekkel ezelőtt felmerült a kérdés, hogy a kettő közül melyik paraméter veszi ki nagyobb mértékben a részét az egyén jellemző hangszínezetének kialakításában. Egy magyar fonetikai kísérletben (Olaszy 1978) azt mutatták ki, hogy az artikulációs csatorna jellegzetességei nagyobb jelentőségűek, mint a zönge rezgésformája. Igaz, ezt a megállapítást egyetlen magánhangzó vizsgálata alapján tették. Természetesen szupraszegmentális tényezők is meghatározzák, hogy valakinek a hangját mennyire biztosan ismerjük fel (hangsúlyozás, tempó, dallamos vagy kevésbé dallamos a beszéd, szünettartás, tagolási formák stb.). Ilyen hangváltoztatást akkor alkalmaznak a beszélők, amikor érzelmeket vagy szituációt akarnak kifejezni (mesemondás, prózafelolvasás, veszekedés). Az ilyen hangzásokat a szakirodalom egyenlőre csak metaforákkal tudja jellemezni: érdes, bársonyos, lágy, éles, dühös, rekedtes, tompa, csengő, mérges, rémült, szomorú, levegős, suttogó, vidám, halovány stb. Mindezeket a hangzásokat a spektrum különbözőségével le lehetne írni. Így fizikai paraméterekkel is jellemezhetnénk az egyes hangtípusokat. Ilyen téren komoly eredmények még nem születtek, csak kísérleteznek a kutatók. A beszédtechnológia tehát még nem tudta modellezni a hangszínezet hatásmechanizmusát. Egyes beszédszintézis-technológiák már tudják imitálni az adott beszélő hangszínezetét, de csak a tőle vett minták alapján, nem modellezéssel. Az ilyen statisztikai alapú eljá-

rások azonban nem nyújtanak olyan ismereteket, amelyekkel parametrikus szinten lehetne kategorizálni a különbségeket.

BESZÉDTECHNOLÓGIA

7. fejezet
A beszédtechnológia tudománya

A beszédtechnológia a beszédkutatás és a technológia ötvöződéséből kialakult új tudomány, amely a 20. század utolsó harmadában indult komolyabb fejlődésnek. A beszédtechnológia definíciószerű meghatározása a következő: az a tudomány, amelyik az emberi beszédtevékenység körfolyamatából valamely komponens(ek) modellezésével és gépi megvalósításával foglakozik. A sikeres beszédtechnológiai fejlesztésekhez komplex, több tudományterületet átfogó szaktudás szükséges (nyelvészet, matematika, fonetika, beszédakusztika, jelfeldolgozás, számítástechnika, pszichológia stb.).

Az információs társadalomban számítógépek vesznek körül bennünket, a mikrochiptől kezdve a mobil eszközökön keresztül az ipari rendszerekig. Alapvető követelménnyé vált, hogy ezekkel a rendszerekkel szóban kommunikáljunk. A beszédtechnológia lényeges szerepet kap az infokommunikációs rendszerekben, egyrészt automatizált gépi szolgáltatásokat tesz lehetővé (automatikus tudakozó), másrészt kényelmi szolgáltatásokat is nyújthat (beszédalapú tárcsázás, sms-felolvasás, sms-diktálás). A tömeges nyelvi információ (beszéd és szöveg) automatikus vizsgálatára is egyre nagyobb szükség van (keresés, kivonatolás, lényegkiemelés). Az emberi beszéd részletes vizsgálata új orvosdiagnosztikai eszközök létrehozását is előrevetíti. Ezzel a három példával csak érzékeltetni szerettük volna, hogy milyen széles területet fed le a társadalomban a beszédtechnológia tudománya.

7.1. A beszéd számítógépes feldolgozása

Szaszák György

A beszédkutatást jelentősen fellendítette a digitális jelfeldolgozás fejlődése. Ennek a fejlődésnek az egyik új ága a beszédtechnológia is. A számítógépes feldolgozás azt jelenti, hogy digitálisan tároljuk a beszédjelet (átalakítjuk bitsorozattá) és ettől kezd-

ve ezzel a formával dolgozunk. A visszaalakítás analóg jellé csak akkor történik, ha hallhatóvá akarjuk ismét tenni a jelet. A számítógépek műveleti gyorsaságának rohamos növekedése egyre tágabb teret ad a digitális beszédfeldolgozás legkülönfélébb megoldásainak megvalósítására. A jelfeldolgozási algoritmusok fejlődése is felgyorsult az utóbbi évtizedekben, egyre nagyobb számítási igényeket elégítenek ki valós időben a számítógépek. A számítógép ma már jelen van életünk minden területén (mobil, GPS, PDA stb.). Ez lehetővé teszi azt, hogy beszédtechnológiai fejlesztések eredményeit széles körben használjuk. A következő fejezetekben egyrészről összefoglaljuk a számítógépes beszédfeldolgozás témakörének legfontosabb elemeit, másrészről látni fogjuk azt is, hogy milyen tág tere van a beszédtechnológia alkalmazásának a mindennapi életben.

7.1.1. Mintavételezés, kvantálás, visszaállítás

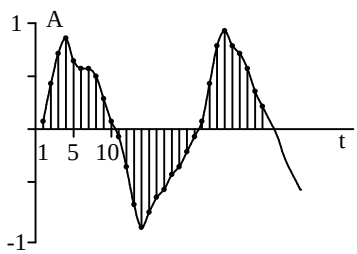
Az emberi beszéd analóg jel, azaz időben folytonosan értelmezett, amplitúdóban pedig folytonos az értékkészlete. A digitális jel ezzel szemben amplitúdóját tekintve véges értékkészletű, időben pedig vagy véges, vagy legfeljebb megszámlálhatóan végtelen pontban értelmezett. A beszédjel digitális feldolgozásához tehát mind az idő, mind az amplitúdó értékeit diszkrétté kell alakítani. Az időben történő diszkretizálás neve mintavételezés, az amplitúdóbeli diszkretizálásé pedig kvantálás. Az időtengelyt tehát a mintavételezési gyakoriság osztja diszkrét intervallumokra, az analóg jel amplitúdóját pedig a kvantálási szintek. Ily módon a mintavételezett, majd kvantált jel egy, a diszkrét mintavételi időpontokhoz tartozó, a diszkrét (kvantált) amplitúdót megadó számok sorozataként értelmezhető. Ha a mintavételezett, tehát időben diszkrét értékkészletű jelből az időben folytonos analóg jelet szeretnénk nyerni, akkor azt folytonos értelmezési tartományúvá kell visszaalakítanunk, ennek neve visszaállítás. A visszaállítás a mintavételezés ellentett művelete, míg a kvantálás esetében nem létezik ilyen egyértelmű ellentétes átalakítás. A kvantálás tehát (és a gyakorlati megvalósításban kisebb mértékben a mintavételezés is) azzal jár, hogy információt veszítünk az eredeti jelből.

A gyakorlatban a mintavételezést, a kvantálást és az esetleges további kódolást az analóg-digitális átalakítók (Analog-Digital Converter, ADC) végzik, amelyek gyakran integráltan tartalmazzák a mintavételezést végző mintavevő és -tartó, valamint a kvantáló/kódoló egységet. Ezt az átalakítást nevezik együttesen digitalizálásnak, amely tehát mintavételezésből és kvantálásból áll. Az ellentétes irányú átalakítást a digitális-analóg átalakító (Digital-Analog Converter, DAC) végzi, amely a diszkretizált amplitúdót reprezentáló számsorozatnak megfelelő, időben folytonos, amplitúdóban kvantált kimeneti jelet állít elő, majd ezen jelből az úgynevezett visszaállító (interpoláló) szűrő hozza létre az időben is és amplitúdóban is folytonos jelet. A min-

tavételezés és kvantálás során figyelemmel kell lennünk a hű, de legalábbis a lehető legjobb visszaállíthatóság követelményére. Másképpen fogalmazva úgy kell elvégeznünk a mintavételezést és a kvantálást, hogy az eredeti jelhez a lehető legközelebb álló jelet kapjunk vissza. Mintavételezés esetén elvileg lehetséges a veszteség nélküli visszaállítás, amennyiben teljesül, hogy a jel sávkorlátozott. Kvantálásnál viszont az eredeti amplitúdóértéket kerekítjük, ebből elkerülhetetlenül adódik valamennyi *kvantálási hiba*, ami a visszaállítás után is megmarad.

7.1.1.1. Mintavételezés

A mintavételezést jellemzően fix időeltolással hajtják végre, azaz a jelből $t_0 + kT$ időközönként vesznek mintát (k= 0, 1, 2, ...). A T az úgynevezett mintavételi frekvencia (gyakoriság) reciprokának megfelelő periódusidő. A mintavételi pontokon a kapott amplitúdóértékek még pontosan tükrözik az eredeti (analóg) jel mintavételi pillanatra vonatkozó amplitúdóját (7.1. ábra). A mintavételi frekvencia (sampling ra-



T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
0,1	0,38	0,75	0,85	0,68	0,63	0,62	0,49	0,3	0,1	-0,08	-0,43

T = mintavételi időpont

7.1. ábra. A mintavételezéssel az analóg jelet időben diszkretizáljuk. A számsorozat az analóg jel első 12 mintavételi pontjának amplitúdóértékeit mutatja

te) megadásával ekvivalens, ha az egységnyi időre (jellemzően egy másodpercre) eső minták számát adjuk meg (samples per second). A mintavételezést akkor végezzük helyesen, ha a mintavételezett jel a mintákból egyértelműen visszaállítható. Könnyen belátható, hogy túl ritkán vett mintákból az eredeti jelformát nem lehet visszaállítani. A kérdés tehát az, hogy milyen feltételeknek kell teljesülnie ahhoz, hogy az eredeti jel visszaállítható legyen. Ennek elégséges feltételét mondja ki a Nyquist és Shannon nevéhez fűződő *mintavételi tétel*. Itt lényeges szerepet kap a jel sávszélessége, azaz a spektrumának legalacsonyabb, illetve legmagasabb frekvenciájú összetevőjének a különbsége. A jel sávszélességét (B)-vel jelölik. A tétel így szól: bebizonyítható, hogy ha a mintavételezést úgy végezzük, hogy a mintavételi frekvencia (f_s) nagyobb, mint a jel sávszélességének kétszerese (alapsávi jel esetén a mintavételi frekvencia legalább kétszerese a jelben előforduló legmagasabb frekvenciájú összetevő frekvenciájának), akkor az eredeti jel helyesen visszaállítható, tehát:

$$f_s \geq 2B \tag{7.1}$$

A matematikai okfejtést a tétellel kapcsolatosan Gordos–Takács (1983) tárgyalja. A mintavételi tétel egyúttal azt is megköveteli, hogy a mintavételezendő jel sávkorlátozott legyen. Ez azt jelenti, hogy meg kell adni a mintavételezésre kijelölt alapsávi analóg jel legmagasabb frekvenciakomponensét (a B felső frekvenciaértékét, a *felső határfrekvenciát*). A gyakorlatban ezt egy megfelelően választott aluláteresztő szűréssel szokták biztosítani. A beszédre optimális esetben 10 kHz lehet a szűrő vágási frekvenciája. Ezzel a beállítással minden lényeges frekvencia-összetevőt átviszünk. Az 1900-as évek elejének műszaki korlátai miatt elhatározták, hogy a távközlési beszédátviteli rendszerekben a 300–3400 Hz-es sávot viszik át. Ez ma is így van. A szűkített átviteli sáv különösen a zár- és réshangok, valamint a zár-rés hangok érthetőségét csökkenti.

A mintavételi tétel vázlatos bizonyításához tekintsük az $X(\omega)$ spektrumú, sávkorlátozott jelet. A mintavételezés maga az $x_\delta(t) = \sum_{n=-\infty}^{\infty} \delta(t-nT)$ Dirac-impulzussorozattal való szorzásként is felfogható. A mintavételezett eredeti $x(t)$ jel $X_s(\omega)$ spektruma ekkor:

$$X_s(\omega) = \int_{-\infty}^{\infty} \left\{ \sum_{n=-\infty}^{\infty} \delta(t-nT) \right\} x(t) e^{-j\omega t} dt \tag{7.2}$$

$$X_s(\omega) = \left(\int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(nT) e^{-j\omega t} dt \right). \tag{7.3}$$

$\omega_s = \frac{2\pi}{T}$ definíciószerűen a körfrekvencia.

Az előbbi (7.2) összefüggésbe a mintavételnek megfelelő impulzussorozat

$$x_\delta(t) = \frac{1}{T} \sum_{k=-\infty}^{\infty} e^{-jk\omega_s t} \tag{7.4}$$

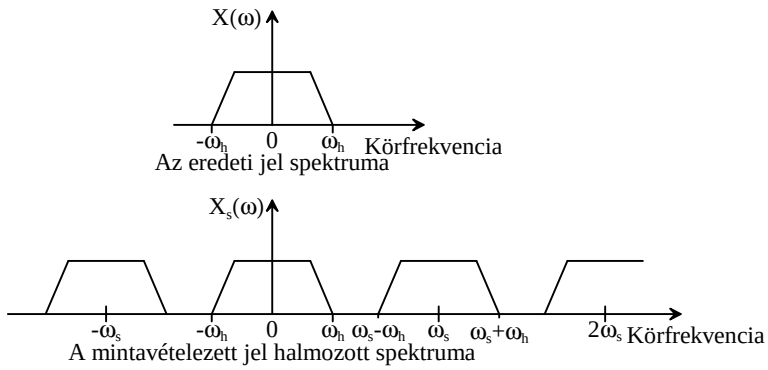
szerint meghatározott Fourier-sorát írhatjuk, így:

$$X_s(\omega) = \int_{-\infty}^{\infty} \left\{ \sum_{k=-\infty}^{\infty} \frac{1}{T} e^{-jk\omega_s t} \right\} x(t) e^{-j\omega t} dt \tag{7.5}$$

$$= \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \frac{1}{T} x(t) e^{-j(k\omega_s - \omega)t} dt \tag{7.6}$$

$$= \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\omega + k\omega_s) \tag{7.7}$$

Ez utóbbi összefüggés adja meg a mintavételezett jel spektrumát, amely tehát a következő komponensekből áll: alapsávi jel, vagyis az eredeti jel spektruma ($X(\omega)$), valamint az alapsávi jel spektrumának a mintavételi frekvencia egész számú többszöröseivel való eltolása (lásd a 7.2. ábrát). Az egészet szokták halmozott spektrumnak nevezni. Innen már jól látható, hogy a visszállított spektrum átlapolódás-



7.2. ábra. Az eredeti (fent) és a mintavételezett (lent) jel spektruma

mentességéhez (ahhoz, hogy ne csússzanak egymásba a halmozott spektrum frekvenciasávjai) az szükséges, hogy a következő feltétel teljesüljön: $f_s > 2B$.

A mintavételezés alkalmazható az oldalsávba eső jelekre is, ekkor a sávkorlátozás és a visszaállítás sávszűrőkkel történik. Részsávos kódolóknál vagy MP3-nál is több frekvenciasávra bontjuk az analóg jelet, és ezeket a sávokat külön kódoljuk (lásd a 7.2. fejezetben). Ekkor természetesen nem a teljes jel, hanem a kiválasztott részsáv átlapolás-mentességét kell biztosítani.

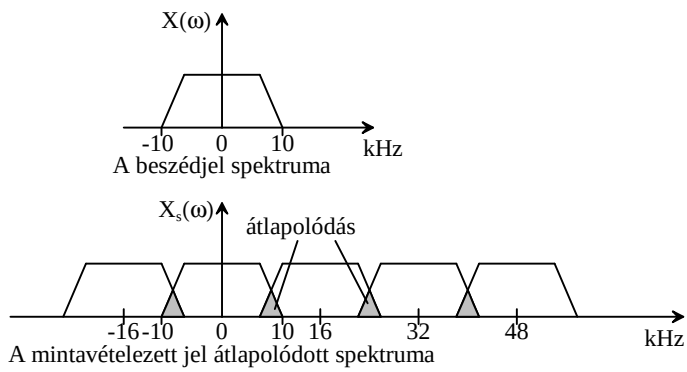
Megjegyzések:

- Az oldalsáv fogalma a modulációs eljárásoknál kerül elő, a moduláció eredményeképpen ugyanis az eredeti tiszta szinuszos vivőfrekvencia két oldalán (tehát a vivőnél valamelyest kisebb, illetve magasabb frekvenciákon is) megjelennek spektrális összetevők, amelyek a vivőfrekvenciának megfelelő érték alatt és felett egy-egy intervallumot határoznak meg. Ezek az oldalsávok.
- A részsáv lényegében egy frekvenciaintervallumnak felel meg, tehát a spektrum egy adott frekvenciaintervallumba eső összetevőit foglalja magába.

Összefoglalva tehát, a mintavételezésnél a visszaállíthatóság feltétele az átlapolásmentesség, továbbá annak az információnak az ismerete, hogy melyik oldalsávban van a hasznos jel, mert a minták ezt az információt nem tartalmazzák, hiszen minden oldalsáv egymásra lapolódik.

A mintavételezés tervezésénél eszerint két út van a mérnök számára: vagy az adott frekvenciasávú analóg jelhez kell igazítani a mintavételi frekvenciaértékét, vagy pe-

dig, ha kötött a mintavételi frekvencia (ilyen a távközlés), akkor az analóg jelet kell megszűrni, hogy ne tartalmazzon magasabb frekvenciakomponenst, mint a mintavételi frekvencia fele. A beszédfeldolgozásban elterjedten használt mintavételi frekvenciák a 8 és 16 kHz. Amennyiben nem tartjuk be a mintavételi tételt, a halmozott spektrumban átlapolódások jöhetnek létre, ami jeltorzítást eredményez. Minél nagyobb az átlapolódás, annál jobban torzul az alapsávi jel. Vegyük a következő példát. Ha a beszédet (amely mintegy 10 kHz-ig tartalmazhat frekvenciakomponenseket) 16 kHz mintavételi frekvenciával digitalizáljuk, akkor a mintavett jel halmozott spektrumában az alapsávi jel 10 kHz-ig fog komponenseket tartalmazni, a 16 kHz-re eltolt spektrum pedig 6 kHz-től 26 kHz-ig. Belátható, hogy az eltolt spektrum átlapolódik az alapsávi jel 6–10 kHz-es sávjával (7.3. ábra). Ez torzítást fog okozni az alapsávi jelben, ami hallható. Az átlapolódásból származó jeltorzulás utólagosan semmilyen eljárással nem korrigálható, mert nem választhatók szét az eredeti és az átlapolódó komponensek. Az átlapolódás megszüntetésére az analóg jelet minden



7.3. ábra. Az eredeti (fent) és a mintavételezett (lent) jel átlapolódott spektruma

esetben célszerű egy átlapolódásmentetsítő aluláteresztő szűrővel (antialiasing filter) megszűrni, mielőtt mintavételezzük. Ennek a szűrőnek a vágási frekvenciáját a mintavételi frekvenciához igazítják. A 16 kHz-es mintavételezéshez célszerű 7,7 kHz-es, a 8 kHz-eshez legfeljebb 3,7 kHz-es vágási frekvenciát beállítani. A vágási frekvenciák azért alacsonyabbak, mint a mintavételi frekvencia fele, mert a szürő vágási meredeksége nem lehet végtelen. A hagyományos vezetékes távközlés (PSTN) felső sávkorlátja ilyen okból 3400 Hz. Abban az esetben, ha nem áll rendelkezésre kellő meredekségű aluláteresztő szűrő, szándékosan túlmintavételezést (oversampling) is el lehet végezni. Ha a mintavételezett jelből időben analóg jelet szeretnénk visszaállítani, akkor a jel értékét meg kell határoznunk a mintavételi időpontok közötti időpontokban is. Bebizonyítható, hogy a

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT) \frac{1}{\pi(t-nT)} \sin\left(\frac{\pi(t-nT)}{T}\right) \tag{7.8}$$

összefüggéssel minden köztes t időpontra meghatározott értékek együttesen az eredeti jelet szolgáltatják vissza, amennyiben a mintavételezett jel sávkorlátozott volt, és betartottuk a mintavételi frekvencia megválasztására vonatkozó feltételt is. Ennek a műveletnek a neve *interpoláció*, amelyet a műszaki gyakorlatban aluláteresztő szűrővel, az úgynevezett interpoláló szűrővel valósítanak meg. Az interpoláló szűrő tehát egy $\frac{sin(t)}{t}$ impulzusválaszú (azaz ideális aluláteresztő) szűrő. A visszaállított jel úgy áll elő, hogy a mintavett jelsorozatot az interpoláló szűrő impulzusválaszával konvolváljuk:

$$x(t) = \left(\sum_{n=-\infty}^{\infty} x(nT)\right) * \left(\frac{1}{\pi t} \sin\left(\frac{\pi t}{T}\right)\right). \tag{7.9}$$

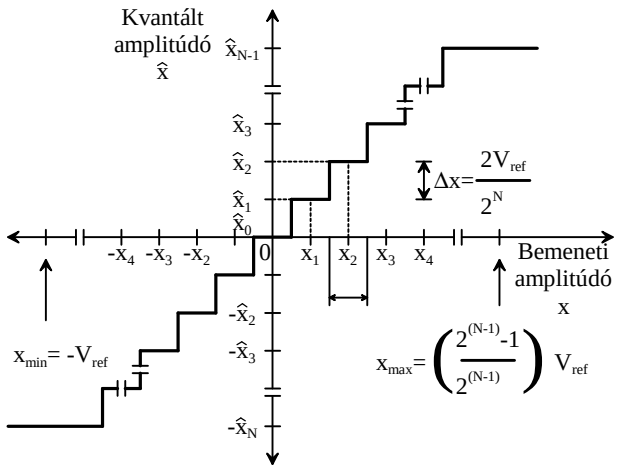
A konvolúció eredményeként a mintavételezett eredeti jelet pontosan visszakapjuk. Az interpoláló szűrő feladata, hogy az egyes mintavételi időpontok között a jelet regenerálja, azaz folyamatosan értékekkel töltse ki a minták közötti tartományt.

Megjegyzés: a konvolúció egy matematikai művelet, amely definíció szerint az aláábiakban határozható meg $f(t)$ és $g(t)$ függvények konvolúciójára:

$$f(t) * g(t) = \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau \; = \; \int_{-\infty}^{\infty} f(t-\tau) g(\tau) d\tau. \tag{7.10}$$

7.1.1.2. Kvantálás

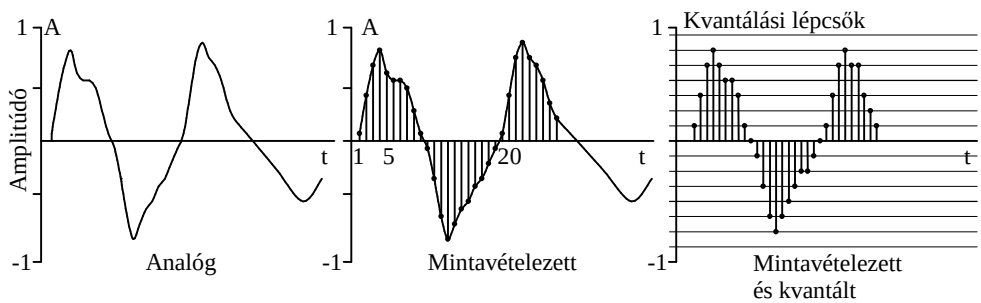
A kvantálás a jelamplitúdó diszkretizálására szolgál. A mintavett jel amplitúdója ugyanis bármely értékű lehet, ez csak az analóg jeltől függ. Ahhoz, hogy egységesíteni lehessen bármilyen analóg jel digitális feldolgozását, az amplitúdó vonatkozásában is előre meghatározott értékkészletet kell megadni. Az amplitúdó teljes kivezérlési tartományát diszjunkt intervallumokkal fedjük le, ezeket kvantálási lépcsőknek nevezik. A mintavett jel pillanatnyi amplitúdóértékét a továbbiakban a hozzá legközelebb eső kvantálási lépcső fogja képviselni. Ha az amplitúdóérték-készletet egyenlő hosszúságú intervallumokra bontjuk, és a kvantált érték az egyes intervallumok átlagértéke, akkor a kvantálás egyenközű. A kvantálási lépcsőkhöz tartozó amplitúdótartományok a kvantálási szintek. Mivel jellemzően binárisan kódoljuk a kvantált értékeket, ezért N biten történő kvantáláskor az értékkészletet 2^N lépcsőre osztjuk. 2-es komplemens kód használatakor a pozitív értékeket $2^{N-1}-1$, a negatívakat 2^{N-1} szinten kvantálhatjuk. Ha feltételezzük, hogy a kvantálandó jel legnagyobb, illetve legalacsonyabb értéke V_{ref}, illetve. $-V_{ref}$, akkor lineáris kvantálás esetén egy kvantálási lépcső (Δx):



7.4. ábra. N bites egyenközű (lineáris) kvantálás kvantálási karakterisztikája

$$\Delta x_i \;=\; \hat{x}_{i+1} - \hat{x}_i \;=\; \frac{2V_{ref}}{2^N} \tag{7.11}$$

minden i-re. A kvantálás során tehát egy adott kvantálási intervallumba eső minden értéket azonos szintre kvantálunk (az i-edik kvantálási intervallumba, azaz az $[x_i, x_{i+1}]$ intervallumba eső valamennyi értéket $\hat{x}_i$ szintre). A 7.4. ábrán ezt a bemeneti amplitúdó tengely x_2 értéke alá rajzolt tartománnyal érzékeltetjük. A mintavételezéskori amplitúdóértékek tehát meg fognak változni a kvantált értékekre (7.5. ábra). A kvantált jelet soha nem lehet visszaállítani a mintavett jel eredeti formájára. Az eredeti és a kvantált jel között irreverzibilis különbség, úgynevezett kvantálási hiba jelentkezik, ami torzításként fogható fel. N-bites egyenközű kvantálás esetén a



7.5. ábra. Az eredeti analóg jel, a mintavételezés utáni állapot (középen) és a mintavételezés után elvégzett kvantálás eredménye

kvantálási hiba értéke abszolút értékben legfeljebb $\frac{V_{ref}}{2^N}$. Kellően sok kvantálási szint esetén a kvantálási hiba közel egyenletes eloszlásúnak feltételezhető, melynek sűrű-

ségfüggvénye $f(e) = \frac{2^N}{2V_{ref}}$ (konstans). A kvantálás négyzetes hibája (e^2) ekkor:

$$e^2 = \int_{-2-nV_{ref}}^{+2-nV_{ref}} e^2 f(e) de \tag{7.12}$$

$$= \int_{-2-nV_{ref}}^{+2-nV_{ref}} e^2 (\frac{2^N}{2V_{ref}}) de \tag{7.13}$$

$$= \frac{2^{-2N} V_{ref}^2}{3}. \tag{7.14}$$

Amennyiben a kvantálandó jelet a kvantálási tartományt teljesen kihasználó (maximálisan kivezérelt) szinusznak feltételezzük (amplitúdója V_{ref}), akkor a jel-zaj viszony:

$$SNR = \frac{x^2}{e^2} = \frac{\frac{V_{ref}^2}{2}}{\frac{2^{-2N} V_{ref}^2}{3}} = 1{,}5 \cdot 2^{2N}. \tag{7.15}$$

Decibelben kifejezve:

$$SNR_{dB} = 10 \lg(1{,}5 \cdot 2^{2N}) = 10 \lg 1{,}5 + 2N \lg 2 = 1{,}76 + 6{,}02N. \tag{7.16}$$

A kvantálási hiba tehát 6 dB/bit-tel csökkenthető a bitek számának eggyel való növelésével. Ez azt jelenti, hogy a telefóniában a 8 kHz-es, 8 bites mintavételezés 48 dB-es jel/zaj viszonyt biztosít egyenközű kvantálás és szinuszos jel teljes kivezérlése esetén. A egyenközű kvantálás nagy hátránya, hogy a kisebb jelszintű részeket ugyanakkora hibával kvantálja, mint a nagyobbakat, azaz a kvantálási hiba nem arányos az amplitúdóval (a relatív hiba a kis jelszintű részeken nagyobb). Ez a beszédben tipikusan a kisebb energiájú beszédhangok egyenközű kvantálása során zavaró, hiszen a kvantálási hiba esetleg összemérhető lesz a jelamplitúdóval, ami torzulásként jelentkezik. A kivezérlés sem teljes minden esetben, ebből adódik, hogy a jeltartomány egésze nincs kihasználva, az amplitúdótartomány alsó és középső szintjei sokkal gyakrabban használtak, mint a maximális szint. A kvantálási hiba arányos elosztására jött a gondolat, hogy nem lineárisan (egyenközzel) kell kvantálni. Ott, ahol nagy a relatív hiba, kisebb kvantálási lépcsőket kell alkalmazni, máshol nagyobbakat. A gondolat meg is fordítható. A kvantálási intervallumok módosítása helyett az eredeti, kvantálandó analóg jelet is módosíthatjuk (előtorzíthatjuk) oly módon, hogy az alacsonyabb jelszintű szakaszokat jobban, míg a nagy energiájú részeket kevésbé erősítjük (dinamikakompresszió). Ezután pedig alkalmazhatjuk már az egyenközű kvantálót. Arról azonban gondoskodnunk kell, hogy a jel vissza-

állításakor az inverz amplitúdótorzítást is el kell végeznünk. Az előtorzítást végző függvényt gyakran *kompresszor*nak, míg a torzítás korrigálását végzőt *expander*nek hívják. A beszédjel nem egyenközű kvantálásához a telefóniában logaritmikus kompresszorfüggvényeket alkalmaznak (logaritmikus kvantálás). A világon két különböző, de széles körben használt karakterisztika terjedt el: a μ-törvényű, és az A-törvényű. Előbbit például az Amerikai Egyesült Államokban és Japánban, utóbbit Európában használják. Az A-törvényű karakterisztika kompresszorfüggvénye ($g(x)$) például:

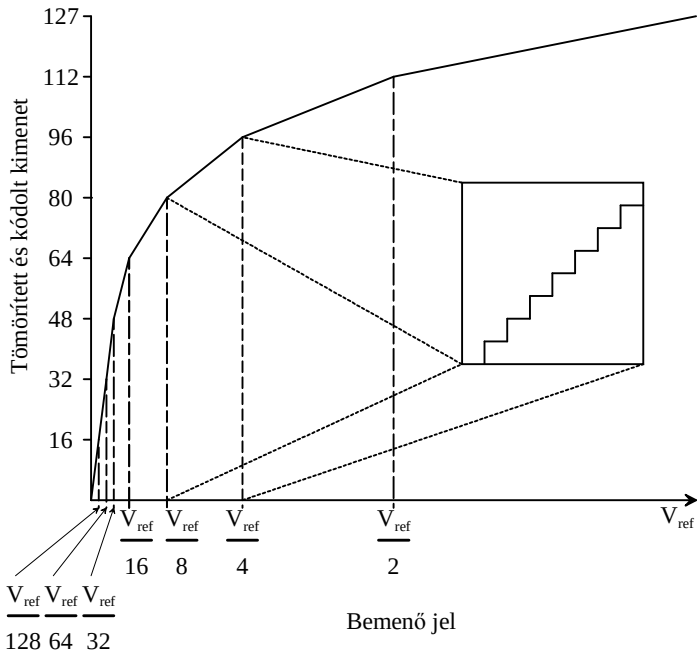
$$g(x) \;=\; sgn(x) \begin{cases} \frac{A|x|}{1+\ln A} & \text{, ha } |x| < \frac{1}{A} \\ \frac{1+\ln(A|x|)}{1+\ln A} & \text{, ha } \frac{1}{A} \leq |x| \leq 1 \end{cases} \tag{7.17}$$

Az expander:

$$g^{-1}(y) \;=\; sgn(y) \begin{cases} \frac{|y|(1+\ln A)}{A} & \text{, ha } |y| < \frac{1}{1+\ln A} \\ \frac{e^{|y|(1+\ln A)-1}}{A} & \text{, ha } \frac{1}{1+\ln A} \leq |y| \leq 1\ . \end{cases} \tag{7.18}$$

A fenti összefüggésekben az A konstans a kompresszió mértékét befolyásolja, értékét az ITU-szabvány G. 702 ajánlása $A = 87,56$-nak rögzíti, a $sgn(.)$ pedig az előjelfüggvény ($sgn(x) = \frac{x}{|x|}$, ha $x \neq 0$ és $sgn(0) = 0$) (Gordos–Takács 1983). A gyakorlatban sokszor a $g(x)$-függvényt szakaszonként lineáris függvények sorozataként hozzák létre (lásd a 7.6. ábrán). Az A-törvényű 8 bites logaritmikus kvantálás megfelel egy 12 bites egyenközű kvantáló minőségének (az ábrán a 0 és $\pm\frac{V_{ref}}{64}$ közé eső jelszintekre). A legmagasabb jelszintekre (abszolút értékben $\frac{V_{ref}}{2}$ és V_{ref} között) a felbontás már csak 6 bites egyenközű kvantálással egyenértékű, azonban a teljes beszédjelre vetített jel-zaj viszony, illetve a visszaállított beszéd minősége jobb, mint amit egyenközű, más néven lineáris kvantálással kapnánk.

A 8 bites logaritmikus kvantálást alkalmazzák a hagyományos digitális telefonvonalakon, a 8 kHz-en mintavételezett és 8 bitre kvantált beszédjel így 64 kbps átviteli kapacitást igényel, és 72 dB-es jel/zaj viszonyt jelent. A nem egyenközű kvantálás tehát 24 dB-es javulást eredményez. Mind a lineáris, mind a logaritmikus kvantálásra igaz, hogy nem képesek kezelni az egyes beszélőn belüli hangerőváltozásokat, illetve a beszélők közötti jelszint-, azaz amplitúdóbeli változásokat. Optimális esetben a legnagyobb elfogadható/várható hangerőhöz kellene igazítani a kvantálási tartományt. Hasonló a problematika a beszéddel magával is, amelyben kisebb és nagyobb amplitúdójú részek váltakoznak, az egyes minták közötti korreláltság azonban igen nagy (a beszédet folyamatosan képezzük, a beszédben az amplitúdó változásai is leginkább folyamatos erősödés vagy lecsengés jellegűek). A beszéd e mindkét, az imént felvetett jellegzetessége egyben sugallja is a lehetséges megoldást: a kvantálási tartományt dinamikusan változtatva azt adaptívvá tehetjük az aktuális amplitúdóviszonyokhoz, illetve a magas korreláltság miatt érdemes lehet a minták közötti



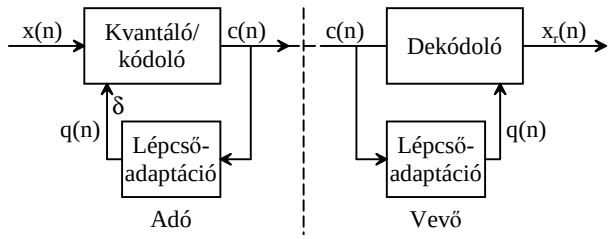
7.6. ábra. Az A-törvényű logaritmikus kvantálás kompresszorfüggvényének gyakorlati megvalósítása szakaszonkénti lineáris közelítésel, a szakaszokon belül 8 egyenközű kvantálási lépéssel

különbséget kvantálni, így ugyanannyi biten jóval finomabb felbontást nyerhetünk. Az előbbi eljárást nevezik *adaptív* kvantálásnak, az utóbbit *differenciális* kvantálásnak. A két, alábbiakban bemutatandó eljárás ötvözhető is, így *adaptív-differenciális* kvantálást is végezhetünk.

Az **adaptív kvantálás**nál (lásd a 7.7. ábrát) tehát a kvantálási lépcsők méretét igazítjuk. Ez történhet mintáról mintára, de – ismét a beszéd mintáinak magas korreláltságát is kihasználva – célszerűbb lehet ezt nagyobb időközönként, 10–30 ms-onként végezni (ilyen rövid idő alatt a beszédjel statisztikai jellemzői általában majdnem állandónak tekinthetők, azaz a jel *kvázistacionárius*). A kvantálási lépcsők igazítása jellemzően az előző mintán, esetleg mintákon alapul

$$\Delta x_i(k) \;=\; K \cdot \Delta x_i(k-1), \tag{7.19}$$

azaz a k-adik mintára alkalmazandó lépcsőket az előző, $k-1$-edik mintánál használt lépcsőkből egy K-szorzóval transzformálva kapjuk, ahol K természetesen az előző $k-1$ minta (illetve az előző, figyelembe veendő minták) amplitúdójától függ. Ha az amplitúdó kicsi, akkor $K < 1$, ha nagy, akkor $K > 1$. Mivel K értéke az előzőleg kvantált mintából kiszámítható, ezért a visszaállítás során is rendelkezésre áll, nem szükséges tárolni, átvinni.



7.7. ábra. Az adaptív kvantálás blokkvázlati szintű menete

A **differenciális kvantálás** alapötlete, hogy ha a minták egymáshoz közeli amplitúdójúak, akkor elegendő az amplitúdóváltozást kvantálni, az ugyanis sokkal kisebb (kisebb dinamika tartomány, kisebb szórás), mint maga az amplitúdó, így azonos számú bitre kvantálva jóval finomabb felbontást érhetünk el. A kvantálás menete és alapelve tehát megegyezik a lineáris prediktorral való beszédkódolással (lásd a vonatkozó fejezetet), azaz a megelőző, visszaállítás után előálló minták $(\tilde{x}(k-1), \tilde{x}(k-2), \ldots)$ alapján becslést készítünk (predikció) a következő mintára $(\hat{x}(k))$, például lineáris prediktorral:

$$\hat{x}(k) \;=\; \sum_{i=1}^{N} a_i \tilde{x}(k-i), \tag{7.20}$$

majd a tényleges mintaérték $(x(k))$ és a becslés $(\hat{x}(k))$ különbségét

$$d(k) \;=\; x(k) - \hat{x}(k) \tag{7.21}$$

kvantáljuk. A visszaállított $\tilde{x}(k)$ jelet a becslés és a hibajel összeadásával kapjuk:

$$\tilde{x}(k) \;=\; \hat{x}(k) + d(k) \tag{7.22}$$

A predikcióban célszerű a visszaállítás után előálló minták értékét alapul venni, mert azok a visszaállítást végző egységben is rendelkezésre állnak (míg az eredeti minták nem). A fenti 7.20. összefüggésben a_i-k a lineáris prediktor együtthatói.

7.1.2. Spektrális tulajdonságok meghatározása

A beszédjel-feldolgozás, illetve a beszédjel-analízis alapvető technikája a frekvenciatartománybeli elemzés, hasonlóan más jelfeldolgozási feladatokhoz. Azon paraméterek zömét, amelyekkel a beszédjel jellemezhető, frekvenciaelemzéssel kaphatjuk meg (vannak olyan jellemzők is, amelyek vizsgálata az időtartományhoz kötött: nullátmenetek száma, autokorreláció, jelenergia stb.). A frekvenciaösszetevők (más szóhasználattal spektrális tartomány) vizsgálatához a beszéd amplitudó-

időfüggvényét (hullámformáját) áttranszformáljuk frekvencia-időfüggvénnyé (frekvenciatartomány). A beszéd frekvencia-időfüggvénye megadja, hogy milyen frekvenciakomponensekből áll össze a jel, és hogy ezek hogyan változnak az idő függvényében. Jól illeszthető egymáshoz a beszédhangok egyfajta osztályozása és a spektrális jellemzés három módszere. Az egyszerű szerkezetű zöngés hangok jól jellemezhetők a Fourier-sorral. Az egyszeri lefutású jelek vizsgálatára használatos frekvenciatranszformáció jól alkalmazható a lökéshullám jellegű beszédhangrészek (zárfelpattanás utáni időszakasz) vizsgálatára. Az ergodikus sztochasztikus folyamatok spektrális jellemzésére kidolgozott teljesítménysűrűség-függvény pedig jól illeszkedik a turbulens áramlással létrehozott beszédhangok tulajdonságaihoz. A következőkben röviden áttekintjük a frekvenciatartománybeli elemzésre alkalmazott ezen eljárásokat (Gordos–Takács 1983).

7.1.2.1. Fourier-sor

A Fourier-tétel azt mondja ki, hogy ha egy $x(t)$ jel periodikus (és egy periódusára teljesül a négyzetes integrálhatóság), akkor előállítható $k\frac{2\pi}{T}$ frekvenciájú harmonikusok összegeként, ahol T a periódusidő, k természetes szám, $k = 1, 2, \ldots$. Ebben az esetben a Fourier-sor:

$$x(t) = c_0 + \sum_{k=1}^{\infty} c_k \cos(k \cdot \frac{2\pi t}{T} + \phi_k). \tag{7.23}$$

Mindez szemléletesen azt jelenti, hogy tetszőleges periodikus jel (amely a négyzetes integrálhatóságot teljesíti) előállítható c_k amplitúdójú és ϕ_k fázisszögű szinusz- (koszinusz-) függvények szuperponálásával. Ha az egyes komponensek amplitúdóit (illetve az amplitúdók abszolút értékeit) frekvenciájuk függvényében ábrázoljuk, akkor az amplitúdóspektrumot kapjuk. Hasonlóan, a fázisszögek is ábrázolhatók a komponens frekvenciája mint független változó mentén, ez a fázisspektrum. Ha csak spektrumról beszélünk, akkor általában az amplitúdóspektrumot értjük alatta.

Vegyük észre, hogy a Fourier-sorfejtés az összetett rezgésekből álló jelet komponenseire bontja, ha pedig a jel egyetlen alapharmonikusból és annak felharmonikusaiból áll, akkor a spektrumban az egyes komponensek az alapperiódus reciprokának egész számú eltolásával követik egymást. Az ilyen spektrumot szokták szemléletesen *vonalas spektrum*nak hívni (lásd a 3.3. fejezetben). Az egyes vonalak távolsága megegyezik az alapfrekvencia (F_0) értékével. Az emberi érzékelés valószínűleg ezt a tulajdonságot is felhasználja a beszélő személy hangmagasságának meghatározásakor. Mivel a hangszalagokkal képzett zönge közel periodikus (kváziperiodikus), ezért a zönge spektruma vonalas spektrummal jól közelíthető. A zöngére jellemző vonalas spektrum a zöngés felpattanó zárhangok képzésekor, a fojtott zönge fázisból

származtatható (egyéb esetben a zönge eredeti spektruma az artikulációs csatornában jelentősebben módosul).

Megjegyezzük, hogy a Fourier-sorfejtés feltételezi, hogy teljes pontossággal ismerjük a periódusidőt. Ezt valós beszédjelnél a 7.1.3. fejezetben leírtak szerint véges hibával tudjuk közelíteni.

7.1.2.2. Fourier-transzformáció

A Fourier-transzformáció a Fourier-sorfejtés általánosítása: bebizonyítható, hogy nem periodikus jeleket (sőt, impulzusszerű jeleket) is felbonthatunk szinuszos komposensekre, ekkor azonban a megszámlálhatatlanul sok különböző frekvenciájú komponens végtelenül közel kerül egymáshoz. A spektrum ebben az esetben tehát folytonos. A nem periodikus jelekre általánosított eljárás neve a Fourier-transzformáció, amelyet a Fourier-integrál segítségével végezhetünk el:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt; \tag{7.24}$$

Így kapjuk a komplex spektrumot ($\omega = 2\pi f$). A jel az inverz Fourier-transzformációval nyerhető vissza:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega. \tag{7.25}$$

- A Fourier-transzformáció (7.24. összefüggés) akkor végezhető el, ha az integrál véges ($x(t)$ abszolút integrálható), illetve ha $x(t)$ legfeljebb véges számú szakadást tartalmaz tetszés szerint kijelölt véges intervallumban.
- A Fourier-integrálban a szummázás helyett integrálást végezhetünk, kihasználva, hogy $e^{j2\pi\theta} = \cos(2\pi\theta) + j \cdot \sin(2\pi\theta)$ (Euler-formula). A komplex tartományra való áttérés miatt a „frekvencia” negatív is lehet. Ez természetesen absztrakció, a valóságban értelmezhetetlen. Az $a + jb$ alakú komplex spektrumból az amplitúdóspektrumot az $\sqrt{a^2 + b^2}$, a fázisspektrumot $arctg\frac{b}{a}$ adja meg.
- Az összetett szerkezetű beszédhangok impulzus jellegű hangrészlete folytonos spektrummal közelíthetől jól.

A frekvenciaelemzésnél alapvető fontosságú a rövid idejű elemzés. A beszédjel ugyanis folyamatosan változik, így rövid időintervallumokra (jellemzően 10–30 ms) bontva többnyire teljesül, hogy a beszédjelet közel stacionáriusnak, azaz kvázistacionáriusnak tekinthetjük. Mindez arra vezethető vissza, hogy ilyen rövid idő alatt az emberi beszédképzésben résztvevő szervek konfigurációja sem képes jelentős vál-

tozásra. A folyamatos beszédjel rövid idejű feldolgozását úgy biztosíthatjuk, ha a beszédből rövid részt vágunk ki (ablakolunk), és erre a kivágott mintára végezzük el a frekvenciaelemzést. Ezt az ablakot aztán eltoljuk a beszédjel mentén, majd a következő beszédjelrészletre is elvégezzük a transzformációt. Az ablak eltolásakor érdemes némi átlapolódást biztosítani az egyes transzformálandó beszédjelrészletek között. Ily módon tehát a beszédjel mentén folyamatos, rövid idejű spektrumokat kapunk, ezt nevezik gördülő spektrumnak. A gördülő spektrumot az idő mint független változó mentén megjelenítve kapjuk a hangspektrogramot, mely már 3-dimenziós ábra: a vízszintes tengelyen az idő, a függőlegesen a frekvencia található, a harmadik dimenzió pedig az amplitúdó, amelyet szürkeárnyalatos skálával vagy színkódokkal jelenítenek meg (a korábbi fejezetekben számos ilyen ábra látható).

Mivel a beszédfeldolgozás ma már szinte kizárólag digitalizált (mintavett és kvantált) beszédjelen történik, ezért az alábbiakban a diszkrét Fourier-transzformáció származtatását mutatjuk be röviden. A diszkrét Fourier-transzformáció (DFT) a Fourier-transzformáció digitalizált jelekre alkalmazható közelítése. A digitalizált beszédjel elemzésekor a hullámformát rövid részekre bontjuk az alábbiak szerint. Ha feltételezzük, hogy egy N mintából álló, $0..N-1$-ig indexelt T mintavételi idővel mintavételezett beszédrészletünk van, akkor a fenti integrálba a mintaértékek időben megfelelően eltolt Dirac-impulzussal való szorzatösszegét is beírhatjuk:

$$X(\omega) = \int_{-\infty}^{\infty} \left\{ \sum_{n=0}^{N-1} \delta(t-nT)x(t) \right\} e^{-j\omega t} dt; \tag{7.26}$$

Vegyük azt a gyakorlati esetet, amikor a spektrumot N, egymástól egyenlő $\Delta\omega$ távolságra elhelyezkedő frekvenciára szeretnénk meghatározni a 0 és ω_s közötti intervallumban ($N\Delta\omega = \omega_s$). Ekkor $\omega = k\Delta\omega$, $k = 0,1,2,...,N-1$ esetén:

$$X(\omega) = \sum_{n=0}^{N-1} x[nT]e^{-j\omega nT}, \tag{7.27}$$

azaz

$$X(k\Delta\omega) = \sum_{n=0}^{N-1} x[nT]e^{-jk\Delta\omega nT}. \tag{7.28}$$

$$X(k\Delta\omega) = \sum_{n=0}^{N-1} x[nT]e^{-jkn\frac{2\pi}{N}}. \tag{7.29}$$

Az utolsó lépésben felhasználtuk, hogy $\omega_s = \frac{2\pi}{T}$ és hogy $\omega_s = N\Delta\omega$. Az utóbbi (7.29) összefüggés a diszkrét Fourier-transzformációt (DFT) megadó összefüggés. Az inverz transzformáció alakjához hasonló megfontolásokkal juthatunk:

$$x[nT] = \frac{1}{N} \sum_{k=0}^{N-1} X(k\Delta\omega) e^{jnk\frac{2\pi}{N}}. \tag{7.30}$$

A gyakorlatban a DFT helyett a sokkal gyorsabb, úgynevezett FFT (Fast Fourier Transform) algoritmussal számítják a spektrumot (Bogert et al. 1963). FFT esetén a transzformáció a DFT-hez szükséges N^2 számítási komplexitás $N \cdot log_2 N$ komplexitásúra csökken, ráadásul az FFT nem is közelítő eredményt ad, hanem a DFT-vel azonosat. Mindössze annyi a megkötés, hogy a számítás szempontjából optimális, ha N a 2 valamilyen egész kitevőjű hatványaként adódik (például 512 vagy 1024 pontos FFT).

7.1.2.3. Teljesítménysűrűség-függvény

Bizonyos beszédhangok nem jellemezhetők jól a periodikus, illetve periodikus összetevőt tartalmazó jelekre bevezetett eszközökkel. A zörejes gerjesztésű hangok – például a zöngétlen réshangok – sokkal jobban modellezhetők a valószínűségszámítás eszközeivel, hiszen ezek nagyjából azonos körülmények között ugyan, de véletlenszerűen, végtelen sok akusztikai realizációban megjelenhetnek. Az ilyen típusú beszédhangok jellemezhetők jól a teljesítménysűrűség-, vagy más néven spektrálsűrűség-függvénnyel.

Tekintsük a ξ valószínűségi változóval leírt sztochasztikus folyamat autokorrelációs függvényét (R_ξ), ami a következő alakban írható fel:

$$R_\xi(t_1, t_2) = E(\xi_1, \xi_2), \tag{7.31}$$

ahol E(.) a várható érték képzését jelenti, ξ_i-k a folyamat egyes realizációihoz x_i értékeket rendelnek. Például az [ʃ] hangok ejtésekor a t_i időpontokban a hangokhoz hozzárendelt valószínűségi változók (7.8. ábra). A sztochasztikus folyamatot *stacionáriusnak* nevezzük, ha az autokorrelációs függvény a vizsgálati időpontok távolságától függ, vagyis

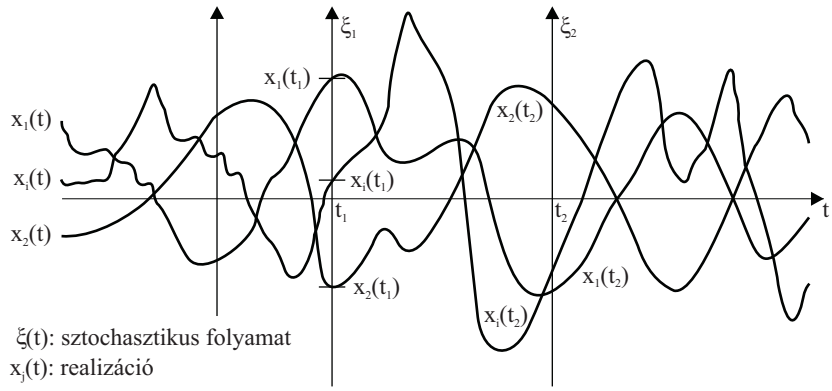
$$R_\xi(t_1, t_2) = R_\xi(\tau), \tag{7.32}$$

ahol $\tau = t_2 - t_1$, valamint a folyamat várható értéke definíció szerint időfüggetlen, azaz

$$E(\xi) = konstans. \tag{7.33}$$

Megjegyzések:

- A stacionaritás további feltétele még, hogy $R_\xi(\tau)$ a $\tau = 0$ pontban folytonos legyen.



7.8. ábra. Sztochasztikus folyamat értelmezése Gordos–Takács (1983) alapján

- Ha a folyamat invariáns az időeltolásra, azaz a folyamatot jellemző eloszlás időinvariáns, akkor *erős stacionaritás*ról beszélhetünk. Ebben az esetben a stacionaritás biztosan teljesül (elégséges feltétel).

Stacionárius folyamatokra jellemző a spektrális sűrűség (spectral density, $s_\xi(\omega)$) függvény. Ez az autokorrelációs függvény Fourier-transzformáltja:

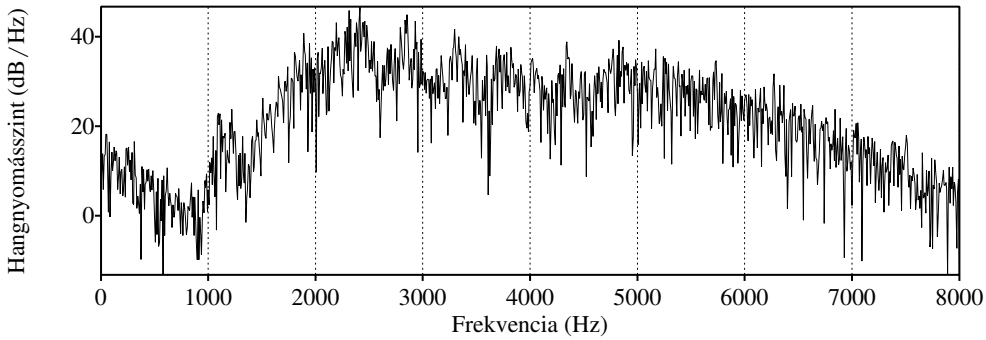
$$s_\xi(\omega) = \mathscr{F}\left\{R_\xi(\tau)\right\} = \int_{-\infty}^{\infty} R_\xi(\tau) e^{-j\omega\tau} d\tau \tag{7.34}$$

$$R_\xi(\tau) = \mathscr{F}^{-1}\left\{s_\xi(\omega)\right\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} s_\xi(\omega) e^{j\omega\tau} d\omega \tag{7.35}$$

Ha a sztochasztikus folyamat nem csak stacionárius, hanem az autokorrelációs függvényre nézve *ergodikus* is, azaz

$$R_\xi(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x_j(t) x_j(t+\tau) dt, \tag{7.36}$$

ahol x_j a ξ folyamat bármilyen realizációja, akkor a spektrálsűrűség-függvény a jel teljesítményének spektrális felbontásaként értelmezhető ($s_\xi(\omega) = |X(\omega)|^2$ értelemben). Például az [ʃ] réshang esetében a teljesítménysűrűség-függvény megadja, hogy egy adott frekvencia környezetébe a jel teljes teljesítményének mekkora hányada esik (7.9. ábra). A függvénynek a példában vett magyar átlagos réshangra a 3000 Hz-es frekvencia környékén van maximuma.



7.9. ábra. Az [ʃ] hang teljesítménysűrűség-függvénye (100 ms-os időtartományra, egyetlen realizációból)

7.1.2.4. Ablakoló függvények

A DFT/FFT esetében láttuk, hogy a teljes beszédjelből csak egy részletet ragadunk ki, amelyet kvázistacionáriusnak tekintünk, és ezen végezzük a transzformációt. Ezt úgynevezett ablakoló függvénnyel való szorzással hajtjuk végre, ennek legegyszerűbb formája a négyszögablak, amely egyenértékű azzal, mint ha a beszédjel valamennyi mintáját nullával tennénk egyenlővé, kivéve azt a részletet, amelyet elemezni kívánunk, ezeket a mintákat ugyanis változatlanul hagyjuk, azaz ha az i és $i+N-1$ közötti N mintát vizsgáljuk, akkor a négyszögablak:

$$E[nT] = \begin{cases} 1 \text{ , ha } i \leq n \leq i+N-1 \\ 0 \text{ egyébként} \end{cases} \tag{7.37}$$

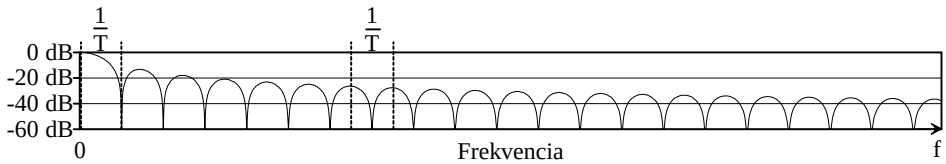
az n index pedig végigfut a teljes, vizsgált beszédjelrészleten, ilyenformán az elemzendő ablakozott jel n-edik mintája az eredeti teljes beszédjelből az ablakfüggvénnyel szorozva adódik: ezt a szorzást tekinthetjük egy ilyen súlyfüggvényű szűrőn való áthaladással egyenértékűnek.

$$\hat{x}[nT] = x[nT]E[nT]. \tag{7.38}$$

Ekkor a négyszögablak átviteli karakterisztikáját a 7.10. ábra mutatja.

$$H_E(\omega) = NT \cdot \frac{sin(\frac{1}{2}\omega NT)}{\frac{1}{2}\omega NT}. \tag{7.39}$$

A diszkrét Fourier-transzfomációt megadó (7.29) összefüggés által szolgáltatott N darab spektrális komponens úgy is felfogható, mint ha azt egy N darab, $\omega_i = \frac{2\pi i}{NT}$ középfrekvenciájú, finom sávszűrőkből álló szűrősor kimenete adná. A (7.38) összefüggés által megadott időtartománybeli szorzás a frekvenciatartományban konvolú-



7.10. ábra. A négyszögablak átviteli karakterisztikája

ciónak felel meg, így lényegében úgy tekinthetjük, hogy a 7.10 ábrán szereplő karakterisztika az egyes sávszűrők karakterisztikáját jelenti, az eredeti origót a frekvenciatengely mentén természetesen a középfrekvenciára eltolva. Ebből pedig azt látjuk, hogy ezek a sávszűrők igen gyenge elnyomást adnak a szomszédos frekvenciákra, azaz egy adott spektrális koponenst nem tudunk kellő szelektivitással vizsgálni, mert a szomszédos összetevők beszivárognak, bekeverednek az éppen vizsgált frekvenciakomponens mellé. Ennek a nem kívánatos jelenségnek a neve a *spektrumszivárgás* (leaking), és éppen erre vezethető vissza, hogy a beszédjel-feldolgozásban a gyakorlatban a négyszögablakot nem alkalmazzák. A négyszögablak helyett tehát olyan ablakfüggvény alkalmazása célszerű, amellyel a spektrumszivárgást a lehető legkisebb szintre tudjuk csökkenteni. Ilyen esetben az elemzendő beszédjelrészletet nem csupán kiragadjuk a beszédjelből (a négyszögablak lényegében ezt jelentette), hanem az egyes mintákat súlyozzuk is, mégpedig úgy, hogy középen kiemeljük, a széleken pedig elsimítjuk. Számos ilyen jellegű ablakfüggvény létezik, ezek közül az egyik leggyakrabban alkalmazott a Hamming-ablak:

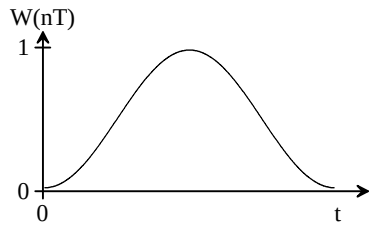
$$W[nT] = \begin{cases} 0,54 - 0,46cos(\frac{2\pi n}{N-1}) & \text{, ha } i \leq n \leq i+N-1 \\ 0 & \text{egyébként} \end{cases} \tag{7.40}$$

A Hamming-ablak időfüggvényét a 7.11. ábra mutatja, frekvenciaválaszát pedig a 7.12. ábra. Utóbbit az előző 7.10. ábrával összehasonlítva látjuk, hogy egyrészt az első zérus pontig tartó frekvenciatartomány kétszer olyan széles, emiatt nagyobb a frekvenciafelbontásban jelentkező úgynevezett *elkenési hatás*. Másrészt a többi frekvencia-összetevőre mintegy 20 dB-lel nagyobb csillapítást kapunk, ezért a Hamming-ablak segítségével az ideális spektrumot jobban közelítő eredményre jutunk. Egy ablakolási művelet eredményeit mutatja be a 7.13. ábra.

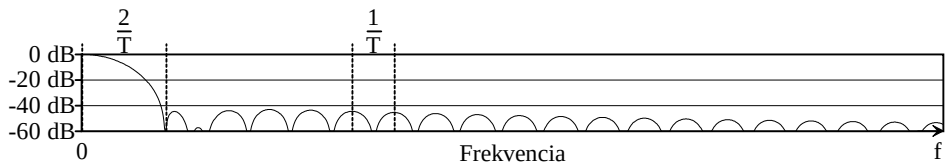
A beszédfeldolgozásban számos más ablakfüggvényt is használnak. Ezekből mutatunk be még hármat.

Hann-ablak (játékosan nevezik Hanning-ablaknak is, az elnevezés nyelvi játékból származik):

$$W[nT] = \begin{cases} 0,5(1 - cos(\frac{2\pi n}{N-1})) & \text{, ha } i \leq n \leq i+N-1 \\ 0 & \text{egyébként} \end{cases} \tag{7.41}$$



7.11. ábra. A Hamming-ablak időtartománybeli képe



7.12. ábra. A Hamming-ablak átviteli karakterisztikája

Ez az ablakfüggvény annyiban tér el a Hamming-ablaktól, hogy a két végpontján nullára csökken.

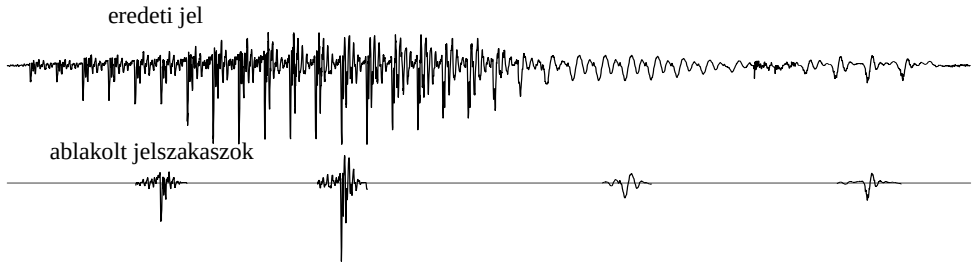
Blackman-ablak:

$$W[nT] = \begin{cases} a_0 - a_1 cos(\frac{2\pi n}{N-1}) + a_2 cos(\frac{4\pi n}{N-1}) & \text{, ha } i \leq n \leq i+N-1 \\ 0 & \text{egyébként} \end{cases} \tag{7.42}$$

ahol általában $a_0 = \frac{1-\alpha}{2}$, $a_1 = \frac{1}{2}$ és $a_2 = \frac{\alpha}{2}$, illetve $\alpha = 0,16$,
Bartlett-ablak:

$$W[nT] = \begin{cases} a_0 - a_1 \left| \frac{n}{N-1} - \frac{1}{2} \right| - a_2 cos(\frac{2\pi n}{N-1}) & \text{, ha } i \leq n \leq i+N-1 \text{ .} \\ 0 & \text{egyébként} \end{cases} \tag{7.43}$$

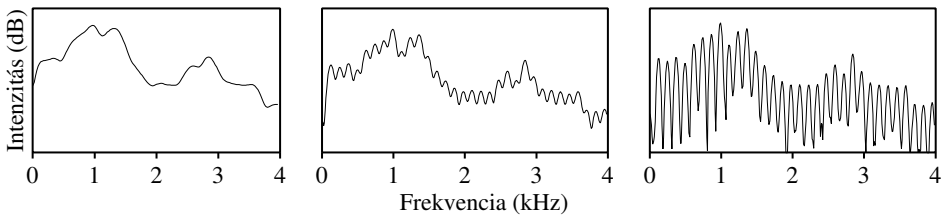
ahol általában $a_0 = 0,62$, $a_1 = 0,48$ és $a_2 = 0,38$. A gyakorlatban az elemzendő beszédjelrészletet kiválasztó és simító ablakot tulajdonképpen végigcsúsztatjuk a teljes beszédjelen, emiatt nevezik csúszóablakos elemzésnek is, jóllehet az ablakot inkább fix időközönként léptetik, semmint csúsztatják. A léptetés tipikus időtartama 10 ms, de mindenképpen olyan, hogy az egyes elemzendő beszédjelrészletek között a kellő (50%-nál nagyobb) átfedés biztosított legyen. Az így kapott spektrum a *gördülő spektrum*.



7.13. ábra. Az *ág* szó (306 ms) időfüggvényéből ablakfüggvénnyel különböző időpontokban kiragadott hullámformarészek. A kétszeres periódusidőnyi hosszúságú (nagyságrendileg 10–20 ms) ablakfüggvényeket a hangperiódusok maximális amplitúdójú részére helyeztük el

7.1.2.5. Idő- és frekvenciabeli felbontás

A beszédjel spektrumának meghatározásához az ablakfüggvény alakja mellett fontos paraméterek még az ablak hossza és a végrehajtott FFT esetében a minták száma (N). A két paraméter nem független egymástól: minél finomabb frekvenciafelbontást szeretnénk elérni, annál nagyobb ablakra van szükségünk, hogy a kellő számú minta a rendelkezésünkre álljon. Ha nagyobb ablakot választunk, akkor az időbeli felbontás lesz gyengébb (nagyobb jelszakaszra számítjuk az FFT-t), sőt akár nem is teljesül már, hogy a beszédjelet az elemzett időablakban kvázistacionáriusnak tekinthetjük. Azaz a frekvencia- és az időbeli felbontás szorzata konstans, valamely paraméter szerinti felbontás növelése csak a másik paraméter szerinti felbontás csökkenése árán lehetséges (vö. 3.3. fejezet). A két ellentmondó követelmény figyelembevételével kell az optimális értékeket beállítani. Az ablak hossza kapcsán említettük már, hogy jellemzően 25 ms körüli értékeket használnak, mivel ekkora szakaszon teljesül, hogy a beszédjel kvázistacionernek tekinthető. Ez azt is jelenti, hogy 8 kHz-en mintavételezett beszédjel esetén 256, 16 kHz-en mintavételezett beszédjel esetén legfeljebb 512-pontos FFT elvégzésének van értelme (7.14. ábra), mivel ez a legközelebbi 2-hatvány az adott időintervallumban elférő minták számát tekintve az adott mintavételi frekvencia mellett. Az N-pontos FFT-vel elérhető elvi frekvenciafelbontásbeli korlát a $(0, \frac{f_s}{2})$ frekvenciaintervallumban $\frac{N}{2}$, mivel valós értékű jelekre – amilyen a beszédjel is – az FFT az $\frac{N}{2}+1$ pontra komplex-konjugált szimmetriát mutat, így $\frac{N}{2}+1$ független amplitúdóértéket kapunk. Ennek megfelelően a frekveciafelbontás elvi felső korlátja mindkét esetben $\Delta f = \frac{f_s}{N} = 31,25\ Hz$. Ez azt jelenti, hogy a megjelenített hangspektrogramon a zöngés hangok felhangjai jól elkülönülve, keskeny horizontális irányú csíkokként látszanak (lásd a 3.24. ábrán korábban). A 7.2. fejezetben az LPC spektrumbecslési módszert ismerhetjük meg.



7.14. ábra. FFT felbontási beállításának hatása a spektrumra. A felbontás növelésével a spektrumkép egyre zajosabbá válik

7.1.3. Zöngés-zöngétlen detekció

A beszéd egyik lényeges jellemzője a zöngésség, illetve a zöngétlenség. A zöngés és zöngétlen szakaszok automatikus detektálása fontos feladat, jóllehet jelentőségéből sokat vesztett a beszédfelismerésben a rejtett Markov-modelles beszédfelismerők megjelenésével, ugyanakkor például a beszédszintézisben és a beszédoktató alkalmazásokban kiemelt szerepe van. A zöngés/zöngétlen osztályozást végző eljárások nagy része egyben a zöngeperiódus, azaz az alapfrekvencia meghatározására is alkalmas, és mint ilyen, jelenleg is különös fontossággal bír. A feladat többnyire már inkább az alapfrekvencia meghatározásaként fogalmazódik meg, lényegében ezt a célt szolgálják az alapfrekvencia-követő (pitch tracker) alkalmazások is. Fontos megjegyeznünk, hogy a feladat látszólagos egyszerűsége ellenére igen nagy kihívást jelent, megbízható osztályozást, illetve alapfrekvencia-követést általában csak olyan kombinált megoldásokkal kaphatunk, amelyek több beszédjellemző együttes alakulása alapján végeznek döntést. Jó összefoglalását adja a témakörnek Hess (1983) munkája. A következőkben röviden áttekintjük a zöngés/zöngétlen különbségtételre és az alapfrekvencia meghatározására alkalmazható automatikus eljárásokat.

A zöngés/zöngétlen különbségtétel egyik alapja lehet, ha megfigyeljük a zöngés és a zöngétlen beszédhangok tipikus tulajdonságait. A rövid beszédszegmensre tekintett úgynevezett gördülőenergia-függvény (Gordos–Takács 1983) – azaz az energia rövid átfogású csúszóablakkal történő számítása – alapján megfigyelhető, hogy a zöngés hangok tipikusan nagyobb, míg a zöngétlenek jellemzően kisebb energiájúak (vö. a specifikus intenzitásokkal a 5.1.1.2. fejezetben). Többé-kevésbé pontos zöngés/zöngétlen detektálás végezhető tehát a gördülő energia alapján, a módszer hátránya, hogy lehetnek (és vannak) a beszédben kisebb energiájú, de zöngés, illetve nagyobb energiájú zöngétlen beszédhangok is, illetve hogy az eljárás rendkívül zajérzékeny.

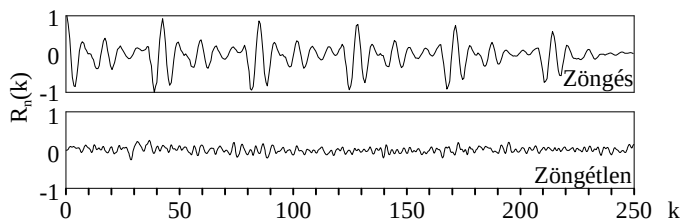
Egy másik lehetőség valamilyen optimálisan választott időegységet tekintve a nullátmenetek száma alapján történő különbségtétel (nevezhetnénk gördülő nullátmenet függvénynek is). E módszer alapja, hogy a zöngés beszédhangok képzésekor a zönge dominál (magánhangzók). Még a vegyes gerjesztéssel képzett beszédhangoknál

is jóval nagyobb energiát képvisel a zöngés komponens, mint például a turbulens áramlás által keltett zörej. Egy zöngés hangperiódus alatt pedig jellemzően csupán néhány nullátmenet történik. Ezzel szemben a tisztán turbulens vagy lökéshullámgerjesztésű zöngétlen mássalhangzók esetében a zajszerű gerjesztés miatt lényegesen több nullátmenet esik egy időegységre.

A kifinomultabb, az alapfrekvencia meghatározására is alkalmas eljárások közül elsőként az autokorrelációs függvény alapján dolgozó módszert említjük, melynek meghatározása a következő:

$$R_n(k) \;=\; \frac{1}{N} \sum_{i=n-N+1}^{n} x(i)x(i-k). \tag{7.44}$$

A beszédjel autokorrelációs függvényét gördülő jelleggel kiszámítva a zönge periódusának megfelelően csúcsok (alapfrekvencia és felharmonikusai) jelennek meg (periódusidőnyi, illetve annak többszörösével végzett eltolás esetén a minták között erős a korreláció), míg zöngétlen beszédszakaszokon a gördülő autokorrelációs függvény gyakorlatilag csúcsmentes. A csúcsok meghatározásával a zöngés beszédszakasz behatárolható, a zönge alapfrekvenciája pedig a periódusidő reciprokát véve kiszámítható (7.15. ábra).



7.15. ábra. Az autokorrelációs függvény zöngés (fent) és zöngétlen (lent) beszédjelekre számolva

A napjainkban széles körben alkalmazott alapfrekvencia-követők az autokorrelációs függvényhez nagyon közel álló átlagos magnitúdó (abszolút érték amplitúdó) különbség függvényt (AMDF, Average Magnitude Difference Function) veszik alapul. N átlagolási intervallumot alapul véve az n-edik beszédmintánál végződő, N mintából álló beszédjel szakaszra a gördülő AMDF-függvény felírható az alábbi alakban:

$$D_n(k) \;=\; \frac{1}{N} \sum_{i=n-N+1}^{n} |x(i)-x(i-k)|, \tag{7.45}$$

ahol $x(i)$ a beszédjel i-edik mintáját jelenti, $i > k$.

A fenti (7.45) meghatározás alapján az AMDF-függvény kváziperiodikus jelekre a periódusidő többszöröseinél minimumot ad, a feladat tehát annak a k értéknek a meghatározása, amelyre az AMDF-függvény értéke minimális. A k mint periódusidő

reciproka adja meg az alapfrekvenciát. Ha nem találunk megfelelően éles minimumot, akkor a beszédrészlet zöngétlennek feltételezhető. Mivel normál beszédben az alapfrekvencia durván az 50–500 Hz tartományba esik, k értéke 2–20 ms között változhat. Emiatt N-t a legnagyobb várható k értéknél valamivel nagyobbra, jellemzően 25 ms nagyságúra választhatjuk. Az AMDF-alapú alapfrekvencia-detektálást széles körben használják, a módszer az egyik legmegbízhatóbb, egyetlen hátránya, hogy esetenként nem az alapperiódust, hanem annak a felét vagy valamelyik felharmonikusát találjuk meg az AMDF-függvény minimumaként, ilyenkor az alapfrekvenciakövető „oktávot ugrik", jellemzően a tényleges alapfrekvencia kétszeresére (vagy a felére) téveszt. Ez ellen a beszélőre jellemző alapfrekvencia-tartomány megadásával (ez k-ra nézve is megszorítást jelent), vagy utólagos szűréssel szoktak védekezni, például a medián szűrő jól kiküszöböli a néhány mintára kiterjedő oktávugrás jellegű hibákat. Ezek a simító szűrők jellemzően az alapfrekvencia-követők szerves részét képezik.

Az alapfrekvencia meghatározása, illetve a zöngés/zöngétlen detekció történhet lineáris prediktor hibajelének alapján is, ez ugyanis zöngétlen beszédszakaszokon zajhoz hasonló, míg zöngés szakaszokon periódusidőnként hirtelen hibanövekedést tapasztalhatunk. A hibajelben a kiemelkedő részeket azonosítva meghatározható az alapfrekvencia.

Lehetőség van az alapfrekvencia meghatározására az úgynevezett harmonikus szorzatspektrum (Harmonic Product Spectrum, HPS) alapján is (Noll 1969). A módszer alapötlete, hogy az alapfrekvenciát a felharmonikusai az alapfrekvencia egész számú többszörösei szerint követik, ennek tehát a spektrumban is „nyoma kell legyen". Tehát egy-egy beszédrészletre (illetve gördülő jelleggel a teljes beszédmintára) meghatározzuk a spektrumot ($X(\omega)$), majd ebből az alábbi ($Y(\omega_i)$) értékeket számítjuk minden alapfrekvenciaként szóba jöhető ω_i-re:

$$Y(\omega_i) = \prod_{n=1}^{N} X(n\omega_i), \tag{7.46}$$

ahol N a figyelembe veendő harmonikusok száma. Az alapfrekvenciát ezután egyszerűen a

$$\hat{Y} = \max_{\omega_i} \{Y(\omega_i)\} \tag{7.47}$$

összefüggésben $\hat{Y}$-t maximalizáló ω_i adja meg.

7.1.4. Jelfeldolgozás prozódiai módosításokhoz

A beszédtechnológiai feldolgozásoknál sokszor van szükség arra, hogy egy adott beszédjel prozódiai jellemzőit megváltoztassuk. Az amplitúdó- és az időtartam-

módosításokat akár az időtartományban is megtehetjük, a nagyobb problémát az alapfrekvencia (dallam) megváltoztatása okozza. Az F_0-t úgy kell megváltoztatni, hogy formánsok, illetve a beszéd magasabb frekvenciájú spektrális komponensei ne torzuljanak (a hangszínezet és hangkarakter változatlan maradjon). A probléma megoldására többféle eljárást is kidolgoztak. Itt kettőt ismertetünk részletesen.

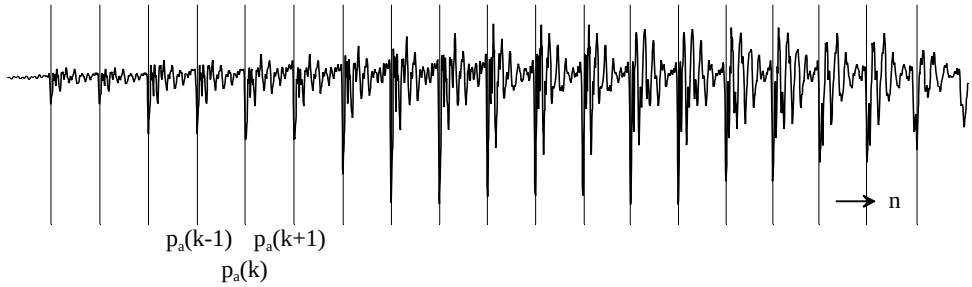
Napjainkban is az egyik legnépszerűbb az úgynevezett zöngeszinkron átlapoló összegző eljárás (Pitch Syncronous OverLap-Add, PSOLA), amelyet a nyolcvanas évek végén publikáltak (Hamon et al. 1989). Az eljárást beszédszintézisben és beszédmanipulációban használják. A jelfeldolgozó algoritmussal a beszéd prozódiai szerkezetét lehet módosítani úgy, hogy a kapott beszédminőség jó marad. Automatizálni is lehet, ami azt jelenti, hogy nincs szükség emberi felügyeletre az algoritmus használatakor. Az eljárás alapelve, hogy a beszéd zöngés periódusait egyenként úgy tekintik, mint egy impulzusra adott válaszfüggvényt. Mivel az alapfrekvencia-változtatás a zöngés periódusokat érinti, a PSOLA-eljárásban a beszédet periódusonként ablakozva „szétszerelik" a kijelölt szakaszon (mondat, szó, hullámformarész). Ezt nevezik analízisnek. Minden zöngés periódusnak az időtengelyen saját címkéje van, ez határozza meg az eredeti beszédjelben a helyét. A módosítás (szintézis) során a zöngés periódusrészeket sorrendben újból összerakják. Ha mindegyik komponens az eredeti időtengelyi címke szerint kerül vissza a helyére, akkor visszakapjuk az eredeti jelet. Ha azonban az időtengelyen kissé eltolva rakjuk őket össze (például sűrűbb időosztással), akkor változni fog az alapfrekvencia anélkül, hogy a beszéd hangszínezete megváltozna. Ennek az eljárásnak mutatjuk meg most a matematikai részleteit.

Előfeldolgozás. A beszédjel szakaszokra bontása úgynevezett zöngemarkerek elhelyezésén alapul (vö. 7.16. ábra). Zöngés szakaszon periódusonként egy markert helyeznek el (a nullátmenetre vagy a periódus maximumhelyére), zöngétlen szakaszokon pedig fix, 10 ms hosszúságú szakaszokra bontják a jelet a markerek segítségével. A továbbiakban $s(n)$ jelöli a mintavett beszédjelet, $p_a(k)$ pedig az analízis során elhelyezett k-adik zöngemarkert. A zöngemarkerek elhelyezése történhet automatikusan is zöngés-zöngétlen detektálás és maximum/nullátmenet-detektálás alapján (lásd az azonos című pontban).

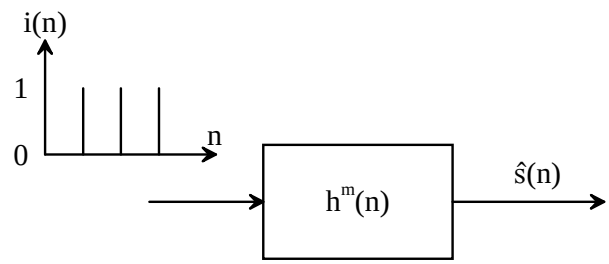
Analízis. Az analízis során a beszédjelet egy időben változó karakterisztikájú véges impulzusválaszú szűrő (FIR szűrő) kimeneteként fogjuk fel (7.17. ábra), amelyből szeretnénk visszanyerni a képzeletbeli szintézisszűrő $i(n)$ bemeneti jelét, azaz:

$$i(n) = \sum_{k=-\infty}^{+\infty} \delta(n - p_a(k)) \tag{7.48}$$

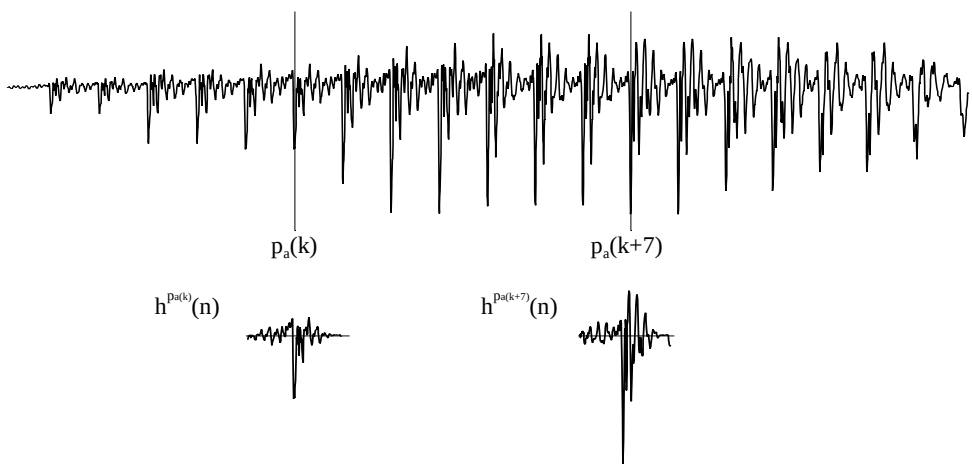
A szűrő impulzusválaszát a $p_a(k)$ markernél jelölje $h^{p_a(k)}(n)$, ezt az $s(n)$ beszédjel egy ablakozott részének tekintjük (lásd 7.18. ábra):



7.16. ábra. A zöngemarkerek elhelyezése a beszédjelben



7.17. ábra. A beszéd szintézismodellje zöngés hangokra a szintézisszűrővel



7.18. ábra. Beszédjel analízise: a k-adik és a k+7-dik zöngemarkerek és a hozzájuk tartozó impulzusválaszok

$$h^{p_a(k)}(n) = s(n) \cdot w^{p_a(k)}(n), \tag{7.49}$$

ahol $w^{p_a(k)}(n)$ az ablakoló függvény, például az alábbi aszimmetrikus alakban:

$$w^{p_a(k)}(n) = \begin{cases} 0 & \text{, ha } n \le p_a(k-1) \\ 0,5 - 0,5 \cdot \cos\left(\frac{\pi(n-p_a(k-1))}{p_a(k)-p_a(k-1)}\right) & \text{, ha } p_a(k-1) < n \le p_a(k) \\ 0,5 - 0,5 \cdot \cos\left(\pi + \frac{\pi(n-p_a(k))}{p_a(k+1)-p_a(k)}\right) & \text{, ha } p_a(k) < n < p_a(k+1) \\ 0 & \text{, ha } n \ge p_a(k+1) \end{cases} \tag{7.50}$$

Az ablakfüggvény átfogása a következő: a kiválasztott markernél van a maximuma, a tőle jobbra és balra következőnél van az átfogási sáv vége. Mivel a beszédben a zöngés periódusok periódusideje periódusról periódusra változhat (a beszéddallam és a hangsúlyozás miatt), az alkalmazott ablakfüggvény nem tekinthető szimmetrikusnak. Az analízis eredménye tehát sok-sok ablakolt zöngés jelszakasz, mindegyik hozzá van rendelve az időtengely adott pontjához.

Módosítás. Az analízis során alkalmazott megkötések miatt a beszédjel előállítása (módosított) a következők szerint zajlik:

$$\hat{s}(n) = \sum_{m=-\infty}^{+\infty} i(m) h^m(n) \tag{7.51}$$

$$= \sum_{m=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \delta(m - p_a(k)) h^m(n) \tag{7.52}$$

$$= \sum_{k=-\infty}^{+\infty} h^{p_a(k)}(n) \tag{7.53}$$

$$= s(n) \cdot \sum_{k=-\infty}^{+\infty} w^{p_a(k)}(n) \tag{7.54}$$

$$= s(n), \tag{7.55}$$

hiszen $h^{p_a(k)}(n)$ a szintézisszűrő impulzusválasza.

Az alapfrekvencia módosítása. Az alapfrekvencia módosításának első lépése, hogy a zöngemarkerek ($p_p(k)$) időtengelyi helyét módosítjuk. Az új pozíciókat a kívánt alapfrekvenciaváltozásból számítjuk ki. Ezzel a szintézisszűrő kívánt bemeneti jelét határozzuk meg célértékek ($\hat{i}(n)$) formájában:

$$\hat{i}(n) = \sum_{k=-\infty}^{+\infty} \delta(m - p_p(k)), \tag{7.56}$$

így a szintézis után nyert beszédjelünk:

$$\hat{s}(n) = \sum_{m=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \delta(m - p_p(k)) h^m(n) \tag{7.57}$$

$$= \sum_{k=-\infty}^{+\infty} h^{p_p(k)}(n). \tag{7.58}$$

Mivel a szintézisszűrő impulzusválaszai a $p_p(k)$ markerek által adott időpontokban nem ismertek, ezért a szűrő $h^{p_p(k)}(n)$ impulzusválaszát a k időponthoz időben legközelebb álló időponthoz tartozó, ismert impulzusválasszal helyettesítjük, természetesen az időkülönbséggel eltolva. (Alkalmazható lenne interpoláció is a két szomszédos, az eredeti jelből meghatározott impluzusválaszra.) Formálisan megadva az eddig elmondottakat: definiáljuk a $C(p_a, p_p(k)) = C^a_{p_p}(k)$ értéket úgy, mint azt az l-et, amely az $|p_a(l) - p_p(k)|$ időbeli távolságot minimalizálja (azaz l a legközelebbi analízis-zöngemarker indexe). Ezzel a szintetizált beszédjel:

$$\hat{s}(n) = \sum_{k=-\infty}^{+\infty} h^{p_a(C^a_{p_p}(k))} \left(n + p_a \cdot \left(C^a_{p_p}(k) - p_p(k) \right) \right). \tag{7.59}$$

Ha közelebb hozzuk egymáshoz az „összeszereléskor” (szintézis) az eredeti időtengelyi zöngemarkerek értékét, akkor növekedni fog az alapfrekvencia, és fordítva. Belátható, hogy nem hozhatók tetszőleges közelségbe a zöngemarkerek, mert egy bizonyos közelség után olyan nagyfokú torzítás fog fellépni, ami eltorzítja az eredeti hangot. Az alapfrekvenciaemelés ebben az eljárásban tehát maximum 20 százalék lehet hallható torzítás nélkül. Alapfrekvencia csökkentésnél nincs ilyen határ, akár 50 Hz-re is csökkenthetjük azt, a hangszínezet megmarad. Ilyen esetben a zöngés periódusok végén csendszakaszok is láthatók a manipulált időfüggvényen. Az intenzitást természetesen korrigálni kell, hogy az átlagos energia állandó maradjon.

Az időtartam módosítása. Az időtartam módosítására két okból lehet szükség. Az egyik, amikor dallammódosítást hajtottunk végre. Ekkor az alapfrekvencia módosulásával megváltozik a beszédhang időtartama, amit korrigálni kell, hogy ne torzítsuk az eredeti beszédjel időszerkezeti képét. Amennyiben az időtartam-módosulás kisebb, mint egy hangperiódusnyi idő, akkor nincs szükség időkorrekcióra. Amennyiben nagyobb, akkor egy-egy periódust megismételnek. Azt, hogy melyiket, az időtengelyi osztás mondja meg. A hangperiódusok ugyanis a szintéziskor fokozatosan csúsznak el az eredeti helyükhöz képest. Ha az elcsúszás nagyobb, mint a soron következő periódus ideje, akkor az ottani aktuális periódust még egyszer meg kell ismételni. Ezzel helyerállítottuk az eredetihez hasonló állapotot. A másik ok lehet, hogy meg kívánjuk változtatni a hangidőtartamokat (lassítjuk,

gyorsítjuk, nyújtjuk a beszédet vagy annak bizonyos szakaszait). Ekkor is a fenti módon járunk el. Az időtartam módosításához egy $n_t = \theta(n_a)$ vetemítőfüggvényt definiálunk, amely az eredeti n_a időindexű mintát n_t időpontba transzformálja. A vetemítőfüggvény inverze: $n_a = \theta^{-1}(n_t)$. A vetemítőfüggvény a kívánt időszerkezetet adja meg. A szintetizált (szintézisszűrő kimenetén) megjelenő beszédjel ekkor:

$$\hat{s}(n) = \sum_{m=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \delta(m - p_t(k)) h^{\theta^{-1}(m)}(n) \tag{7.60}$$

$$= \sum_{k=-\infty}^{+\infty} h^{\theta^{-1}(p_t(k))}(n), \tag{7.61}$$

ahol a szintézis-zöngemarkereket rekurzívan határozhatjuk meg:

$$p_t(k) - p_t(k-1) = p_a\left(C^a_{\theta^{-1}(p_t)}(k-1) + 1\right) - p_a\left(C^a_{\theta^{-1}(p_t)}(k-1)\right). \tag{7.62}$$

Az első, k_0 indexű szintézis-zöngemarkerre általában: $p_t(k_0) = p_a(k_0) = 1$, azaz az első analízis-zöngemarkerhez rögzített. Ezután, akárcsak az alapfrekvencia-módosításánál, meghatározzuk a szintézisszűrő impulzusválaszait:

$$h^{\theta^{-1}(p_t(k))}(n) = h^{p_a(C^a_{\theta^{-1}(p_t)}(k))}\left(n + p_a\left(C^a_{\theta^{-1}(p_t)}(k)\right) - p_t(k)\right), \tag{7.63}$$

ekkor

$$\hat{s}(n) = \sum_{k=-\infty}^{+\infty} h^{p_a(C^a_{\theta^{-1}(p_t)}(k))}\left(n + p_a\left(C^a_{\theta^{-1}(p_t)}(k)\right) - p_t(k)\right) \tag{7.64}$$

Megjegyzések:

- Az alapfrekvencia és az időtartamok egyszerre is változtathatóak, ennek formális leírásától most eltekintünk, hiszen az eddig bemutatott két módszer egyszerű ötvözéséről van szó.
- Zöngétlen szakaszokon az alapfrekvencia módosítása nem értelmezhető, így itt az analízis- és szintézis-zöngemarkerek megegyeznek, a szintetizált beszédszakasz megegyezik a kiindulási beszédszakasszal.
- Az eljárással a hangerő változtatása is lehetséges. Az ablakolt $h^{p_a(k)}(n)$ beszédszakaszokat valamilyen $\alpha(k)$ szorzóval módosítjuk a szintézis előtt.

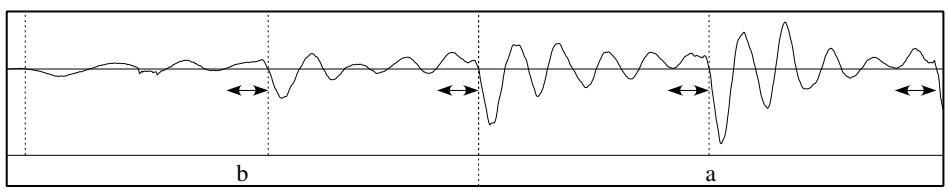
7.1.4.1. Fonetikai alapú prozódiamódosítás

Olaszy Gábor

Egyszerű eljárást dolgoztak ki a magyar hullámforma elemösszefűzéses beszédszintetizátorok prozódiai moduljának támogatására, amely nem használ ablakozást

(Olaszy et al. 2000b). A megoldás azon a fonetikai gondolaton alapul, hogy a hangszalagok záródási periódusának végén már lezajlott a hangképzési periódus lényeges szakasza, ezért a hangperiódus végén végzett időtartománybeli közvetlen beavatkozás nem változtatja meg jelentősen az eredeti hangszínezetet. A beavatkozás a mintavett jel mintáinak a szintjén történik, tehát azokat ritkítjuk, illetve bővítjük (ismétléssel), ezzel változtatjuk a periódus idejét, vagyis az alapfrekvencia módosul, a hangszínezet nem változik.

Előkészítés. Az eljárás alkalmazásának egyetlen kritériuma, hogy a beszédhullám zöngés hangjainak a zöngecímkéit szinkronban kell elhelyezni a hangszalagok nyitódási kezdeti pontjával (például a hangfelvétel készítésekor glottográfos jelet is rögzítünk a hanggal párhuzamosan). Igy automatikusan detektálhatjuk a hangszalag záródási szakaszát. A ProfiVox beszédszintetizátor adattárában így vannak ellátva zöngecímkékkel a hangok (7.19. ábra). Minden hangperiódus címkéjének elhelyezése az időtengelyen fizikailag is megfelel a zönge működésének. Az ilyen címkézés (és csak az ilyen) eredménye az, hogy a fent leírt módosítás nem fogja elrontani a hangszínezetet.



7.19. ábra. A hangszalagok nyitódásával szinkronban elhelyezett zöngemarkerek a hullámformán a [b ɔ] hangkapcsolat néhány periódusán bemutatva. A hangszalag záródási szakaszainál a jelfeldolgozási tartományt a vízszintes nyilak mutatják

F_0 módosítás. Az alapfrekvencia-módosítás algoritmusa a következő: a hangperiódus végén minták szintjén beszúrunk, illetve kiveszünk mintákat, attól függően, hogy mélyíteni, vagy magasítani akarjuk a hangmagasságot.

Határfeltételek: a teljes hangperiódus 75%–100% közötti részén hajtunk végre módosítást. A beszúrás, illetve kivágás legkisebb lépcsője: minden második minta. Tehát két egymás utáni mintát nem vághatunk ki.

Az algoritmus lépései 1. Meghatározzuk az F_0 módosítást a periódus idejére vetítve. 2. Kiszámítjuk hány mintát kell kivenni, illetve betoldani. 3. Kiszámítjuk, hogy minden hányadik mintát érinti a művelet. 4. Elvégezzük a manipulációt. A gyakorlati percepciós mérések azt mutatták, hogy az F_0 mélyíte tágabb határok között végezhető, mint az alapfrekvencia emelése. Így egy 120 Hz alaphangmagasságú férfi hangnál 135 Hz-90 Hz-ig lehetett elhanyagolható torzítással az alapfrekvenciát változtatni. Ez elégséges az összes szükséges F_0 módosításhoz.

Időtartam-módosítás. Az időtartam-módosítás algoritmusa a következő: hangperiódus-betoldást, illetve -kivágást hajtunk végre a beszédhang közepétől számítva a hang szélei felé szimmetrikusan (Olaszy–Olaszi 1998). Így a hangátmenetek formánsmozgásai csak minimális mértékben torzulnak. Határfeltételek: A beszédhang teljes terjedelmére alkalmazható. Két egymás melletti periódust soha nem vágunk ki. Az algoritmus lépései 1. Kiszámítjuk, hogy a kívánt időtartammódosítás hány hangperiódusnyi időt jelent. 2. Kijelölünk annyi periódust a beszédhangon belül, ahányat érint a manipuláció. 3. Elvégezzük a kivágást, illetve betoldást. A gyakorlati tapasztalatok azt mutatják, hogy ezzel az eljárással felére is le lehet lassítani a beszéd sebességét minőségromlás nélkül. A beszédjel nagymértékű felgyorsítása is jó hatásfokkal elvégezhető anélkül, hogy az érthetőség romlana. A vak számítógép-felhasználókat segítő Jaws for Windows képernyőolvasó-programba ezért választották ezt a magyar beszédszintetizátort honosításhoz (lásd a 12.7.1. fejezetben). A fonetikai alapú eljárás előnyei: gyors futást biztosít, egyszerű az algoritmusa, gyakorlati alkalmazásban jól teljesít. Hátrányok: speciális címkézést kíván, viszonylag szűk a módosítási tartománya.

7.1.5. Kepsztrum

Szaszák György

A hagyományos jelfeldolgozásban jól megszokott az idő-, illetve a frekvenciatartománybeli elemzés. Az alábbiakban bemutatunk egy újabb ilyen „tartományt", méghozzá a kepsztrális tartományt. Bár a kepsztrum nem társítható közvetlenül más fizikai jellemzőkhöz (úgy, mint például a spektrum a frekvenciához), nem járunk messze az igazságtól, ha a kepsztrumot mint a spektrummal rokon jelfeldolgozási eszközt képzeljük el. Magát a kepsztrum elnevezést is a spektrum szó első betűinek felcserélésével hozták létre Bogert és Tukey (Bogert et al. 1963). Bár a kepsztrumot eredetileg szeizmikus jelenségek tanulmányozása kapcsán vizsgálták, jól alkalmazható a beszédfeldolgozásban is. Az eredeti cél ugyanis az volt, hogy jól elszeparálhatóvá tegyék egy olyan összetett jel komponenseit, amely egy adott jelből és annak visszhangjából áll:

$$x(t) = s(t) + \alpha s(t-\tau). \tag{7.65}$$

Ennek a jelnek a teljesítményspektruma:

$$|X(\omega)|^2 = |S(\omega)|^2[1+\alpha^2+2\alpha\cos(\omega\tau)], \tag{7.66}$$

melynek logaritmusát képezve az alábbi összefüggésre jutunk:

$$|C(\omega)|^2 \;=\; \log|S(\omega)|^2 + \log[1+\alpha^2+2\alpha\cos(\omega\tau)]. \tag{7.67}$$

Ez utóbbi összefüggést időfüggvényként felfogva (ezúttal f független változóval) az utolsó tagban τ éppen a periódusidő reciprokának megfelelő frekvenciát (alapfrekvenciát) adna meg. Ha tehát (7.67) egy időfüggvény lenne, akkor Fourier-transzformáció után a spektrumban a τ-nak megfelelő frekvencián egy csúcsot látnánk. A kepsztrum számításához éppen ezt végezzük el: ismételt Fourier-transzformációt végzünk, mivel azonban a kiindulási alapunk nem időfüggvény volt, ezért a kepsztrum ábrázolásakor a független változónk nem a frekvencia, hanem valamilyen más dimenziójú mennyiség lesz, amelyet játékosan *kefrenciának* neveztek el:

$$C(k) \;=\; \mathscr{F}\{\log|\mathscr{F}\{x(t)\}|^2\}. \tag{7.68}$$

Az ismételt Fourier-transzformáció ($\mathscr{F}$) helyett inverz Fourier-transzformáció ($\mathscr{F}^{-1}$) is elvégezhető, sőt, ez utóbbit gyakrabban is használják, illetve a teljesítményspektrum analógiájára négyzetre emelés után a kepsztrum alakja:

$$C(k) \;=\; |\mathscr{F}^{-1}\{\log|\mathscr{F}\{x(t)\}|^2\}|^2. \tag{7.69}$$

A kepsztrum előnyös tulajdonsága, hogy az időtartománybeli konvolúció (ez spektrális tartományban szorzással egyenértékű) az összeadással lesz egyenértékű. Márpedig az emberi beszédképzésben a beszédproduktumot a gerjesztőjel (például zönge) és a vokális traktus átviteli függvényének konvolúciója határozza meg, kepsztrumtranszformációval a kettő jól szétválasztható egymástól. Ha tehát ismerjük például a gerjesztőjel kepsztrumát, akkor a beszédproduktum ismeretében a vokális traktus átviteli függvényét is meg tudjuk határozni. A kepsztrumtranszformáció ezen tulajdonságai motiválják a kepsztrum alkalmazását a beszédfeldolgozásban. Ugyancsak az eddig elmondottakból következik, hogy a kepsztrális elemzés jól használható zöngés/zöngétlen detektálásra, illetve az alapfrekvencia meghatározására is, ugyanis a zöngés szakaszokon nagyon jól kiemeli a kváziperiodikus gerjesztő összetevőt az összetett beszédjelből.

7.1.6. MFCC-paraméterek

Az MFC rövidítés a *Mel Frequency Cepstrum* kifejezésnek felel meg, azaz az MFC-együtthatók mel frekvenciás kepsztrális együtthatók (Coefficients). A 3.4. fejezetben már találkoztunk a mel (melodikus hangmagasság) fogalmával. Az MFC-együtthatók kiszámítása egy úgynevezett lényegkiemelési eljárás, amelyet igen széles körben alkalmaznak a beszédtechnológiában. Lényegkiemelés alatt azt értjük, hogy a beszédjelből az információtartalom szempontjából releváns paramétereket

kiemeljük, a többi, lényeges információt nem hordozó vagy redundáns jellemzőt pedig eldobjuk, lényegében tehát beszédtömörítés történik. Az eljárás alapja az emberi hallásmechanizmusra épít. Feltételezzük ugyanis, hogy ha az ember képes megérteni a beszédet, akkor az emberi hallást kellő hűséggel közelítő jelfeldolgozó algoritmusok alkalmazásával nem veszítünk információt (hiszen azt úgysem hallanánk), viszont – amint látni fogjuk – jelentős tömörítést érhetünk el. Az MFC-együtthatók számításának alapja tehát pszichoakusztikai eredetű (lásd a 3.2. fejezetet). Az emberi fül átviteli karakterisztikája ugyanis jól modellezhető egy sávszűrőkből összeállított úgynevezett *szűrősor*ral (más néven *szűrőbank*kal). Az MFC-együtthatókat tehát úgy nyerjük, ha a beszédjelet (helyesebben annak egy rövid, ablakozott szeletét) Fourier-transzformáljuk, majd rajta a fenti szűrősoros elemzést végezzük el, azaz az összetevőket mel sávok szerint összegezzük. Ennek eredményeképpen a 24 sávszűrő kimenetén egy-egy számszerű érték jelenik meg, amely az adott *kritikus sáv*ba eső intenzitás összegzett értéke. Ez azt jelenti, hogy 24 számszerű adattal leírtunk egy rövid beszédszegmenst, a 24 adatot vektorba foglalva kapjuk a jellemzővektort, a vektor elemei pedig az együtthatók. Ha a szűrősoros elemzés után a *kepsztrum-transzformáció*t is elvégezzük, akkor kepsztrális együtthatókat kapunk, ezeket nevezzük MFC-együtthatóknak. Nem nehéz kitalálni, hogy az így kapott együtthatók nem függetlenek egymástól, tehát korreláltak (gondoljunk bele például, hogy egy adott alapfrekvenciájú gerjesztő hang felharmonikusai – felerősödve vagy elnyomva – is megjelennek a beszédben, amelyek már más-más kritikus sávba esnek). Az együtthatók igen hatásosan dekorrelálhatók, ha a kepsztrum számításában alkalmazott Fourier- (vagy inverz Fourier-) transzformáció helyett diszkrét koszinusztranszformációt végzünk, ily módon tehát további tömörítést érhetünk el. A diszkrét koszinusztranszformáció az eredeti együtthatók értékének megfelelő pontokat koszinuszfüggvények szuperpozíciójává transzformálja (tömöríti), például:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right), \tag{7.70}$$

ahol x_n az eredeti, x_k a transzformált együtthatókat jelenti, 24 sávszűrő adatát felhasználva $N = 24$. Az MFC-együtthatókat természetesen gördülő jelleggel érdemes kiszámolnunk, azaz a gördülő spektrumot ablakszélességnyi keretenként konvolváljuk a hallást modellező szűrősor átviteli függvényével, ennek eredményeképpen bizonyos fix időközönként (például a gördülő spektrum számításakor 10 ms-os ablakeltolást végezve) egy-egy 24 dimenziós jellemzővektort nyerünk. Ha a beszédjelet 16 kHz-en mintavételezzük – ez a gyakorlatban elégnek is bizonyul –, az utolsó 4 kritikus sávot figyelmen kívül hagyhatjuk, mivel azok már a mintavételezés megkövetelte 8 kHz határfrekvenciájú sávkorlátozás miatt eleve kiesnek a feldolgozott tartományból. Így 20 szűrőkimenetünk lesz. Ebből a diszkrét-koszinusztranszformáció

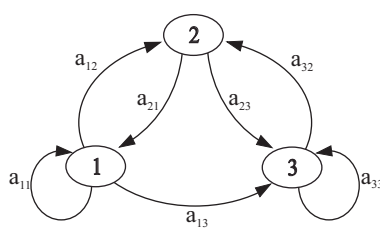
révén 12 dimenziós jellemzővektorokat kapunk, e vektorok számelemei az MFC-együtthatók.

7.1.7. Rejtett Markov-modellek

A Markov-folyamat egy sztochasztikus (véletlen) folyamat, amelyet állapotok egy halmazával, illetve az ezek közötti átmenetekkel, az átmenetek valószínűségével jellemezhetünk. A folyamat ezen modellje a *Markov-lánc*, amelyet Markov-modellnek is neveznek. Egy sztochasztikus folyamat akkor Markov-folyamat, ha teljesül rá a Markov-tulajdonság, azaz a folyamat múltbeli állapotainak megfigyelése alatt nyert információt a folyamat jelen állapota magába sűríti, azt tartalmazza, más szavakkal a folyamat jövője a múlttól nem, csak a jelen állapottól függ. Ezt nevezik *memóriamentes*ségnek is. Feltéve, hogy a folyamat a megfigyelés kezdete óta az $X_1, X_2, ..., X_n$ állapotokon haladt keresztül, a Markov-tulajdonság tehát:

$$P(X_{n+1}|X_1, X_2, ..., X_n) \;=\; P(X_{n+1}|X_n), \tag{7.71}$$

azaz a múlt ismerete semmiféle többletinformációt nem jelent, a folyamat következő állapota (X_{n+1}) csakis a jelentől (X_n) függ. Arra, hogy a folyamat melyik állapotban van, megfigyelés (obsevation, jele:o) alapján következtethetünk, a megfigyelés lehet közvetlen („látjuk", hogy melyik állapotban van), vagy közvetett (valamilyen realizáció alapján következtetünk az állapotra). A Markov-lánc egyes állapotait q_i-vel, az i állapotból a j állapotba történő átmenet valószínűségét $a_{i,j}$-vel szokás jelölni, tehát $a_{i,j} = P(X_{n+1} = q_j|X_n = q_i)$. Grafikus reprezentációnál az állapotokat körökkel vagy ellipszisekkel jelöljük, a lehetséges állapotátmeneteket pedig (ahol $a_{i,j} > 0$) nyíllal. Diszkrét idejű Markov-folyamat esetén a folyamat csak meghatáro-



7.20. ábra. Háromállapotú Markov-modell

zott időpontban válthat állapotot, vagy helyben is maradhat, de a diszkrét időpontok között egy állapotban tartózkodik. A diszkrét idejű Markov-láncokat (is) gyakran használják valamilyen fizikai folyamat modellezésére, ahol különböző megfigyelések alapján kell szimulálni-modellezni a folyamatot. Ha a megfigyelés egyértelműen

azonosítja, hogy a folyamat milyen állapotban van, akkor a használt modellt megfigyelhető Markov-modellnek vagy egyszerűen Markov-láncnak nevezzük. Ilyen eset például a kockadobálás, hiszen összesen 6 állapot lehetséges, a dobás végrehajtása után végzett megfigyelés pedig egyértelműen azonosítja az állapotot (tulajdonképpen az állapot és a megfigyelés nem is válik szét egymástól).

Számos olyan folyamat létezik viszont, melyekre ugyan az állapotok jól definiálhatók, rájuk a megfigyelések alapján mégsem következtethetünk egyértelműen. Vegyük a következő egyszerű példát: Van három dobozunk, mindegyikben 6 db labda. Az elsőben 3 fehér, 2 piros, 1 fekete, a másodikban 3 piros, 2 fehér, 1 fekete, a harmadikban 4 fekete, 1 piros, 1 fehér. Legyen a megfigyelésünk (jelöljük o-val) az, hogy egy dobozból kihúzunk egy labdát és megnézzük, milyen színű. A megfigyelés alapján el szeretnénk dönteni, melyik állapotban vagyunk, vagyis, hogy mi a sorszáma a doboznak, amelyikből húztunk. Világos, hogy bár az első dobozban főleg fehérek vannak, a másodikban főleg pirosak, harmadikban meg feketék, a megfigyelés nem determinálja az állapotot. Csak valószínűségeket tudunk mondani – előzetes ismereteink birtokában. Legyen például a gondolatkísérletben a megfigyelésünk a piros labda. A következő megállapításokat tehetjük: feltéve, hogy az 1. dobozból húztuk, az esemény valószínűsége 1/3. Feltéve, hogy az 2. dobozból húztuk, a valószínűség 1/2, míg ha azt tesszük fel, hogy a 3. dobozból húztuk, akkor 1/6. Ha tudjuk, hogy piros labdát húztunk, akkor tudjuk, hogy legnagyobb eséllyel a 2. dobozból húzhattunk piros labdát, de nem lehetünk biztosak benne, hogy valójában a 2. dobozból került a kezünkbe a piros, hiszen a másik kettőben is van. Az ilyen és ehhez hasonló folyamatok modellezésére alkalmasak a rejtett Markov-modellek. A rejtett szó arra utal, hogy a megfigyelő nem látja az állapotokat, azok mintegy rejtve maradnak – esetünkben sosem tudjuk meg biztosan, melyik dobozból húztuk a piros labdát. A rejtett Markov-modellben minden állapothoz tartozik egy diszkrét vagy folytonos eloszlásfüggvény, ami azt mutatja meg, hogy az állapotban értelmezett egyes (diszkrét vagy folytonos) megfigyeléseknek mi a valószínűsége. A j állapotban az o valószínűségi változó megfigyelésének valószínűségét – vagy egy azzal arányos mennyiséget – $b_j(o)$-val jelöljük. A fenti példán keresztül szemléltetve az első doboz (állapot) eseményeinek (pirosat, fehéret vagy feketét húzunk) valószínűségeloszlása: $b_1(piros) = 1/3$; $b_1(fehér) = 1/2$; $b_1(fekete) = 1/6$. Hasonlóan, a többi állapot eloszlása is könnyen számolható, ezt az olvasóra bízzuk. Természetesen a megfigyelésnek nem kell diszkrétnek lennie. Lehet például valós szám N-es – azaz vektor –, ilyenkor $b_j(\mathbf{o})$ egy vektor-skalár függvény, egy N-dimenziós eloszlás sűrűségfüggvénye. Lényegében ugyanarról van szó mint korábban, a megfigyeléshez egy valószínűségi mértéket rendelünk. Az utóbbi típusú modelleket folytonos megfigyelési-sűrűségfüggvényű rejtett Markov-modelleknek hívjuk, ilyeneket használnak a korszerű beszédfelismerőkben is. A Markov-modellekkel a későbbiekben részletesen foglalkozunk a gépi beszédfelismerés tárgyalásánál.

7.2. A beszéd tömörítése és átvitele

Tatai Péter

A kis sebességű beszédkódolás, más néven beszédtömörítés, a jelfeldolgozás igen fejlett ága, amelynek sokféle alkalmazása széles körben elterjedt. Ilyenek a mobiltelefon, az internetes (VoIP: Voice over Internet Protocol) telefon, a hangrögzítők, a hangposta stb. Az utóbbi 1-2 évtizedben a mobil- és VoIP-telefonálás rohamos terjedése rendkívül felgyorsította a beszédkódolás technikájának a fejlődését. Az átviteli sebesség csökkentése ezeken a területeken különösen fontos, mert ez korlátozza az egyidejű beszélgetések számát, miközben a kisebb sebesség mellett a jó beszédminőséget is a lehetőség szerint meg kell őrizni, ami egyre komplexebb algoritmusokat igényel.

A módszerek, elvek és algoritmusváltozatok rendkívül nagy száma miatt csak néhány sikeres és a gyakorlatban is alkalmazott kódolási eljárást mutatunk be, és mindenütt az elv megértésére helyezzük a hangsúlyt. Az érdeklődő olvasó a megvalósítási részletekről és ötletekről, valamint a szabványos kódolókról tengernyi cikket és számos könyvet is találhat (Chu 2003, Kondoz 2004, Vary–Martin 2006, Hanzo et al. 2001). Jelen fejezet célja ezért a téma áttekintése mellett az irodalom megértésének és az abban való tájékozódásnak a megkönnyítése.

A tömörítő beszédkódolóktól elvárt főbb jellemzők:

- kis bitsebesség (alkalmazástól függően a 2–32 kbit/s sávban)
- jó beszédminőség (jó érthetőség és lehetőleg kellemes hangzás)
- érzéketlenség mind a beszélő személye, mind a használt nyelv iránt
- kis kódolási késleltetés (2–30 ms) a telefonos alkalmazásoknál
- robusztusság a csatorna bithibáival és a csomagvesztéssel szemben
- skálázhatóság, azaz igény szerint változtatható átviteli bitsebesség
- elfogadható átvitel a nem beszéd jelekre (főleg a DTMF-jelekre, amelyeket a tone üzemmódú telefonokban használnak)
- kis bonyolultság, azaz kis memória- és számításikapacitás-igény

Tekintettel arra, hogy egyetlen kódolási módszerrel sem lehet minden fenti igényt kielégíteni, illetve a különböző alkalmazások igényei eltérőek a minőség, a hibatűrés, a bitsebesség, a bonyolultság, a késleltetés stb. terén, ezért sokféle kódoló típus alakult ki, sőt egy-egy adott alkalmazás esetén sem egyszerű az optimális megoldás kiválasztása. A típusokat sokféleképpen osztályozhatjuk. Bitsebesség alapján az alábbi osztályozást tehetjük

- kis sebességű (2–5 kbit/s)
- közepes sebességű (5–15 kbit/s) és
- viszonylag nagy sebességű ($>$15 kbit/s)

7.2.1. Kódolási alapelvek

A kódolási technikákat három csoportba szokták sorolni:

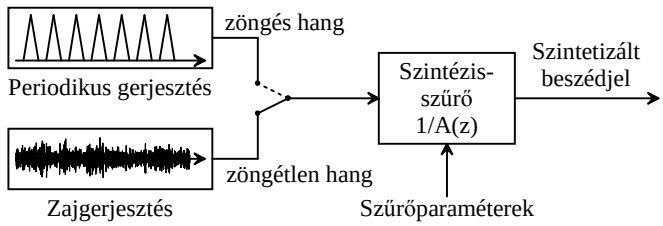
- Hullámforma-kódolás, amely bármely sávkorlátozott jelre alkalmas, és az alakhűség megtartására törekszik, a jel redundanciájának csökkentésével tömörít.
- Parametrikus vagy forráskódolás, amely a forrásmodellnek, beszédjel esetén a beszédkeltési modellnek a paramétereit viszi át, emiatt másfajta jelekre nem alkalmazható.
- Hibrid kódolás, amely mind a csökkentett redundanciájú jelet, mind a forrásmodell paramétereit átviszi valamilyen formában, így ötvözi a fenti típusok előnyeit.

7.2.1.1. A hullámforma-kódolás

Ebbe a kategóriába tartozik a PCM (Pulse Code Modulation), valamint a DPCM (Differential PCM) és az ADPCM (Adaptive DPCM). Ezek minőségét elsősorban a kvantálási jel-torzítás viszonnyal jellemezzük, amely 30–38 dB között van 32–64 kbit/s átviteli sebesség mellett. Ez jó telefonminőséget jelent, akár többszöri átkódolás (analóg-digitál átalakítás, kódolás, digitál-analóg átalakítás) után is. A digitális telefonhálózatban a 64 kbit/s sebességű PCM az általánosan alkalmazott beszédkódolás (ez pontosan 64 000 bit/s, nem 64x1024). A 16, 24 és 32 kbit/s-os, fix vagy igény szerint változtatható sebességű ADPCM-kódolás egyes átviteli és hangrögzítő rendszerekben terjedt el. Mind a PCM, mind az ADPCM elterjedését alapvetően a digitális áramköri technika aktuális fejlettsége tette lehetővé. A PCM esetében ez az 1960-as évek digitális technikájának, az ADPCM esetén pedig a korai digitális jelfeldolgozó processzoroknak (DSP: Digital Signal Processor) volt köszönhető. Hasonlóképpen a komplexebb jeltömörítési eljárások fejlődése is szorosan kötődik a DSP áramköri technika, valamint a jel- és beszédfeldolgozási algoritmusok fejlődéséhez.

7.2.1.2. Parametrikus kódolás

Ennél az eljárásnál nem a jel hullámformáját visszük át, hanem a jel úgynevezett forrásmodelljének a paramétereit. A beszéd esetén a legegyszerűbb forrásmodell egy zajjal, illetve periodikus impulzussorozattal gerjesztett időben változó szűrő, ahol a gerjesztés a tüdőből kiáramló levegő és a hangszálak működését, a szűrő pedig az artikulációs csatornát modellezi. Amennyiben tehát nem tetszőleges jelek tömörítő kódolása a cél, hanem elegendő a beszédjelre koncentrálni, akkor a fenti modellezésen alapuló megoldás 2–5 kbit/s-mal is jó érthetőséget nyújt, ha csak a zönge periódusidejét, a zöngésség/zöngétlenség információt, az erősítést, valamint a szin-



7.21. ábra. Egyszerűsített beszédkeltési modell

tézisszűrő paramétereit visszük át. A beszédkódolásra szolgáló parametrikus kódolót hagyományosan Vokódernek nevezik. Mivel a modellezés a beszédjelre van szabva, eltérő tulajdonságú jelekre a Vokóder gyakorlatilag nem alkalmas, és ezen a bitsebesség növelése sem segít, mert a forrásmodell korlátai miatt a hullámforma ettől alig javul, ellentétben a hullámforma-kódolókkal, amelyek a bitsebesség növelésével egyre jobban közelítik a kódolandó jelet.

7.2.1.3. Hibrid kódolás

Ez a szokásos elnevezése a komplexebb kódolási módszereknek, amelyek a hullámforma és a parametrikus kódolás előnyeit egyesítik. Ily módon, akár igény szerint változtatható (skálázható) bitsebesség is megvalósítható, ami optimálisan illeszthető egy-egy adott alkalmazáshoz a rendelkezésre álló sávszélesség és a beszédminőség közötti kompromisszumnak megfelelően. Igény szerint ezekkel elfogadható hullámforma-reprodukció is biztosítható, például a DTMF-telefonjelzések átvitelére.

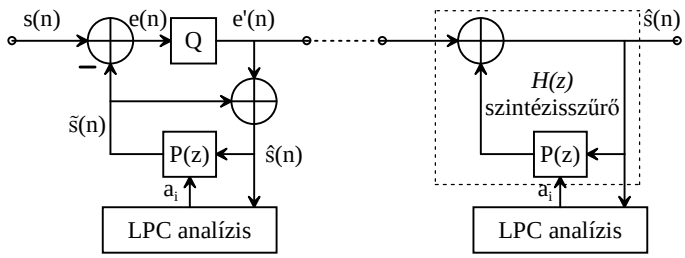
A hibrid kódolás esetén mind a gerjesztőjelet, mind a szintézisszűrő paramétereit kvantáljuk, kódoljuk és átvisszük, illetőleg felhasználjuk a dekódolás során. Tekintettel arra, hogy néhány kivételtől eltekintve a tömörítő beszédkódolók a hibrid kategóriába tartoznak, ezért a logikus áttekintés kedvéért a következőkben nem a fentiek szerint, hanem a kódolási struktúra alapján osztályozzuk a módszereket. Ennek alapján megkülönböztetünk három fő csoportot:

- Adaptív differenciális kódoló (ADPCM) predikciós struktúra
- Nyílt hurkú, AaS (Analysis and Synthesis) predikciós kódoló struktúra
- Zárt hurkú, AbS (Analysis by Synthesis) predikciós kódoló struktúra

Itt jelezzük, hogy a következőkben ismertetett megoldásokban nem foglalkozunk a jel erősségével. A gyakorlati megvalósítások során a legtöbb struktúrához hozzátartozik egy skálatényező (scale factor, gain) paraméterbecslése is, amelyet – a visszacsatolt struktúrákat kivéve – szintén át kell vinni a vételi oldalra.

7.2.2. Adaptív differenciális, predikciós kódoló

Az előrejelző módszerek (predikció) a jelek feldolgozására kifejlesztett matematikai eljárások klasszikus eszközei. A predikció lényege, hogy a jelenség valamely időpillanatra vonatkozó értéke megbecsülhető a korábbi időpillanatokban tapasztalt értékekből. Lineáris a predikció, ha a becslés a becsléshez felhasznált korábbi értékek lineáris függvénye. A lineáris predikció alkalmazása a beszédfeldolgozásban azt eredményezi, hogy kevesebb paraméterrel tudják jellemezni (átvinni, közvetíteni) a beszédet a hangminőség lényegi romlása nélkül. Amennyiben ezt a kódolást végrehajtjuk, az átviteli csatorna másik végén hasonló elvű dekódolást kell biztosítani. Ennek a teljes rendszernek a blokkjait mutatja a 7.22. ábra (adaptív differenciális PCM-kódolás). A bemeneti *s(n)* beszédjelet az előző becsült minták alapján az $\tilde{s}(n)$



7.22. ábra. ADPCM kódoló és dekódoló

jellel közelítjük, amelyet a *P(z)* lineáris prediktor állít elő

$$\tilde{s}(n) = \sum_{i=1}^{p} a_i \hat{s}(n-i), \tag{7.72}$$

ahol p a predikció fokszáma. A lineáris prediktor tehát lényegében egy véges impulzus válaszú (FIR: Finite Impulse Response) szűrő, amely a bemenetére jutó minták lineáris kombinációját képezi, és z-transzformáltja ennek megfelelően

$$P(z) = \sum_{i=1}^{p} a_i z^{-i}. \tag{7.73}$$

A prediktor bemenetére az előző becsült $\tilde{s}(n)$ mintának és az előző predikciós hiba kvantált $e'(n)$ értékének összege kerül, amely azonos a dekódoló kimeneti jelével, ha az átvitel hibátlan:

$$\hat{s}(n) = \tilde{s}(n) + e'(n). \tag{7.74}$$

Tekintettel arra, hogy a beszéd és az egyéb információs jelek mintái nem függetlenek, a prediktor a_i lineáris predikciós paramétereinek (LPC: Linear Predictive Coeffici-

ents) helyes megválasztásával elérhető, hogy a differenciális jel, vagyis a predikciós hibajel, *e(n)*, energiája kisebb legyen, mint a bemeneti jelé. Mivel a jelek statisztikai jellemzői változnak, a DPCM továbbfejlesztésének tekinthető ADPCM esetén a paramétereket is folyamatosan becsüljük és a jelhez igazítjuk. (A paraméterek becslésére, az analízisre a következő fejezetben visszatérünk.) Ily módon a Q kvantálóba jutó differenciális jel dinamika tartománya csökken, amivel a kvantálási hiba is csökken, ha finomabb kvantálást alkalmazunk, vagy kevesebb bittel érhető el azonos torzítás. Az adaptivitást érdemes kiterjeszteni a kvantálóra is, vagyis a kvantálási szinteket a jelszinttel együtt változtatva jelentősen tovább csökkenthető a hiba (ezt nem ábrázoltuk a blokksémán). Ha a 7.21. ábra beszédkeltési modelljét összevetjük a 7.22. ábra szerinti ADPCM struktúrával, akkor látható, hogy a dekódoló oldal egy hibajellel gerjesztett szintézisszűrőnek tekinthető, amelynek transzferfüggvénye

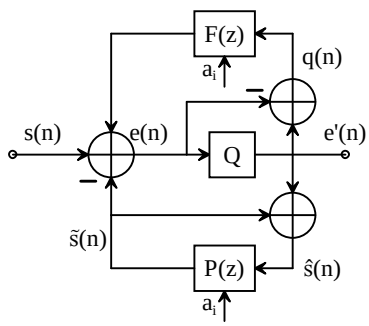
$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - P(z)} = \frac{1}{\sum_{i=1}^{p} a_i z^{-i}}. \tag{7.75}$$

Kísérleti eredmények szerint a beszédjel spektruma, azaz az artikulációs csatorna frekvenciafüggése aránylag jól modellezhető egy ilyen, csak pólusokat tartalmazó (All-Pole) szintézisszűrő karakterisztikával. A tényleges beszédjel keltésének pontosabb fizikai modellje természetesen ennél lényegesen bonyolultabb, akár a gerjesztést tekintjük, akár a frekvenciamenetet. Például a nazális hangoknál jellegzetes átviteli zérusok is vannak, azonban a fenti egyszerűsített szintézisszűrő-modell nemcsak jó gyakorlati közelítés, hanem a lineáris predikció (LP) elmélete révén matematikailag és számítógépes algoritmusokkal is igen hatékonyan kezelhető, ezért széles körben alkalmazzák. Szinte minden kódolási algoritmus bizonyos mértékig erre épül, az LPC (Linear Predictive Coding) számtalan változatával találkozhatunk. Nem szokott félreértést okozni, de az LPC rövidítést értelemszerűen egyaránt használják a kódolásra is és az együtthatókra is (LP Coefficients).

Megjegyezzük, hogy a beszéd formánsalapú szintézisénél és a beszédfelismerés akusztikus paramétereinek előállításánál is kulcsszerepe van a lineáris predikciós együtthatóknak, és ezek meghatározását nevezzük LPC-analízisnek. A beszédre az 1970-es években kezdték alkalmazni a predikciós módszert (Markel–Gray 1976). Az eljárás előnyei a következők. Nem kell a prediktor a_i paramétereit feltétlenül átvinni a dekódoló oldalra, mert ugyanezek az adatok ott is előállíthatóak a dekódolt jelből, továbbá, hogy a kódolás nem okoz számottevő késleltetést, szemben a későbbiekben tárgyalt struktúrákkal, amelyeknél egy eltárolt jelszakaszra történik az optimalizálás, ezáltal jelentős bufferelési késleltetés lép fel, tipikusan 10–30 ms, viszont a bufferelléssel az analízis sokkal hatékonyabbá válik. (Utóbbi esetben már csak formálisan beszélünk predikcióról, jóslásról, hiszen egy *ismert* jelszakasz spektrális analízisét végezzük el, csak a matematikai módszer azonos a predikcióval.) Az eljárás hátrá-

nya, hogy az alkalmazott úgynevezett visszairányú (backward) adaptáció nem hatékony. Egyrészt azért, mert a predikció, valamint az analízis nem az eredeti jelet használja, másrészt azért, mert az analízis nem az aktuális jelszakaszra, hanem a jel korábbi szakaszára vonatkozik.

Az ADPCM kvantálása során keletkező kvantálási hiba *teljes egészében*, változatlan spektrummal megjelenik a kimeneten. Feltételezve, hogy az átvitel során nem lépnek fel bithibák, a dekódoló oldalra átvitt hibajelből visszaállított $\hat{s}(n)$ beszédjel csak az $e(n)$ hibajel kvantálásából származó $q(n) = e'(n) - e(n)$ hibával tér el az eredeti jeltől. Ennek spektruma (lásd a Kvantálás c. fejezetet) közel egyenletes, ezért fehérzajjal is szokás közelíteni. A beszédminőség szempontjából ez hátrányos, mert amíg a spektrális csúcsoknál a beszédjel részben vagy egészben maszkolja a fehérzaj jellegű kvantálási torzítást, addig a spektrális völgyekben ez jól hallható maradhat. Ezt a kellemetlen hatást jelentősen lehet csökkenteni zajformálással (7.23. ábra). Az



7.23. ábra. Zajformáló ADPCM kódoló

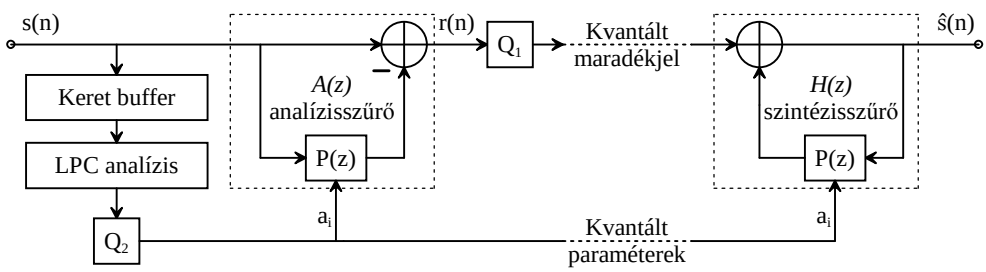
$F(z)$ zajformáló a $q(n)$ kvantálási hibajelet a beszédjel spektrumának megfelelően formálja (a formálás mértékét egy paraméter jelzi, amelyre később visszatérünk). Ezáltal a teljes torzítási teljesítményből a beszédspektrum csúcsaihoz nagyobb, a völgyeihez pedig kisebb torzítási teljesítmény jut. Ez fontos, ugyanis a beszédkódolásnál a füllel hallható hibajelet kell minimalizálni, mert ***nem az számít, hogy mekkora a reprodukálás hibája, hanem az, hogy ebből mennyit hallunk***, vagyis az a cél, hogy a hasznos jel minél jobban elfedje, maszkolja az eljárás elkerülhetetlen hibáit. A zajformálás a korszerű kódolóknál nagyon fontos és általánosan használatos, ezért erre a későbbi struktúráknál visszatérünk.

Most csak érdekességképpen említjük meg, hogy a differenciális (DPCM) kódolást egybites változatban, deltamoduláció (DM) néven elterjedten alkalmazták régebbi, főleg katonai telefonrendszerekben, ahol a kvantálási lépcső adaptív változtatásával, igen egyszerű áramkörökkel, aránylag hatékony, bár nem túl jó minőségű beszédátvitel valósítható meg. A DM kisebb módosításával jutunk a delta-szigma modulátorhoz (ez lényegében egy feszültség-frekvencia konverter), amely a min-

tavételi frekvencia jelentős növelésével és igen erős zajformálással nagyon pontos analóg-digitál átalakítást tesz lehetővé. Noha a mintavételi frekvencia minden négyszerezése csak 1-1 bit nyereséget jelent önmagában, zajformálással a hasznos sáv fölé tolhatjuk a kvantálási torzítás hibajelének spektrumát, ahonnan digitális aluláteresztőkkel ez majdnem teljesen eltávolítható, és extrém pontos alkatrészek nélkül az egybites átalakítóból 16–24 bites analóg-digitál és digitál-analóg átalakítók származtathatók, amelyek jól integrálhatók és hatékonyan gyárthatók, így ez a módszer szinte egyeduralkodóvá vált a hangtechnikai berendezésekben.

7.2.3. Nyílt hurkú predikciós kódoló

A nyílt hurkú (AaS) kódolót a 7.24. ábra mutatja. Ennél a struktúránál is szintézisszűrőnek felel meg a dekódoló, amely teljesen megegyezik az ADPCM struktúra dekódolójával. A különbséget a kódoló oldalon látjuk, ahol előre irányuló predikciót, úgynevezett analízisszűrőt alkalmazunk, és a kvantáló (Q_1) kikerült a hurokból. Itt szintén a beszédjel előző p mintájának lineáris kombinációjával becsüljük az $s(n)$



7.24. ábra. A nyilt hurkú (AaS) kódolás elve

mintát. A becslés természetesen nem pontos, hiszen akkor az *n.* minta nem is hordozna új információt, ezért egy *r(n)*, úgynevezett maradékjel (residual signal) marad, mint a predikció hibajele:

$$r(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i). \tag{7.76}$$

Az ábrán *A(z)*-vel jelöltük az

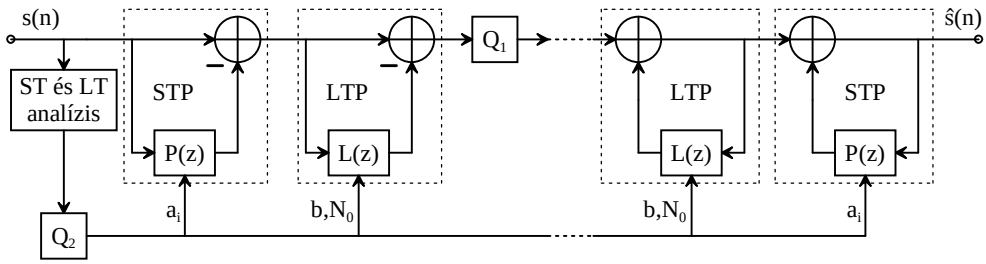
$$A(z) = 1 - P(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{7.77}$$

blokkot, amit analízisszűrőnek nevezünk, mert célja a jel analízise és a jelben lévő redundancia eltávolítása. Ha képesek lennénk az összes redundancia eltávolítására, akkor a maradékjel fehérzaj lenne, azaz az átviteli sávban minden frekvencián ugyanolyan intenzitású frekvenciakomponense lenne, illetve zöngés jelek esetén periodikus impulzussorozat.

Most újra utalunk a 7.21. ábra szerinti forrásmodellre (7.2. fejezet). Ha az ott szereplő szintézisszűrő pontos inverze lenne az analízisszűrő, akkor ennek kimeneteként valóban fehérzaj, illetve periodikus impulzussorozat állna elő. Azonban az analízis, vagyis a redundancia eltávolítása nem tökéletes, ezért a maradékjel nem lesz teljesen korrelálatlan minták sorozata, tehát a jó beszédminőség érdekében célszerű azt is átvinni, és a dekódoló oldalon ezzel a maradékjellel gerjeszteni a szintézisszűrőt. Ezt nevezik nyílt hurkú, illetve AaS (Analysis and Synthesis) módszernek is, utalva a kódoló oldali analízisre és a dekódoló oldali szintézisre. A korlátozott átviteli sávszélesség miatt temészetesen mind a maradékjelet, mind az LPC paramétereket kvantálni kell. Vegyük észre, hogy az analízisszűrő is kvantált paraméterekkel működik (lásd Q_2), noha a kódoló oldalon a kvantálatlan paraméterek is rendelkezésre állnak, viszont így elérhető, hogy a szűrők mindkét oldalon azonos paraméterekkel működjenek, tehát *egymás pontos inverzei* legyenek, ami sokkal jobb az átviteli minőség szempontjából, mint ha a kvantálatlan paramétereket használnánk az analízisszűrőnél. (Megjegyezzük, hogy a digitális jelfeldolgozó eszközök természetesen digitálisan reprezentált, tehát kvantált mennyiségekkel számolnak, de ezek pontossága csak a számábrázolás végessége miatt korlátos, ezért a jelen tárgyalásban pontosnak tekintjük ezeket. A konkrét algoritmusok lebegőpontos megvalósítása során ez jó közelítés, azonban a jelfeldolgozó processzorok gyakran használnak hatékonyabb, fixpontos számításokat, ahol az algoritmusok kerekítési hibáinak hatását is figyelembe kell venni mind az algoritmusok tervezésénél, mind a kimeneti jel torzulásánál.) Vegyük észre azt is, hogy a Q_2 kvantáló kimenetén megjelenő kvantálási hiba a vételi oldalon áthalad a szintézisszűrőn, és így az ADPCM struktúrával ellentétben a spektruma a beszédjeléhez hasonlóra formálódik. Emiatt viszonylag durva kvantálóval is jó szubjektív beszédminőség érhető el, mert a fülünk sokkal kevésbé érzékeny a gerjesztőjel formájára, mint a spektrális burkolóra. Jelentős részben ennek a spektrumformáló hatásnak köszönheti a legegyszerűbb AaS struktúra a kezdeti sikereket a beszédátvitel terén.

A zöngés szakaszok, keretek hatékonyabb modellezése érdekében az analízist célszerű egy rövid idejű és egy hosszú idejű predikciós blokkra bontani. Utóbbi a zönge periódusát modellezi. Többnyire egy mintavételi periódus pontossággal határozzák meg az optimális késleltetést, de használnak ennél pontosabb analízist is, amely a mintavételi időn belüli értékeket is megadja. A 7.25. ábrán az egyszerűség kedvéért mind a rövid idejű STP (Short Time Prediction), mind a hosszú idejű LTP (Long Time Prediction) elemeket előre irányú struktúrában mutatjuk, de nem ritkán alkalmazzák az LTP-t és néha az STP-t is hátra irányú prediktív struktúrában (lásd a 7.2.2

fejezetet), illetve elterjedt a visszacsatolt adaptív vektorkvantálós hosszú idejű predikciós megoldás is, amit majd a CELP struktúránál látunk. Gyakorlati szempontból érdekes, hogy kódolóknál az STP blokk megelőzi az LTP-t, mert a kísérletek szerint így hatékonyabb a predikció, vagyis átlagosan kisebb maradékjel adódik. A fenti



7.25. ábra. AaS rövid és hosszú idejű predikcióval

LTP analízisszűrő-blokk paraméterei a b súlyozó tényező és a zönge periódusának megfelelő késleltetési idő N_0T, ahol T a mintavételi periódusidő. Ezzel az LTP analízisblokkra vonatkozó transzferfüggvény

$$B(z) = 1 - L(z) = 1 - bz^{-N_0}, \tag{7.78}$$

míg a megfelelő szintézisszűrő transzferfüggvénye

$$B(z) = \frac{1}{B(Z)} = \frac{1}{1 - L(Z)} = \frac{1}{1 - bz^{-N_0}}. \tag{7.79}$$

Az optimális b súlyozó tényezőt N_0 ismeretében úgy határozzuk meg, hogy a maradékjel energiája minimális legyen. Ha az LTP bemenő jele $s(n)$, és a kimenetén a maradékjel, vagyis a hibajel $e(n)$, akkor

$$e(n) = s(n) - bs(n - N_0). \tag{7.80}$$

A maradékjel energiájára írhatjuk, hogy

$$\overline{e(n)^2} = \overline{s(n)^2} - 2b\overline{s(n)s(n-N_0)} + b^2\overline{s(n-N_0)^2}, \tag{7.81}$$

ahol a felülvonással az egy keretre vonatkozóan az n index szerinti átlagolást jelöltük. Ezt az egyenletet b szerint deriválva kapjuk az optimális b értéket:

$$b_{opt} = \frac{\overline{s(n)s(n-N_0)}}{\overline{s(n-N_0)^2}} = r(N_0), \tag{7.82}$$

ahol $r(N_0)$ jelöli a normalizált autokorrelációs együtthatót N_0 késleltetés esetén.

A zöngeperiódus értékének meghatározásával az előző fejezet foglalkozott. Elterjedt módszer, hogy minden késleltetési időre kiszámolják az autokorrelációs együtthatót, és a maximumot adó késleltetést keresik meg (peak picking). Mivel a zöngeperiódus gyorsabban változhat, mint a jelspektrum burkolója, ezért a kereteket tipikusan négy alkeretre szokás osztani, és amíg az STP-t keretenként, addig az LTP analízist alkeretenként érdemes újra elvégezni. 8 kHz mintavételi frekvencia esetén a zöngeperiódusra (5–15 ms) általában a 40–120 mintavétel jut, tehát $120-40 = 80$ különböző késleltetési értékre érdemes az autokorrelációs együtthatót meghatározni, amely egy 160 minta hosszúságú (20 ms-os) ablak esetén együtthatónként 160 szorzást és akkumulálást jelent (lásd a 7.37 képletet). A szorzás és akkumulálás a DSP-k esetén egyetlen lépésben történik, ily módon a kiegészítő feladatok elhanyagolásával is 80x160 = 12 800 műveletet kell elvégezni 5 ms-os alkeretenként egyetlen N_0 érték meghatározásához, ami 12800/5ms = 2,56 MIPS (Million Instructions Per Second) sebességet jelent erre az egyetlen részfeladatra.

Összehasonlításul: jelenleg a nagy teljesítményű jelfeldolgozó processzorok többszáz vagy akár többezer MIPS sebességet is nyújtanak, különösen, ha órajelciklusonként párhuzamosan több utasítást hajtanak végre, azonban a gyors órajelnek és a párhuzamos működésnek az ára a nagyobb fogyasztás, ami a mobil eszközöknél hátrányos, ezért is fontos a műveletek számának minél alacsonyabb ($<4\ldots40$ MIPS) értéken tartása, így lassabb processzorokkal is megvalósítható a kódolás. Amennyiben az adott keret zöngétlen jelhez tartozik, az LTP algoritmus akkor is szolgáltat valamilyen „eredményt”, amely ekkor persze nem zöngeperiódus, és így gyakorlatilag nem csökkenti a maradékjel szintjét, de nem is zavaró.

Az LTP szintézisszűrő akkor stabil, ha $b<1$. Ez könnyen belátható, mert ellenkező esetben egyetlen bemenő impulzus hatása sem csillapodna. Viszont az analízisnél gyakran adódik $b>1$ érték a zöngés hangok elején, ezért egy definiált rövid időre az algoritmusok 1,2–2 körüli b értéket is megengednek a jobb predikció kedvéért, mert a hosszú idejű stabilitás így is biztosítható.

Az LTP módszer régóta ismert, már az ADPCM struktúrában is használták, az úgynevezett adaptív prediktív kódoló (APC: Adaptiv Predictive Coder) esetén, ahol a visszacsatoló hátrairányú adaptív hurokban helyezték el a hosszú idejű prediktort, és az első GSM kódolóknál is ilyen megoldást alkalmaztak. Megjegyezzük, hogy hatékonyabb LTP érhető el, ha egy mintavételi periódusnál pontosabban határozzák meg a zöngeidőt, amelyhez egyetlen b paraméter helyett kettő, esetleg három súlyozó tényezőt használnak. Az ilyen algoritmusok a mintavételi frekvencia növelésén vagy interpoláción alapulnak.

Problémát okoz viszont, hogy a magasabb alapfrekvenciájú, női vagy gyermekhangoknál 2 vagy 3 periódus is beleeshet egy keresési ciklusba, így az egymást követő kereteknél az N_0 értéke „ugrálhat”, mert hol az alapnál, hol egy többszörös periódusnál található az abszolút maximum. A többszörös periódusok kiszűrését

érdemes beépíteni a kereső algoritmusba, mert N_0 értékének ugrásszerű változása hallható torzítást okoz.

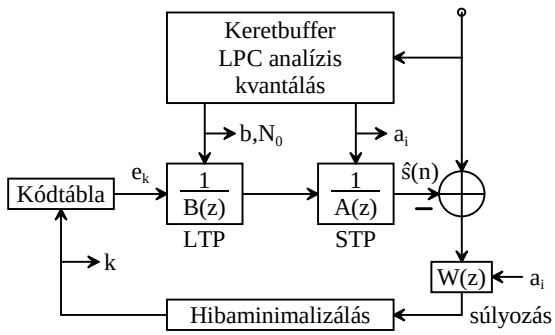
Az LTP-nél bemutatott számítási módszer alkalmazható az STP esetre is, ahol a prediktor fokszáma p. Ennek értékét a statisztikai vizsgálatok alapján 8–12 között érdemes megválasztani. Ennél hosszabb STP már nem csökkenti számottevően a maradékjelet, mert a késleltetés növekedésével természetesen csökken a jelszakaszok korrelációja. Ebből is látható, hogy célszerű az STP és LTP szétválasztása, mert a 12 feletti LP együtthatók közül már csak az LTP együtthatója járul hozzá számottevően a maradékjel szintjének csökkentéséhez. Az LTP-hez hasonlóan az a_i jelű LP együtthatókra egy p ismeretlenes lineáris egyenletrendszert kapunk, amelyet az $i=1...p$ késleltetésekre meghatározott autokorrelációs együtthatók ismeretében oldható meg. Ennek során, az úgynevezett autokorrelációs módszer esetén feltételezzük, hogy az ablakolt jel értéke mindenütt nulla az ablakon kívül. (Létezik egy kevésbé „népszerű", úgynevezett kovariancia-módszer is, ahol kissé mások a matematikai feltételek.)

Az STP analízis 10–30 ms-os ablakolt szegmensekben, keretekben történik, mert ekkora szakaszon a beszédjel viszonylag jó közelítéssel stacionáriusnak tekinthető. Számos matematikai módszer és számítástechnikai algoritmus is létezik az optimális a_i LPC paraméterek meghatározására. Ezek a maradékjel energiáját (négyzetes átlagértékét) minimalizálják a beszédjel véges szakaszaira. A hatékony, főként rekurzív számítási módszereket a vonatkozó irodalom részletesen tárgyalja. (Gordos–Takács 1983, Rabiner–Schafer 1978, Chu 2003). Leginkább a Levinson–Durbin algoritmus terjedt el, azonban fixpontos aritmetika esetén ez dinamikaproblémákhoz vezet, amely gondos programozással vagy lebegőpontos számábrázolással kezelhető. Ebből a szempontból a Leroux–Gueguen-algoritmus kedvezőbb, viszont ez nem az LPC paramétereket, hanem a szintézis rácsszűrős struktúrájának paramétereit adja, amely ugyan a szűrő stabilitásának ellenőrzése szempontjából előnyös, viszont számításigényesebb. A részleteket illetően ismét a bő irodalomra utalunk.

Az egymást követő keretek jelszakaszai többé-kevésbé hasonlóak, azonban ezek LPC paraméterei teljesen eltérőek, így a keretek között hatékonyan nem is interpolálhatók. Ezért a paramétereket érdemes transzformálni olyan kedvezőbb reprezentációkra, amelyek egyrészt folyamatosabban változnak keretről keretre, hogy a szűrőkarakterisztika keretenkénti ugrásait simíthassuk, másrészt, amelyek eloszlása kedvező a paraméterek kvantálása szempontjából. Számos módszert dolgoztak ki az LP együtthatók transzformálására, kvantálására és kódolására. Erre a szakirodalomban találhatunk részletesebb ismereteket.

7.2.4. Zárt hurkú predikciós kodoló

A legsikeresebb AbS (Analysis-by-synthesis) módszer a CELP (Code Excited LP) kódgerjesztésű lineáris predikció. A zárt hurkú AbS kódoló elvét a 7.26. ábra szemlélteti. A kódgerjesztés céljára egy kódtáblában tárolunk 2^k gerjesztő vektort, ame-



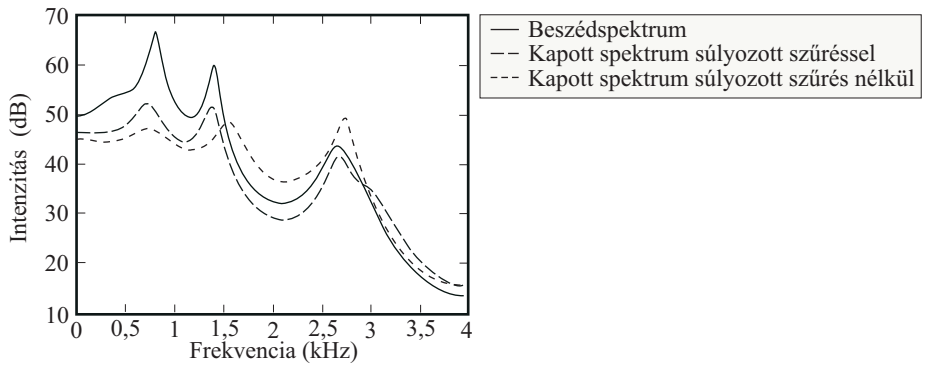
7.26. ábra. CELP kódoló struktúra

lyek közül a k index jelöli ki az aktuális e_k gerjesztő jelsorozatot. A szintézisszűrőt az előző fejezetben tárgyalt LTP és STP blokkok valósítják meg, majd a kapott szintetizált $\hat{s}(t)$ jelet összevetjük az $s(t)$ bemeneti jellel. Ennek az AbS módszernek a lényege az, hogy a szintetizált jel hibájának minimalizálása zárt hurokban történik. A különbségi jelet még egy $W(z)$ súlyozó szűrőn is átvezetjük a spektrum leginkább zavaró tartományainak kiemelésére, majd a kódtábla összes vektorát végigpróbálva megkeressük a legkisebb súlyozott hibát okozó gerjesztő vektort. Az $L(z)$ és $P(Z)$ prediktorok paramétereit a bemeneti jel analíziséből nyerjük. Ezeket tehát nem változtatjuk a hiba minimalizálása során. A vételi oldalra csak az $a_i(i = 1...p), b, N_0$ kvantált paramétereket, valamint a kódtábla k indexét kell átküldeni. A $W(z)$ súlyozó szűrő azokat a spektrális komponenseket csillapítja főleg, amelyek szubjektív hatása a beszédjel maszkoló hatása miatt kevésbé zavaró. Emiatt ez a szűrő is felhasználja az STP céljára meghatározott LP együtthatókat, és transzferfüggvénye

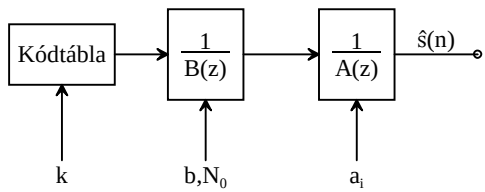
$$W(z) = \frac{A(z)}{A(\frac{z}{\gamma})} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}}, \tag{7.83}$$

ahol a szubjektív tesztek alapján $\gamma = 0,8 - -0,9$ értékeket szoktak választani. Látható, hogy $\gamma = 1$ esetén egyáltalán nincsen súlyozás, míg $\gamma < 1$ esetén a formánsoknál csillapít és a spektrális völgyekben kiemel a súlyozás, ezzel hozzájárul a maszkolási hatáshoz, mert a hibajel spektrumában így a formánsoknál keletkezik kiemelés

(7.27. ábra). A 7.28. ábrán látható dekódolás pontosan megegyezik a kódolóban alkalmazott szintézissel, tehát ugyanolyan kódtábla és szintézisszűrő van a dekódolóban is, mint a kódolóban. Tekintettel arra, hogy egy teljes keresés a kódtáblában



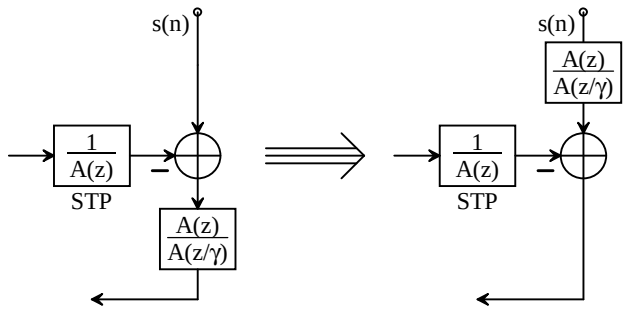
7.27. ábra. A kódolási hiba érzeti súlyozása a frekvencia függvényében



7.28. ábra. CELP dekódoló struktúra

igen számításigényes, hiszen minden kódvektor esetén végig kell számolni a szintézisműveleteket, számos módosítást javasoltak ehhez az algoritmushoz. Kézenfekvő egyszerűsítést jelent, ha a $W(z)$ súlyozó szűrő transzferfüggvényének $A(z)$ számlálóját összevonjuk az STP-vel a 7.29. ábra szerint, ahol kihasználtuk a különbségképzés linearitását. Ily módon a hibaminimalizálási ciklusokban nem kell a súlyozást elvégezni, csupán egyszer, a bemenő jelen, továbbá egy módosított analízisszűrővel kell az STP-t végezni, ami nagyon jelentősen csökkenti a számítási igényt.

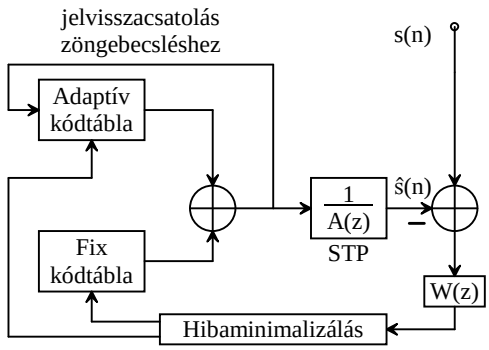
A kódtábla tartalmának meghatározásához hagyományosan tényleges beszédjelet használnak, amelyre minimalizálják a kvantálási hibát. Maga a hiba a kvantált és az eredeti jelvektorok euklideszi távolsága, amelynek szintjét rekurzív algoritmusokkal lehet minimalizálni, vagyis egy adott jelstatisztikához létrehozni az optimális vektorkvantálót (Gersho–Gray 1995). Kisebb méretű kódtáblával jelentősen csökken a keresési műveletek száma, azonban ez a gerjesztőjel pontatlanságához vezet. Jó kompromisszum, ha több kódtábla vektorainak összegeként állítjuk elő a gerjesztőjelet. Széles körben alkalmazzák a kettős, úgynevezett **konjugált kódtáblát**, amelynél két



7.29. ábra. A *W(z)* súlyozó szűrő kivétele a visszacsatoló hurokból

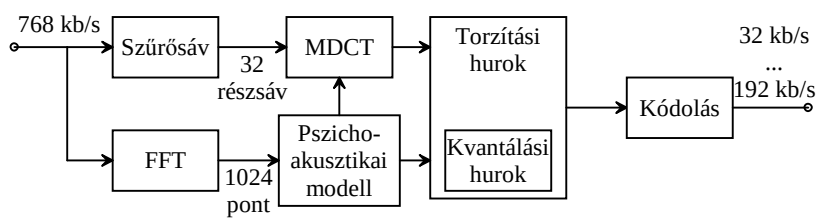
vektor összege a gerjesztő jel, és a vektorokat úgy optimalizálják, hogy ha az egyik index megsérül az átvitel során, akkor a másik vektorral a lehető legkisebb legyen a gerjesztőjel hibája. A kódtáblakeresés igen jelentős egyszerűsítését eredményezte a korszerű kódolókban alkalmazott úgynevezett **algebrai kódtábla**. Ennek gyakorlatban is alkalmazott változatánál a táblában csak +1 vagy –1 amplitúdójú impulzusok vannak, mégpedig minden vektor esetén ezek is csak egyenletesen elosztott fix pozíciókban, a többi pozíció tartalma 0. Több, például 4 ilyen vektor összeadásából képződik végül a gerjesztő vektor, amelynek megadásához eszerint nem egy index szolgál, hanem minden egyes vektornál meg kell adni a polaritást és a pozíciót. Vegyünk egy példát. Ha 40 mintavételi periódusidő (5 ms) esetén egy vektor minden 5. pozíciójában, tehát 8 hely valamelyikén lehet az impulzus, akkor 1+3 = 4 bittel definiálható egy-egy vektor. Noha ez a módszer több bitet használ a kódgerjesztés megadására, mint az egyszerű kódtáblaindex, egy kvantált erősítési tényezőt tartalmazó kódtáblával kiegészítve viszonylag pontos gerjesztő vektorokat szolgáltat, és így a gyakorlatban hatékony és jó minőségű kódolást eredményez.

A korszerű CELP kódolókban az LTP korábban tárgyalt módszere helyett egy külön **adaptív kódtáblát** alkalmaznak, amelybe a jel korábbi mintáit tárolva érhető el egy zöngeperiódusnyi késleltetés. A megoldást a 7.30. ábra mutatja. Az adaptív kódtábla a gerjesztőjel korábbi szegmenseit tartalmazza. Az AbS elvnek megfelelően ezek közül kell kiválasztani azt a szegmens vektort, amely a legkisebb hibával szintetizálja a jelet. Zöngés jel esetén ez éppen a zöngeperiódussal késleltetett vektor lesz. Annak ellenére, hogy a keresési hurokban érzeti súlyozó szűrőt alkalmazunk, a helyreállított beszédben a kódolási hibák hatására hallható torzítás jön létre, amely kissé „durva” színezetet ad a beszédnek. Ugyanis a formánsok közötti tartományban és a zönge harmonikusok között az említett maszkolási hatás nem elegendő. Ezért számos dekódoló kimenetén egy komplex **utószűrőt** helyeznek el, amely az említett frekvenciasávokban tovább csökkenti a kódolási hibát, vagyis ennek spektrumát áttolja a nagyobb energiájú formánstartományokba.



7.30. ábra. Adaptív kódtábla alkalmazása hosszú idejű prediktor helyett

Részsávú és transzformációs kódolás. A fentiekben tárgyalt kódolási módszerek és változatos struktúrák nemcsak a beszédjel egészére, hanem annak részsávjaira külön-külön is alkalmazhatók. Ez a **részsávos kódolás.** Tekintettel arra, hogy a fül érzékenysége eltérő a különböző spektrális tartományokban, jelentős további tömörítés érhető el, ha az egyes sávokban eltérő pontosságú kódolást alkalmazunk. A keskenyebb sávokban a mintavételi frekvenciát is megfelelően csökkenthetjük (decimálhatjuk), de ennek ellenére a részsávokra bontás valamelyest növeli a kódoló bonyolultságát. Általánosan használatos még az úgynevezett **transzformációs kódolás**, főleg zenei és videojelek esetén, amely a részsávos kódolás kiterjesztésének tekinthető, ahol nem sávszűrőkkel, hanem a frekvenciatartományba való transzformáció útján bontjuk szét a beszédjelet, bár a két módszer közötti határvonal nem éles, inkább tervezési kérdés, hogy szűrősávokat vagy transzformációt alkalmazunk. A 7.31. ábra az MPEG Layer-3, ismertebb nevén az MP3-kódoló egyszerűsített blokksémáját mutatja, amely jellegzetes példája a szűrősávval induló részsávos kódolóknak. (Brandenburg–Popp 2000). Ez a struktúra audiojelekre van optimalizálva, ennek megfelelően a bemeneten 768 kb/s sebességű zenei jelet fogad. A jelet egy



7.31. ábra. Az MP3-kódoló egyszerűsített blokksémája

szűrősáv bontja 32 részsávra, majd 18 elemű módosított diszkrét koszinusztranszformációval (MDCT) áll elő a spektrális felbontás 18x32 = 576 eleme. Az MDCT abban különbözik a DCT-től, hogy átlapolt szegmenseket használ, ami simító hatású,

csökkenti a blokkhatárok hatását. A transzformációnál szükséges ablakoláshoz négyféle, rövidebb vagy hosszabb ablakot lehet választani aszerint, hogy milyen gyorsan változnak a jel statisztikai jellemzői. Ezt is a pszichoakusztikai modell alapján határozzák meg, illetve egy kéthurkos szabályozási körben a kiszámított maszkolási hatás alapján több menetben optimalizálják a kvantálási szintek kiosztását és a torzítást, a kívánt kimeneti sebességhez illesztve. Végül a kimeneti jelet egy összetett, kódolási eljárás szolgáltatja. Maga a vonatkozó szabvány csak a jelfolyamot írja elő a kompatibilitás érdekében, és igen nagy szabadságfokot ad a megvalósítási algoritmusra, ezáltal nem állja útját a technológia és a tudomány fejlődésével egyre jobb kódolók és dekódolók kidolgozásának. Figyelem: mivel nem szabványos a megvalósítás, nem minden MP3-lejátszó nyújt azonos minőséget. Fontos azt is megjegyezni, hogy az MP3-kódoló kimondottan nem beszédjelre optimalizált. Ezért az MP3-kódolóval tömörített jelet beszédinformációs rendszerekben (például beszédfelismerésben) kisebb hatékonysággal lehet felhasználni, különösen a beszédkódolásnál használatos bitsebesség-tartományban.

8. fejezet
Adatbázisok a beszédtechnológia szolgálatában

Vicsi Klára

Miért van szükség beszédadatbázisra? Azért, mert a beszéd biológiai produktum, dinamikusan változó akusztikai jel, sokféle változatossággal. Akár ugyanaz a mondat ugyanattól a személytől más-más hullámforma-összetételben jelenhet meg (mérges, szomorú stb.). A különböző beszédhangok létrehozása a beszédképző szervek számos különböző állapotát és mozdulatát foglalja magába, továbbá a hangképző szervek időzítésének és állapotának is nagymérvű szabadsága van. Következésképpen, a körülményektől, a gondolati tartalmaktól függően a beszélő módosíthatja beszédét, eközben a létrehozott akusztikus jel paraméterei is változnak, ugyanakkor az elhangzott közlemény még mindig ugyanazt a nyelvi tartalmat közvetíti a hallgató számára. A változékonyság és a változatlanság témája iránt évtizedek óta nagy az érdeklődés a beszédtudományban. Melyek azok a jellemzők a sok redundáns jellemző közül, amelyek a különböző körülményekre invariánsak? A kutatók egy része ezeket az invariáns jegyeket keresi, míg mások a beszéd változékonyságát vizsgálják, a változékonyság korlátait igyekeznek jobban megérteni. Azonban van, ami közös e két megközelítésben. Gunnar Fant így ír erről a problémáról: *„Több beszélő elegendően nagy adatháttere nélkül nem tudunk betekinteni a változatlanságba és a lényegbe. Viszont a beszédanalízis és a beszédlétrehozás általános ismerete nélkül azt nem fogjuk tudni, hogyan kell releváns adatokat gyűjteni a beszélővváltozatokról. A beszédváltozatok tanulmányozása kikövezi az utat az invariánsok tanulmányozásához, és viszont."* (Fant et al. 1990). Ezek szerint az adatháttér, vagyis a nagyszámú beszédminta az, ami lehetőséget teremt a változékonyság és a beszédben rejlő invariancia tanulmányozására. Ilyen beszédminták feldolgozott gyűjteménye a beszédadatbázis. Mivel a beszéd természetére jellemző a fizikai paraméterek nagymértékű variáltsága egy beszélőn belül, beszélők között, továbbá az akusztikai környezet függvényében is, a beszédadatbázisok feladata az, hogy mintáikkal ezt a nagyfokú változatosságot egy adott szempont szerint minél jobban lefedjék.

Mindezek után definiáljuk a beszédadatbázis fogalmát: a beszédadatbázis nagy méretű, nyelvileg és akusztikailag feldolgozott, tárolt hangadathalmaz, magyarázó

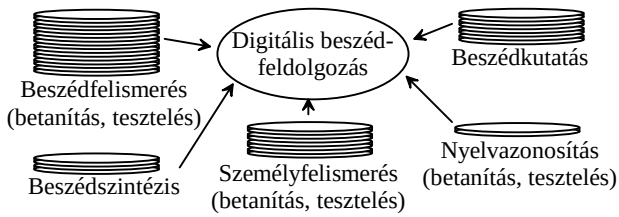
jegyzetekkel, címkézésekkel és átírásokkal ellátva, amelyet sokféle csoportosítás szerint hozhatunk létre. Az adatbázis nagyságát és belső szerkezetét általában a felhasználási terület határozza meg. A rádióból, televízióból, internetről felvett beszéd nem adatbázis, mivel nincs ellátva keresést biztosító belső jelzésekkel. Az adatbázis lényeges tartozéka a precízen leírt dokumentáció a rögzítés technikájáról, a beszélők számáról és típusáról, a nyelvi tartalomról. Az adatbázist felhasználók a gyűjteményre vonatkozó részletes ismertetést ebből a dokumentációból kapják meg. A megfelelő minőségű adatbázisok kulcsfontosságúak a beszédtechnológiában.

A beszédjel változatossága. A beszéd fizikai rezgésének belső tartalmát a beszélő személy és annak közvetlen környezete határozza meg. A beszédképzés folyamán a fonémarealizációk (beszédhangok) paraméterei számos hatás következtében – bizonyos határok között – megváltoznak. Ezt a jelenséget nevezzük a beszédjel változatosságának. Ez több elemből tevődik össze. A folyamatos hangképzőszervi mozgások miatt (koartikuláció) a beszédhangok fizikai tulajdonságai befolyásolják egymást a kapcsolódás folyamán (lásd az 5.2. fejezetben). A beszédstílus szintén befolyásolja a létrehozott hangsorozat fizikai jellemzőit, de változhatnak ezek a jellemzők az érzelmi és egészségi állapot szerint is. Az akusztikai környezet (stúdió, utca, autó) szintén erősen befolyásolja a véglegesen kialakult nyomáshullám (beszéd+környezet) fizikai jellemzőit. A változatosságot okozó tényezők számos módon csoportosíthatók, mégis talán a beszélők szerinti és a környezeti hatások szerinti csoportosítás illik a legjobban a beszédtechnológia témakörbe. A beszélő személy szerint megkülönböztetünk beszélőn belüli (intraindividuális), valamint beszélők közötti (interindividuális) változatosságot. Az egy beszélőn belüli változatosság a személy pillanatnyi jegyeit viseli a beszéléskor. Ilyen a vérmérséklet, az egészségi állapot (megfázás), az érzelmi állapot stb. A beszélők közötti változatosság abból ered, hogy mindannyian mások vagyunk, más-más egyéni jellemzőkkel vagyis egyénenként változik a testi felépítés, a hangképző szervek méretei és állapotuk, az artikulációs vezérlés agyi formái (például gyors beszédű) stb. Itt belép a korbeli és a nemek közötti különbség, valamint a szociolingvisztikai tényező is (tájnyelv). A beszédjel változatosságához hozzá kell sorolni a beszédstílusokat is, vagyis a kontextusváltozatosságot. Az elmondandó nyelvi anyag tartalma magával hozza a hozzárendelt beszédmodort (Olaszy 2005). Ezt főleg felolvasásos beszédnél kell figyelembe venni (hírolvasás, hirdetés, mese, regény, ismertetés, narrátori beszéd).

A környezeti hatások, valamint a rögzítő, feldolgozó berendezések akusztikai jellemzői hozzáadódnak a beszédjelhez, ezzel annak további változatosságát okozhatják (például a zajos, zajtalan környezet, a visszhangok, a termek és távközlési csatornák tulajdonságai, a telefonkészülékek átvitele stb.).

A beszédadatbázisok fő jellemzői. Az utóbbi évtizedekben sokféle beszédadatbázist hoztak létre a beszédkutatás, az oktatás, a fejlesztői munka támogatására. Fő

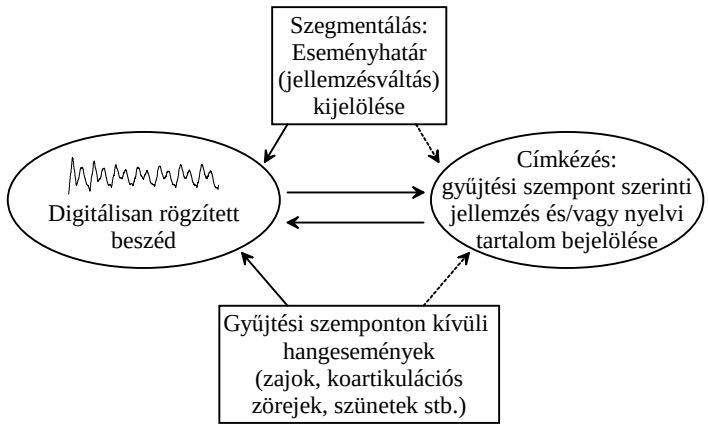
jellemzőjük ezeknek az adatbázisoknak, hogy tartalmilag és belső feldolgozásuk szerint különböznek egymástól. Ez attól függ, hogy milyen felhasználási területre készültek. Ennek függvénye, hogy mekkora a bennük gyűjtött beszédanyag mérete, hány bemondóval készítették, milyenek voltak az akusztikai körülmények stb. (8.1. ábra). A beszédtechnológiával foglalkozó szakemberek számára fontos, hogy



8.1. ábra. Beszédadatbázisokat használó tématerületek és a hozzájuk készített beszédadatbázisok hozzávetőleges méretarányai

ismerjék a már elkészített beszédadatbázisokat, hogy közülük egy meghatározott feladatra a legmegfelelőbbet tudják kiválasztani. Ha nincs még a célnak megfelelő adatbázis, akkor azt létre kell hozni. Ehhez adjuk meg az adatbázis készítésének legfontosabb ismérveit a következőkben.

Beszédadatbázisok tervezése, készítése. Az első szempont, hogy milyen feladat segítésére, elvégzésére kell a beszédadatbázis. Ez határozza meg a gyűjtési szempontot. A beszédadatbázisokban a digitálisan rögzített beszédjelet a gyűjtési szempont szerinti jellemzéssel és/vagy a rögzített nyelvi tartalommal kötik össze. A beszéd digitális tárolása mellett tehát rögzíteni kell a gyűjtési szempont szerinti jellemzést, a verbális és nem verbális információt is. Ez a jellemzés többrétegű lehet. Tartalmazhatja a hanganyag nyelvi tartalmát karakteresen, de tartalmazhatja az adott hangszakasz prozódiai vagy érzelmi tartalom szerinti lejegyzését is stb. Az adatbázisnak tartalmaznia kell továbbá az eseményhatárokat, vagyis az időfüggvénybe vagy a karakterláncba bejelölve azokat a helyeket, ahol a gyűjtési szempont szerinti jellemzés a hanganyagban megváltozik. A leggyakoribb ilyen feladat a mondat- szó- és hanghatárok automatikus bejelölése. Ez például a beszédfelismerés során használt akusztikai-fonetikai modellek betanításának támogatására készített beszédadatbázisokban fontos. Más eseményhatárok vonatkozhatnak például az érzelmi állapot változására. Az ilyen adatbázisban jelölni kell, hogy a hangfolyam mely pontján vált át a beszélő a normál társalgási hangnemből például ingerült hangnemre, esetleg más érzelmeket tükröző beszédstílusra. Az adtatbázisban előfordulhatnak a gyűjtési szemponton kívüli hangesemények is, például zajok, beszéd közbeni zörejek, szünetek. Az adatbázis szegmentálása tehát minden esetben egyedi jelölések elhelyezését jelenti (vö. 8.5. ábra). A digitálisan rögzített beszéd és



8.2. ábra. Beszédadatbázis szegmentálási szempontjai

a hozzá tartozó különböző információk összekötését a hanganyag annotálásával és címkézésével szokás végrehajtani.

Előfeldolgozás. Az előfeldolgozás a hangfelvételek lehallgatásából, ellenőrzéséből és esetleges feldarabolásából áll. Az ellenőrzés körébe tartozhat a tartalmi ellenőrzés (az van-e benne, aminek lenni kell), a tartalom optimalizálása (a hasznos tartalom elejének és végének meghatározása stb.).

Annotálás. Annotálásnak nevezzük azt az eljárást, amikor a hangfelvétel tartalmával kapcsolatos általános információkat adjuk meg. Az annotálás során minden hangfájl mellé készítünk egy címkefájlt, amely információkat tartalmaz a hangfájl paramétereivel és tartalmával kapcsolatosan. Ilyenek például: az elhangzott szöveg ortografikus lejegyzése (ha nincs meg), rendszerint az adott nyelv betűkészletével, valamint SAMPA fonetikai hangkarakterekkel (Wells 1997). A hibás kiejtést, a nem érthető szavakat, szótöredékeket is jelölni kell, valamint a beszélő nem beszédből származó hangjait (csettintés) és az esetleges környezeti zajokat is. Az ortografikus szöveg átírása a kiejtést megadó hangsorozattá az egyik legalapvetőbb annotálási feladat, hiszen ez fogja írott formában megadni, hogy milyen hangoknak kellene lenni a beszédjelben. Az adatbázisban rögzített beszéd ortografikus formájából (a hullámforma figyelembevételével) általában elkészítik annak fonetikai átiratát, ahol már beszédhangok szimbólumaival írják le a beszéd tartalmát. Itt IPA vagy SAMPA nemzetközi hangjelöléseket kell alkalmazni. Az átírásnak számos gyakorlati formája létezik. A szabványosított eljárások a következők.

Kanonikus átírás: Az adott szöveg karaktereinek olyan átírása, amelyben az ortografikus karaktereket beszédhang szimbólumok sorozatára alakítjuk ki, de az adott szövegkörnyezetet nem vesszük figyelembe (ezt fonemikus átírásnak is nevezik). Te-

hát a hasonulás és a koartikuláció (például hiátustöltés) nincs figyelembe véve. Például: *háztető* h A: z t E e t 2:, *2 háztető* k E: t h A: z t E e t 2:. Ebben az átírási formában tehát nem a hanghullám fonémareprezentációs elemeit kapjuk meg. A példákban SAMPA jeleket használtunk.

Fonotipikus átírás A karakterek átírását hangszimbólumokká az adott nyelv fonetikai szabályai alapján végezzük, a hangkörnyezet függvényében, azaz a hasonulási és egyéb hangváltozási szabályokat is figyelembe vesszük (ez tulajdonképpen a fonetikus átírás). Például: *háztető* h A: s t E e t 2:, *2 háztető* k E: t h A: s t E e t 2:. Ez az átírási forma a hanghullámhoz közvetlenül társítható.

A beszéd fonetikus lejegyzését többféle módon lehet végrehajtani. Kétféle manuális eljárást mutatunk be

Hallás alapján történő fonetikai átírás: Ez a lejegyzési forma a legfontosabb a beszédtechnológiában. Ezt az átírási módszert abban az esetben használják, amikor egy elhangzott közlést kell lejegyezni. A lejegyző figyelmesen hallgatja a beszédet, és hallás alapján, hangszimbólumokkal lejegyzi azt. Például, ha a *háztető* szó [s] hangját elnyeli a beszélő, akkor h A: t E e t 2: lesz a lejegyzett hangsor. Az eljárás alkalmazható annak ellenőrzésére is, hogy egy szöveg és a hozzá tartozó hangfájl tartalma (például felolvasás) pontosan megegyezik-e hangszinten (precízen olvasta-e a szöveget a bemondó). Ilyen ellenőrzésre szükség lehet mind beszédfelismerési, mind beszédszintézis-feladatokhoz készített adatbázisok esetén.

Audiovizuális fonetikai átírás: Az átírást a beszéd hallgatása, és az időfüggvény vagy a színkép elemzése alapján hajtják végre. A legtöbb esetben beszédhang vagy a beszédhangnál kisebb egységek alapján történik a lejegyzés (például az összetett szerkezetű beszédhangok belső állapotait is külön jelölhetik). Ez az átírási forma társítható leginkább a beszéd hullámformájához.

A nagy méretű adatbázisokban a fonetikai átírásokat igyekeznek automatizált eljárásokkal megoldani. Igényes adatbázisoknál az automatikus címkézés hibáinak javítására külön félautomatikus eljárásokat is használnak, amelyekben az ember minimális közreműködése is szükséges (Olaszy–Bartalis 2008). Az ilyen, duplán feldolgozott adatbázisok akár 99,9%-os pontosságúak is lehetnek címkézési szempontból. Az átírásoknál mindig felmetül, hogy milyen módon kezeljék a megakadási jelenségeket (nyelvbotlás, szóismétlés szótöredék után stb.), köhögést, nyögést, amik mind szerves részei a hangfelvételnek.

Címkézés. Címkézésnek nevezzük azt az eljárást, amikor címkékkel látjuk el a hanghullámot (szegmentáljuk, azaz jelzésekkel elkülönítjük a kérdéses részeket). A címkék kapcsolódhatnak a beszédhangok jelöléséhez, esetleg a hangsorban talált zajokhoz (ajtócsukódás, köhögés, háttérzaj), valamint gyűjtési szempontokhoz (például dallamformák határai). A címkéket tartalmazó fájlnak szoros szinkronitásban kell lenni a hullámformafájllal. A szegmentálás és címkézés bonyolult feladat, általában manuálisan végzik, de egyre inkább előtérbe kerülnek olyan

szoftverek, melyek ezt a folyamatot igyekeznek automatizálni. (Vicsi–Vig 1998a, Becchetti–Ricotti 1999).

A beszédadatbázisok csoportosítása. Belátható, hogy nem lehet olyan beszédadatbázist előállítani, ami az összes egyénen belüli és egyének közötti beszédvariációt megfelelő számban tartalmazza, ez ma még képtelenség. Minden vizsgálathoz a hozzá tervezett beszédadatbázist kell létrehozni. Különböző adatbázisok készülnek beszédfelismerők betanítására vagy jó minőségű beszéd előállítására is. A jelenleg elérhető adatbázisok 3 alapkategóriába sorolhatók felhasználás szerint, amit a 8.1. táblázat foglal össze. A nemzetközi tudományos életben elfogadott, beszédadatbázisokra jellemző adatokat a 8.2. táblázatban foglaljuk össze Fourcin–Dolmazon (1991) munkája alapján.

8.1. táblázat. Beszédadatbázisok felhasználási terület szerinti felosztása

Analitikus-diagnosztikus	Alapvetően nyelvi és fonetikai kutatások segítését szolgálja	BABEL (EUROM 0, EUROM 1) CV, VC, CC, VV hangkapcsolódások magyarra
Általános adatbázis	Nem specifikus, általános tartalmú, sokfajta felhasználási területre alkalmas	BEA spontán beszéd adatbázis magyar nyelvre
Specifikus adatbázis	Olyan beszédgyűjtemény, amely meghatározott felhasználási területre készül	MRBA, SPEECHDAT, MTBA, Diád-, triádelemek szintézishez, időjárásbeszédkorpusz, hírekbeszédkorpusz

Nemzetközi adatbázis-gyűjtemények. Európában a European Language Resources Association (ELRA) forgalmazza a legtöbb adatbázist, rendszeres kiadványokban tájékoztatva a szakembereket az újonnan megjelent beszéd és szöveges nyelvadatbázisokról. Az internetes oldalukon (`http://catalog.elra.info/`) részletes információkat kaphatunk az elérhető európai nyelvű adatbázisokról. Itt az egyes adatbázisok forgalmazási árai is szerepelnek (a magas összegek mutatják, hogy nagy emberi és szakmai erőforrások használatával készülnek). Az ELRA katalógus négy fő részből áll: beszédadatbázisok, szöveges adattárak, multimodális gyűjtemények és kombinált adatbázisok. Az alábbiakban mindegyiket részletezzük is.

1. Beszédadatbázisok (Spoken Language Recources SLR)
   a) Telefonos felvételek (Telephone recordings).
   b) Mikrofonnal felvett beszédadatbázisok (Desktop/Microphone recordings).
   c) Híranyag-adatbázisok (Broadcast Resources).

8.2. táblázat. Beszédadatbázisok általános nemzetközi jellemző adatai

Típus	Gyűjtési szempont: fonémavariáltság beszédfelismerők betanítására, szószintű és mondatszintű prozódiavariáltság beszédfelismerők betanítására, érzelemvariáltság, hangképző csatorna egészségi állapotának variáltsága, hanganyag statisztikai alapú beszédszintézis számára, hanganyag fonetikai vizsgálatokhoz
Átalakító	Mikrofon, telefon
A rögzítés fizikai leírása	Mintavételezési és kvantálási paraméterek, felvételi körülmények fizikai leírása
Csatorna	Hangtéri: zaj, visszhang stb. átvitel: sávamplitudó, torzítás stb.
Beszélők	Beszélőfüggőség/függetlenség, nem, kor, fizikai és pszichikai állapot
Nyelvi tartalom	Kitartott hangok, szavak, mondatok, jellemző adat az adott nyelv fonetikai lefedettsége
Beszédstílus	Hangszín: meleg, normál, kiabálás, beszédegység, izolált szavak, folyamatos beszéd, spontán beszéd, beszédsebesség: lassú, normál, gyors, érzelmi kifejezés
Méret	Beszélők száma, rögzített anyag időbeli hossza és nagysága, CD-k száma
Szociolingvisztikai jellemzők	Nem, kor, iskolázottság, beszédstílus
Adatbázis feldolgozása	Címkézés, ortografikus és fonetikus átírás, szegmentálás, spektrális elemzés

d) A rádióban, televízióban elhangzott anyagok gyűjteménye (például híradófelvételek).
e) Beszéddel kapcsolatos egyéb források (Speech Related Resources), például különböző fonetikai és kiejtésszótárak.

2. Írott szöveges adatbázisok (Written Language Recources WLR)

a) Nyelvi korpuszok (Corpora). Nagyméretű szövegek gyűjteménye. Tulajdonképpen a Wikipedia is idetartozik.
b) Egynyelvű lexikonok, szótárak (Monolingual lexicons). Például a Magyar értelmező kéziszótár.
c) Többnyelvű szótár/lexikon (Multilingual lexicons).
d) Tipikusan a kétnyelvű nyelvi kézi/nagyszótárak.
e) Terminológiai adatbázisok (Terminological Language Recources TLR).
f) Többnyelvű források, melyek általában nemzetközileg elfogadott, hivatalos fordításokat tartalmaznak.

3. Multimodális és multimédia-adatbázisok (Multimodal/Multimedia LR)
4. Olyan adatbázisok amelyek a kommunikációnak nem csak szigorúan a beszéd részét tárolják, hanem nonverbális jelek szerint lettek feldolgozva (például gesztusok, arcvonások, mimika), és ezek egy kijelzőn vizualizálhatók.

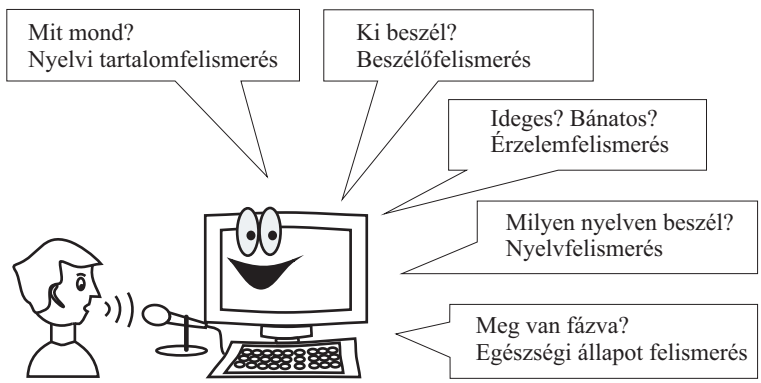
Az Amerikai Egyesült Államokban a Pennsylvaniai Egyetemen, a Linguistic Data Consortium (LDC) gyűjt, forgalmaz és maga is előállít beszédadatbázisokat, szövegkorpuszokat és más egyéb forrásokat, amelyek nyelvészeti kutatásokhoz és fejlesztésekhez szükségesek (`http://www.ldc.upenn.edu/`). Ez egy nyitott konzorcium az

egyetemek, vállalatok és a kormányzat kutatóintézetei számára. Az LDC katalógus hasonlóan az ERLA katalógusához, a beszédadatbázisok és nyelvi korpuszok százait tartalmazza. Magyarországon is készültek beszédadatbázisok, ezekről részletesen később szólunk.

8.1. Tanító adatbázisok gépi beszédfelismeréshez

Vicsi Klára

A számítógépes gépi felismerési feladatok átfogó beszédfeldolgozási témakört jelentenek (8.3. ábra). Az emberi beszédben rejlő bármely információ felismerését jelentheti. A leggyakoribb célkitűzés a nyelvi tartalom kinyerése, de más céllal is készülnek beszédfelismerők. Ilyen egyéb cél lehet a beszélő személy felismerése, a beszédben lévő érzelem felismerése, a beszélt nyelv vagy a beszélő egészségi állapotának a felismerése. Mindegyik felismerési feladat megoldására más-más adatbázist

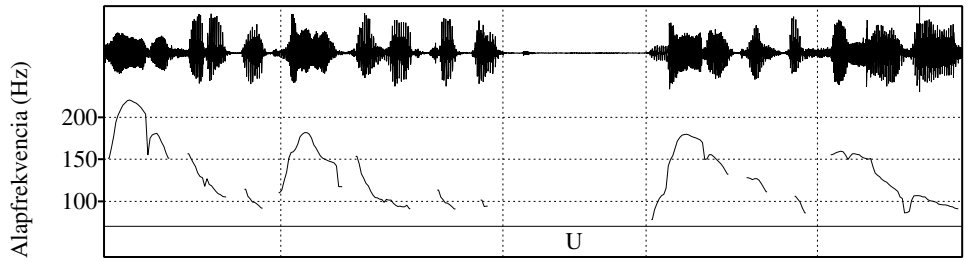


8.3. ábra. A beszéd gépi felismeréséhez kapcsolódó területek

kell alkalmazni a betanításhoz. A gépi felismerési eljárások ez idáig lényegében két jól elkülöníthető elméleti alapra épülnek. Az egyik a szabálybázisú megközelítés (kognitív módszer), a másik a statisztikai elméleti alapú feldolgozás (információelméleti megközelítés). Statisztikai alapú feldolgozást használnak a rejtett Markov-modell (Hidden Markov Model, HMM), vagy neurális hálózatok (Neural Network, NN) használatával megvalósuló felismerők (Levinson et al. 1983). A mai beszédfeldolgozási tudásszinten a gyakorlatban megvalósuló sikeres felismerő rendszerek statisztikai alapokon működnek. A beszéd statisztikai modelljének igazodni kell a beszéd nagyfokú változatosságához, így sok dimenziós paraméter felülettel kell rendelkeznie. Egy pontos paraméterbecslési lépés (betanítási lépés) csak nagyszámú minta vizsgálataalapján lehet sikeres. E minták gyűjteményei – a szükséges jegyzetekkel,

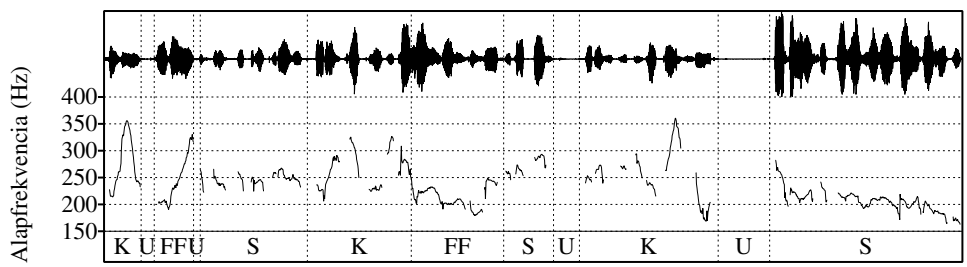
címkézésekkel és átírásokkal ellátva – képezik a tanító adatbázist. A statisztikai alapú gépi felismerésnél a tanítás során egy véletlen folyamat (X) előállít egy diszkrét idejű véletlen jelet ($X_{(n)}$), ahol n a diszkrét időindex. Meg kell becsülni azokat a paramétereket, amelyek a modellt jellemzik. A becslés pontossága arányos a rendelkezésre álló megvalósulások számával. Ha minden lehetséges realizáció rendelkezésünkre állna, akkor ismernénk az X véletlen folyamat együttes felületét. Ebben az esetben a modell teljesen pontos valódi paramétereit becsülhetnénk meg. Kísérletek bebizonyították, hogy nincs értelme ugyanazon folyamat realizációiból túl sokat, hosszú ideig gyűjteni. Lényegében haszontalan ugyanazon beszélő hangjának sokkal többszöri rögzítése, vagy több felvétel készítése, mint amennyi az együttes felület lefedéséhez szükséges. Százhoz közeli, megfelelően kiválasztott beszélő elégséges és hatékony a beszédhang egységű, nyelvi tartalmat felismerő rendszer beszélőfüggetlen betanításához. Tovább növelni számukat haszontalan, sőt néha káros!

A gyakorlatban a tanító adatbázisok tervezésénél tehát azt kell figyelembe venni, hogy az adatbázis létrehozása nem más, mint a véletlenszerű folyamat egyes megvalósulásainak (realizációinak) összegyűjtése egészen a tervezett használati mód közel teljes lefedéséig. Bármelyik megvalósulás statisztikai tulajdonságai egybeesnek a folyamat sokfajta megvalósulásának együttes tulajdonságaival, egy adott időpillanatban (n). Vegyük a beszédfolyamat együttes felületét {x(n, l)}, ahol l jelenti a járulékos függéseket speciális dimenziókon, mint például beszédnél a kiejtés beszélőtől való függését. Itt n az időfüggést, l index pedig a beszédfolyamat járulékos függéseit jelenti ezektől a speciális dimenzióktól. Ilyen speciális dimenzióra vonatkozó index például az, amelyiket a beszélők kiejtésének azonosítására használunk. Az X beszédfolyamat együttes felületét lefedő adatbázis létrehozása megkívánja ezeknek a különböző megvalósulásoknak az összegyűjtését (x(n, l)), minden l járulékos függés esetében (Becchetti–Ricotti 1999). Ez szükséges a megfelelően pontos becslés kialakításához. A fentiek alapján látható, hogy a paraméterbecslés pontossága, tehát a felismerés jósága lényegében a betanításhoz használt adatbázis jóságán múlik. Követelmény tehát, hogy az adatbázis elemei helyesen legyenek kiválasztva, egy-egy elemből megfelelő darabszámú reprezentáns legyen, az elemek minősége megfeleljen az előírásoknak és hasonló legyen ahhoz, amivel majd a felismerés során össze kell hasonlítani. Mind a HMM mind az NN alapú felismerés stacionárius folyamatok sorozataként tekinti a beszédet. Viszont a beszédképző szerveink folyamatos mozgással állítják elő a beszédet, a statisztikai tulajdonságok tehát időfüggők. A beszéd modellezhető úgy, mint részeiben stacionárius folyamatok láncolata, és a HMM nagyon jól alkalmazható ilyen folyamatokra. Ez viszont megkívánja a tanító beszédadatbázis szegmentálását olyan hullámformarészekre, amelyeknek hasonlóak a statisztikai tulajdonságaik. Erre mutatunk néhány példát. A nyelvi tartalmat felismerő rendszer (diktálás) betanításához a hanganyagot beszédhang méretű egységekben kell szegmentálni (lásd az 5.2. fejezetben). A prozódiai egységek felismeréséhez hosszabb hangsorelemeket kell kijelölni (8.4. ábra). A mondat modalitását felismerő



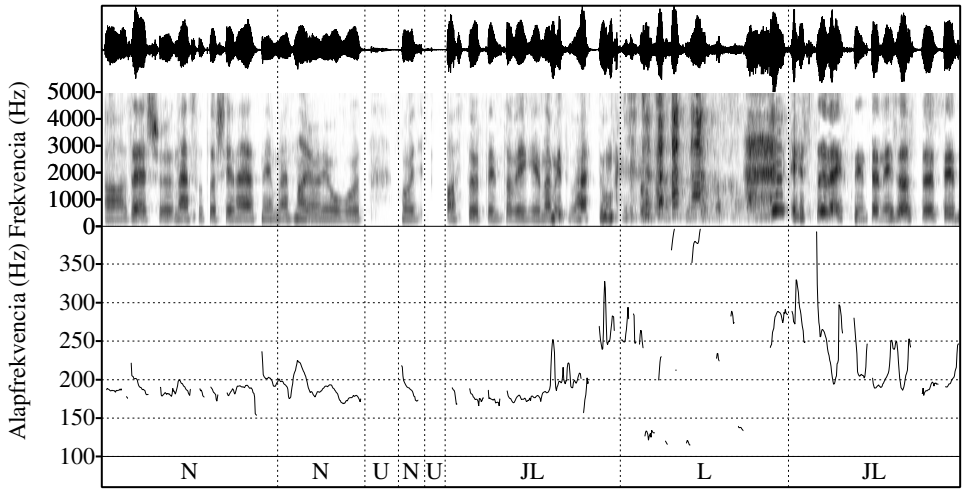
8.4. ábra. Szegmentálás, címkézés prozódiai frázisokra; felül az időfüggvény, alul pedig az alapfrekvencia menete látható. A frázisok határait a függőleges vonalak jelzik. Az U szünetet jelöl

rendszer betanításához mondatnyi, illetve mondatrészegységeket szükséges a tanító beszédadatbázisban bejelölni (8.5. ábra). Az érzelemfelismerő betanításához vi-



8.5. ábra. Mondatszintű szegmentálás és címkézés. Felül az időfüggvény, alul pedig az alapfrekvencia menete látható. Felszólító és felkiáltó mondat: FF, kijelentő mondat: S, kiegészítendő kérdés: K

szont olyan adatbázis kell, ahol a különböző érzelmeket kifejező beszédegységek vannak elkülönítve, ezek határait szükséges a folyamatos beszédbe bejelölni (8.6. ábra). Az ábrán egy telefonbeszélgetés érzelmi beszédrészleteinek határait jelöltük, frázisok bontásában. Az érzelmek bejelölése mellett a hanganyagban a hanggesztusok is bejelölésre kerültek, például a 8.6 ábrán a nevetés, amit nyíllal jelöltünk. Ez a hangszakasz a normál beszédtől abban különbözik, hogy jelllemző rá a nevetési periodicitás, valamint nagyobb a zajtartalma, amint az a spektrogramon látszik, továbbá az alaphang detektáló is emiatt bizonytalanul működik (alul). Összefoglalva azt mondhatjuk, hogy a gépi felismerőket ma még csak egy meghatározott felhasználási területre tudják tervezni. Például olyan felismerő, amely csak egy adott nyelven, telefonon keresztül bemondott számok, szavak felismerésére alkalmas, nem ismer fel mondatokat. Az úgynevezett diktáló rendszerek folyamatos beszédet képesek felismerni megadott nyelven, jól meghatározott témakörön belül, de többnyire csak a felhasználó hangjára működnek elfogadható pontossággal. Ezek a felismerők csak csendes környezetben működnek jól. Azok a felismerők, amelyeket zajtalan környezetre terveztek, nem működnek zajos körülmények között. Az utcazajban működő



8.6. ábra. Érzelem szerinti címkézés. Felül az időfüggvény, középen a spektrum, alul pedig az alapfrekvencia menete látható. Szünet: U, semleges érzelem: N, nevetés: L, nevetve beszél: JL

felismerő rosszul működik, ha személygépkocsiban kívánják használni. A mai gépi felismerők csak azokat a beszéd+zaj mintákat képesek felismerni, amelyeket előzőleg már megtapasztaltak, vagyis amelyekre előzőleg be lettek tanítva. Mindezekből következik, hogy a betanításokat minden esetben célzott beszédadatbázisok segítségével hajtjuk végre, ezért nőtt meg az utóbbi években a jelentőségük. Óriási pénzeket költ ma a világ adatbázisokra. A soknyelvű Európa igen nagy feladat előtt áll, hiszen minden nemzet a saját nyelvén akar bekapcsolódni a nemzetközi kommunikációba, tehát nyelvenként kell sokfajta feladatra alkalmas adatbázisokat létrehozni. Nem biztos, hogy a mai statisztikai megközelítések nyújtják a legmegfelelőbb megoldást a gépi felismerésre, de hogy igen költségesek, az bizonyos. A következőkben csak a szűkebb értelemben vett gépi beszédfelismeréssel, vagyis a beszéd hangsorszintű nyelvi tartalmának a felismerésével foglalkozunk részletesen.

8.1.1. Tanító adatbázisok a nyelvi tartalom gépi felismeréséhez

A beszéd tartalmának gépi meghatározására különböző modelleket használnak. A modellépítésnek számos módja lehetséges, de a mai elterjedt beszéd-szöveg átalakítók mind a rejtett Markov-modelleket használják a felismeréshez (Rabiner 1989, Jelinek 1976). Kétféle modellt alkalmaznak, amelyek egymást segítik a feldolgozás során. Az akusztikai-fonetikai szintet kezelő modellek (akusztikai modellnek is nevezik) magát a beszédhullámot vizsgálják. A nyelvi modellek (a beszédhullámból

kinyert adatokat próbálják statisztikai módszerekkel nyelvi elemekhez rendelni). Ez azt jelenti, hogy a sikeres felismeréshez kétfajta modellt, az akusztikai modellt és a nyelvi modellt együttesen használják. Erre a két modellre kell betanító adatbázisokat készíteni.

A rejtett Markov-modellekkel a beszédfelismerési feladat matematikailag az alábbiak szerint fogalmazható meg: a beszédfelismerőnek azt a legvalószínűbb $\hat{W}^N$ szósorozatot kell meghatároznia, amely az X^T akusztikus jelet adja;

$$\hat{W}^N = \arg\max_{W^n} P(W^n) P\left(X^T | W^n\right), \tag{8.1}$$

ahol W^n egy n elemű szósorozat (mindegyik szerepel a felismerendő szavak listájában), n pozitív egész szám. $P\left(X^T|W^n\right)$-t a nyelvi modellnek kell biztosítania, a $P(W^n)$-t pedig az akusztikai-fonetikai modellnek. A felismerési döntést a két feltétel együttes optimalizálásából kell meghozni: $P(W^n)$ egy szó ($N = 1$), vagy szósorozat elsődleges valószínűsége, ahogy azt a nyelvi modell megadta, és $P\left(X^T|W^n\right)$ a szószekvenciának megfelelő akusztikai jel feltételes valószínűsége, ahogy azt az akusztikai-fonetikai modell megadja.

Mind a nyelvi modell, mind az akusztikai-fonetikai modell nagy mennyiségű, kellően optimalizált és feldolgozott adathalmaz alapján kerül előállításra a rendszer betanítása során. A módszer alapjaiból következik, hogy ha kellően sok minta alapján végezzük a beszédhangjainkat és a beszélt nyelvet leíró modellek előállítását, akkor nagy valószínűséggel pontosak lesznek, azaz lefedik azokat a lehetséges változatokat, amelyeket a tanításhoz használt adatbázisunkba összegyűjtöttünk. Más szavakkal a rejtett Markov-modell és más statisztikai elven működő nyelvi tartalmat felismerő rendszerek jósága azon múlik, hogy milyen jól megválasztott és feldolgozott a beszédadatbázis, amivel az akusztikai fonetikai modellt hozzuk létre (tanítjuk be), és milyen jó az a szövegadatbázis, amivel a nyelvi modellt megalkotjuk. Lényegében véve hasonlóan ahhoz, amit már a statisztikai elvű általános felismerésről elmondtunk.

8.1.1.1. Beszédadatbázisok az akusztikai-fonetikai modell betanításához

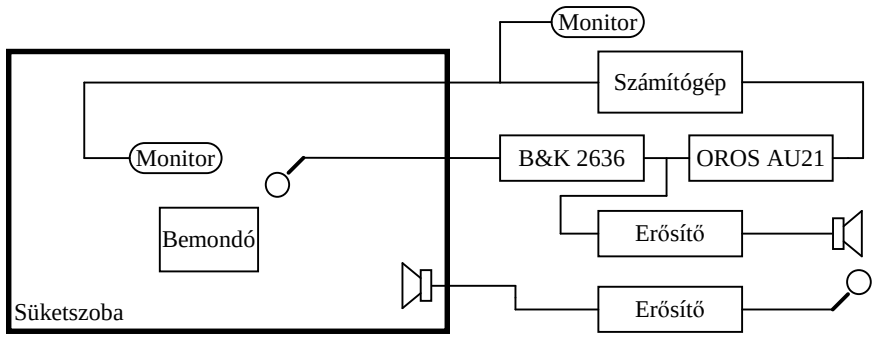
A beszéd nyelvi tartalmát lejegyző gépi rendszerek akusztikai-fonetikai modelljének (a spektrális szerkezetre vonatkozó adatok) betanításához olyan beszédadatbázist kell használni, amelyben az akusztikai jel a nyelvi tartalommal össze van kapcsolva (címkézés). Mivel a beszéd változatosságát okozó tényezők hatással vannak a felismerőt megvalósító rendszerre, olyan beszédanyagot kell összegyűjteni a betanításhoz, amelyben a beszéd a nyelvi tartalommal együtt megfelelő mértékben tükrözi az adott felhasználáskor előforduló változatosságát. Tervezéskor tehát tudni kell, hogy milyen típusú felismerőt szándékozunk létrehozni. Egyfajta osztályozásra ad

példát a 8.3. táblázat. A gyakorlati követelmények jó kiszolgálásához tehát a tanító

8.3. táblázat. Osztályozás a beszéd nyelvi tartalmának gépi felismeréséhez

Beszéd	Beszélőfüggő, -független, beszélőadaptív
Beszédegység	Beszédhang, hangkapcsolat, szó, szókapcsolat
Beszédmód	Izolált szavas, kapcsolt szavas, folyamatos felismerés
Beszédstílus	Olvasott beszéd, előadás, spontán beszéd
Szótárméret	Felismerendő elemek száma
Akusztikus környezet	Csendes helyiség, zajos környezet, telefonvonal
Nem nyelvi elemek	Nyelés, köhögés, hezitálás

beszédadatbázisok sokaságával és sokrétűségével kell számolni. Más adatbázis kell egy beszélőfüggetlen és más egy beszélőfüggő rendszer betanításához. A beszélőfüggő rendszernél elegendő az adott használó személytől hanganyagot gyűjteni, és azzal tanítani a rendszert. Viszont egy beszélőfüggetlen rendszer betanításához nagyszámú beszélő hanganyagával kell tanítani úgy, hogy a használók hangjának változatosságát az összegyűjtött hanganyag lefedje. Amennyiben a használók dialektussal is beszélhetnek, akkor az adatbázisba a használandó dialektusokból is kell mintákat gyűjteni. Amennyiben szépen olvasott szöveget kell felismerni, az optimális felismeréshez szépen olvasott hanganyaggal kell a tanítást elvégezni, spontán beszéd felismeréséhez spontán hanganyaggal kell tanítani. Más típusú beszédadatbázis kell egy izolált szavas (esetleg kapcsoltszavas) felismerő betanításához, és más a folyamatos beszédfelismerőhöz. Az izolált szavas felismerésnél elegendő a használt szavak (szókapcsolatok) megfelelő számú hanganyaga. A folyamatos beszéd felismerésénél az adott nyelv beszédhangjainak és hangkapcsolatainak kell a megfelelő számban a betanító beszédadatbázisban előfordulni. Az adott feladathoz legjobban illeszkedő adatbázis kiválasztásánál (készítésénél) az alábbi szempontokat kell figyelembe venni: a hangfelvételek és a rögzítés pontos fizikai leírását, a felvett anyag nyelvi jellemzőit, az adatbázis méretét, a beszélők szociolingvisztikai adatait, az adatbázis feldolgozási módját. A hangfelvétel pontos leírásának tartalmazni kell a mikrofonok számát, elhelyezését típusát, a környezetet (stúdió, iroda, gépkocsi), a felvétel ellenőrzési módszerét (kézi, gépi), a mintavételi és kvantálási paramétereket. Példaképpen bemutatjuk a BABEL magyar beszédadatbázis felvételi mérési összeállítását (8.7. ábra). A beszédadatbázis méretét a beszélők száma szerint osztályozhatjuk. Kevés beszélővel (1–5 fő) készített adatbázisok például a beszédszintézis fejlesztési céljait szolgálhatják. Több beszélővel készített adatbázisok a beszéd gépi felismerésénél használt akusztikai modellparaméterek becslésére szolgálnak. Ezekben az adatbázisokban éppen ezért a hangkapcsolatok változatossága nagy. Általában csendes helyiségekben történik a felvétel. A beszélők száma kisebb, mint 50. Sok beszélővel készülnek azok a beszédadatbázisok, amelyek a beszélőfüggetlen felismerő algoritmusok betanítására szolgálnak. A beszédstílus és a rögzítési körülmények nagy változatossága szük-



8.7. ábra. A BABEL beszédadatbázis süketszobai felvételeinek mérési összeállítása

séges. Szociolingvisztikai jellemzőket is rögzíteni kell a beszélőkről (férfiak, nők, dohányoznak, nem dohányoznak, anyanyelvükön történik-e a bemondás, nyelvjárás, dialektus van-e rögzítve az adatbázisban, milyen a koreloszlás a bemondók között).

Gyakorlati adatbázisok gépi beszédfelismeréshez. Európában az ESPRIT elnevezésű (Speech Assessment Methods, SAM) projekt keretében szabványos, a beszédre és a nyelvre vonatkozó rögzítési és feldolgozási eljárásokat dolgoztak ki és fogadtak el (Fourcin–Dolmazon 1991). Ezekben a szabványokban a rögzítési eljárásokra, a tárolásokra, az annotálási technikákra, az adatok eloszlására vonatkozó előírások szerepelnek. E projekt keretében hozták létre a fonémarealizációkat meghatározó SAMPA fonetikus beszédhang-szimbólumrendszert. Az itt rögzített szabványeljárások alapján jöttek létre az EUROM 0, EUROM 1, BABEL adatbázisok (Chan–Fourcin 1995), Európa legtöbb nyelvét átfogva. Ezek csendesszobai felvételeket tartalmaznak. A beszélők tervezett hangkapcsolatokat, szavakat, mondatokat, számokat olvastak fel.

A SPEECHDAT 1, 2 és a SPEECHDAT-E adatbázisok vezetékes és mobiltelefonon keresztül rögzített beszédet tartalmaznak (Siemund et al. 2000), szintén Európa számos nyelvére.

A SQUALE projekt keretén belül (`http://www.squale.org/`) olyan adatbázist hoztak létre a kutatók, amely a beszédfelismerők értékelésére használható. Néhány beszédadatbázis jellemző adatait a 8.4. és a 8.5. táblázatban mutatjuk be. Amerikában a TIMIT és az ATIS a legjelentősebb adatbázis, amit gépi beszédfelismerés céljára hoztak létre. Ezek akusztikai modellek felépítésére alkalmasak amerikai angol nyelvre.

A TIMIT adatbázis személyfüggetlen fonetikai beszédfelismerők betanítására és tesztelésére szolgál. Szómodellek felépítésére alkalmatlan, mivel szűkített szótárkészletet használ, fonetikailag gazdag mondatai viszont kiválóan alkalmasak akusz-

8.4. táblázat. Kész beszédadatbázisok gyakorlati jellemzői

Az adatbázis neve	Forrás	Mintavétel	Rögzítési környezet	Bemondás módja	Feldolgozás alapegysége	Átírás
TI Digits	Mikrofon	20 kHz	Csendes szoba	Felolvasás	Szó	Nincs
TIMIT	Mikrofon	16 kHz	Csendes szoba	Felolvasás	Beszédhang	Van
NTIMIT	Telefon	8 kHz	Telefonon keresztül telefonfülke, iroda, lakás, utca stb.	Felolvasás	Beszédhang	Van
ATISO	Mikrofon	16 kHz	Hivatal	Felolvasás	Mondat	Nincs
Switchboard (Credit Card)	Telefon	8 kHz	Telefonon keresztül telefonfülke, iroda, lakás, utca stb.	Spontán beszéd	Szó	Van
Switchboard (Credit Card)	Telefon	8 kHz	Telefonon keresztül telefonfülke, iroda, lakás, utca stb.	Spontán beszéd	Szó	Van
MARSEC	Mikrofon	16 kHz	Változó	Spontán	Beszédhang	Van
ATIS2	Mikrofon	16 kHz	Hivatal	Spontán	Mondat	Nincs
EUROM 1 BABEL	Mikrofon	20 kHz	Csendes szoba	Felolvasás	Beszédhang, szó	Van
SpeechDatSpeech-Dat-E	Telefon	8 kHz	Vezetékes telefon, mobiltelefon, iroda, lakás, utca stb.	Felolvasás, spontán	Beszédhang, szó	Nincs

8.5. táblázat. Kész beszédadatbázisok mennyiségi jellemzői

Adatbázis neve	CD száma	Felvételi idő (óra)	Méret gigabájt	Beszélők száma	Egységek száma
TI Digits	3	~14	2	630	>2500 szám
TIMIT	1	5,3	0,65	630	6300 mondat
NTIMIT	2	5,3	0,65	144	6300 mondat
ATISO	6	20,2	2,38	69	10 722 kiejtés
Switchboard (Credit Card)	1	3,8	0,23	16	35 dialógus
Switchboard (Credit Card)	30	250	15	100	2500 dialógus
MARSEC	1	5,5	0,62	351	53 mondat
ATIS2	6	~37	~5	>124	12 000 kiejtés
EUROM 1 BABEL	3-5	Nyelvfüggő	Nyelvfüggő	100	Számok, fonetikailag kiegyensúlyozott mondatok, hangkapcsolatok, szavak
SpeechDat-SpeechDat-E	4-6	Nyelvfüggő	Nyelvfüggő	Nyelvfüggő 500-5000	Számok, nevek, intézmények, utasítások, fonetikailag gazdag mondatok

tikai (beszédhang) modellek létrehozására. Az adatbázis egy része betanításra, másik része tesztelésre ad lehetőséget.

ATIS (Air Travel Information System) repülőtéri információval kapcsolatos szótárkészleten alapuló adatbázis. Az adatbázisban minden elem spontán társalgásban és olvasva is rögzítésre került hivatali körülmények között.

Magyar nyelvű beszédadatbázisok gépi beszédfelismeréshez. Az utóbbi években sok magyar beszédadatbázis készült el a gépi beszédfelismerés támogatására. Ezek közül a jelentősebbek összefoglaló adatait a 8.6. táblázatban adjuk meg.

BABEL – magyar nyelvű beszédadatbázis.

`http://alpha.tmit.bme.hu/speech/hdbbabel.php`

Ez volt az első magyar beszédadatbázis, amelyik egy többnyelvű közép- és kelet-európai beszédadatbázist létrehozó munkaprogram keretében készült el 1995 és 1998 között. Célja egy közös, egységes elvek alapján felépített, nagy méretű beszédadatbázis létrehozása volt, a beszédakusztikával, fonetikával, digitális jelfeldolgozással, valamint nyelvészettel foglalkozó szakemberek munkájának segítésére (Vicsi–Vig 1995). A BABEL program keretén belül 5 közép- és kelet-európai nyelv beszédadatbázisa készült el, ezek a bolgár, az észt, a magyar, a lengyel és a román. A magyar BABEL adatbázis a hivatalos magyar köznyelvet reprezentáló rendezett hanganyag. Rögzítéskor hangkapcsolatokat, szavakat, számokat, folyamatos szövegben 5 mondatos bekezdések sorozatát olvasták fel, lefedve a magyar beszédhangokat és a leggyakoribb hangkapcsolatokat. Az adatbázis erősen zajcsökkentett környezetben felvett olvasott szöveget tartalmaz, ez az úgynevezett tiszta olvasott beszéd, melyet 60 személlyel, 30 férfivel és 30 nővel vettek fel, kor és foglalkozás szerint széles eloszlásban. A teljes hanganyag 1,8 GB terjedelmű, amelyet 3 CD-n rögzítettek.

SpeechDat-E – magyar telefonbeszéd-adatbázis.

`http://alpha.tmit.bme.hu/speech/hdbspeechdt.php`

Az adatbázis 1000 beszélő által telefonon bemondott, előre megadott szövegből áll. 1999 és 2000 között készült. Különböző típusú telefonos beszédfelismerők betanítására és tesztelésére ad lehetőséget. Ezek lehetnek izolált szavas felismerők, szókeresők és szóazonosítók, dialógusrendszerek, valamint szótárfüggetlen megoldások, amelyeknél a felismerés szónál kisebb felismerési egységek modellezésén alapul. Az összeállított szöveganyag a fenntiekben leírt sokfeladatos elvárásoknak megfelelően igen sokrétű, tehát tartalmaz: parancsszavakat, számjegysorozatokat, telefonszámot, hitelkártyaszámot, PIN-kódot, spontán dátumot, relatív dátumot, parancsszavas kifejezést, számjegyet, betűzött spontán vezetéknevet, betűzött városnevet, betűzött szót, pénzmennyiséget (forint/euró), természetes számot, vezetéknevet, városnevet, cégnevet, vezeték+keresztnevet egyben, eldöntendő kérdést igen/nem válasszal, valamint folyamatos szöveget, amely fonetikailag gazdag mondatokból áll, lefedve a magyar beszédhangokat és a leggyakoribb hangkapcsolatokat (Pollak et al. 2000). A telefon-beszédadatbázis specifikációja az MLAP LRE-63343 SPEECHDAT (M)

8.6. táblázat. Magyar beszédadatbázisok gépi beszédfelismeréshez

Az adatbázis neve	Forrás	Formátum	Rögzítési környezet	Bemondás	Bemondók száma
	Szövegtípus				
	Feldolgozás				
BABEL	Mikrofon	20 kHz, 16 bit	Süketszoba (tiszta beszéd)	Olvasott szöveg	60
	Hangkapcsolatok, számok, szavak, folyamatos szöveg, amely lefedi a magyar beszédhangokat és a leggyakoribb hangkapcsolatokat				
	Fonotipikus átírás, az anyag 10%-ában beszédhangszintű szegmentálás, címkézés				
SpeechDat-E	Vonalas telefon	8 kHz, 16 bit (ISDN)	Iroda, lakás, utca, telefonfülke stb.	80% olvasott, 20% spontán szöveg	1000
	Betűzött szavak, dátumok, pénzösszegek, számok, telefon- és hitelkártyaszámok, szavak, tulajdonnevek, folyamatos szöveg, amely lefedi a magyar beszédhangokat és a leggyakoribb hangkapcsolatokat				
	Leírás karakterekkel, nincs szegmentálás, zajok, hibák jelölése				
MTBA	Vonalas telefon, mobiltelefon	8 kHz, 16 bit	Iroda, lakás, utca, telefonfülke stb.	80% olvasott, 20% spontán szöveg	500
	Betűzött szavak, dátumok, pénzösszegek, számok, telefon- és hitelkártyaszámok, szavak, tulajdonnevek, folyamatos szöveg, amely lefedi a magyar beszédhangokat és a leggyakoribb hangkapcsolatokat				
	Szavak leírása karakterekkel, zajok, hibák jelölése. Folyamatos szöveg beszédhangszintű szegmentálása, címkézés				
MRBA	Mikrofon, hangkártya (PC hangbemenet)	16 kHz, 16 bit	Iroda, lakás,	Olvasott szöveg	332
	Folyamatos szöveg, amely lefedi a magyar beszédhangokat és a leggyakoribb hangkapcsolatokat				
	Az anyag 66%-a karakterekkel leírt, zajok, hibák jelölése. Az anyag 33%-ának beszédhangszintű szegmentálása és címkézése				
SPECO	Mikrofon, hangkártya (PC hangbemenet)	20 050 Hz, 16 bit	Süketszoba	Olvasott, utánmondott szöveg	76
	Kitartott beszédhangok, hangkapcsolatok, számok, szavak, mondatok				
	Fonotipikus átírás, beszédhangszintű szegmentálás				
BEA	Mikrofon	20 kHz, 16 bit	Csendes helyiség	Spontán beszéd, utánmondás, társalgás	500
	Folyamatos beszéd				
	Fonotipikus átírás				

EU projekt javaslata alapján készült. Ez biztosítja azt, hogy a különböző nyelvű adatbázisok igen hasonlóak, és egységes alapot képviselve ugyanazt a beszédtechnológia fejlesztési lehetőséget nyújtják a feldolgozott nyelvhez.

MTBA – magyar telefonbeszéd-adatbázis.

`http://alpha.tmit.bme.hu/speech/hdbMTBA.php`

Az adatbázis 500 beszélő hanganyagát tartalmazza, ezekből 297 vezetékes, 203 pedig mobiltelefon-felvétel. A telefon-beszédadatbázis specifikációja az MLAP LRE-63343 SPEECHDAT (M) EU projekt javaslata alapján készült (Vicsi et al. 2002). Az adatbázis a statisztikai feldolgozási módszereken alapuló, telefonon keresztül működő beszédfelismerő rendszerek betanítására és tesztelésére ad lehetőséget (lásd a SpeechDat-E-nél). Mindezeken felül általános fonetikai, nyelvészeti kutatásokhoz is használható. Az adatbázis két részből áll.

Az első rész parancsszavakat, különböző számokat (telefonszám, hitelkártyaszám, dátum stb.), tulajdonneveket (személy-, cég- és városnevek) tartalmaz. Ebben a részben van rögzítve továbbá a bemondott szöveg magyar betűkkel történő leírása, a felvétel jellemző adatainak a rögzítése és a hibás ejtések, valamint a felvételben hallható zajok jelölései.

A második rész folyamatos szöveget tartalmaz, ami fonetikailag gazdag mondatokból áll. Statisztikailag kiegyensúlyozott anyag a beszédhangok, a kettős- és hármas hangkapcsolódások szintjén, hogy a felolvasandó szöveg kellően változatos legyen és lefedje a magyar nyelv sajátosságait. Az anyagban a magyar nyelvben ritkán előforduló beszédhangok is kellő számban fordulnak elő. Összesen 1992 mondatból és 2000 szóból áll a szöveg. Az így elkészült adatbázisban a magyar nyelv leggyakoribb kettős hangkapcsolatainak a 98,8%-a le van fedve. Ez a rész a fent említett szabvány előírásain túllépve beszédhangszintű szegmentálást és címkézést is tartalmaz.

TESZTEL – telefonbeszéd tesztadatbázis.

`http://alpha.tmit.bme.hu/speech/hdbtesztelen.php`

A TESZTEL egy magyar nyelvű, mobiltelefonon keresztül, erősen zajos környezetben (utcán, metróban, városi közlekedésű buszon) rögzített telefonos beszédadatbázis, amely 2003-ban készült. Az adatbázisba 100 beszélő hangfelvétele került rögzítésre, előre meghatározott szavakat és mondatokat olvastak fel (az MTBA-hoz hasonlóan). Ebben az adatbázisben megadták a felvételek zajossági fokát, azaz a telefonos felvételek jel-zaj viszony adatait. Ez a paraméter függ egyrészt a környezeti zaj nagyságától és milyenségétől, másrészt az átviteli csatorna által indukált zajoktól is. A kapott értékek eléggé széles skálán mozogtak, a nagyon rossznak tekinthető 5 dB körüli jel-zaj viszonytól (például csúcsidőben, forgalmas főutak mellett) a 20–25 dB-ig (késő esti órákban, utcán vagy tömegközlekedési eszközön).

MRBA – magyar referencia beszédadatbázis.

`http://alpha.tmit.bme.hu/speech/hdbMRBA.php`

Ez egy általános célú, otthoni/irodai környezetben felolvasott folyamatos szöveget

tartalmazó adatbázis, melyet a BME TMIT Beszédakusztikai Laboratóriuma a Szegedi Tudományegyetem (SZTE) Informatikai Tanszékcsoportjával együttműködve hozott létre 2004-ben (Vicsi et al. 2005). A létrehozott adatbázis különböző típusú beszédfelismerők betanítására és tesztelésére ad lehetőséget. A készítők alapos statisztikai vizsgálatokat végeztek a beszédhangok, a kettős és hármas hangkapcsolódások szintjén, hogy a felolvasandó szöveg kellően változatos legyen és lefedje a magyar nyelv sajátosságait. Az így elkészült adatbázisban a magyar nyelv leggyakoribb kettős hangkapcsolatainak a 98,8%-a le van fedve. 300 beszélő 12 különböző mondatot és 12 különböző, a mondatoktól független szót olvasott fel. A nyelvi változatosság lefedésére négy különböző tájegységben levő városban (Budapest, Szeged, Győr és Miskolc) is készültek hangfelvételek. Az életkor és a nemek szerinti változatosságot a 8.7. táblázat mutatja. Az adatbázis teljes anyaga annotált, az adatbázis

8.7. táblázat. Az MRBA beszédadatbázis adatközlőinek életkor és nem szerinti eloszlása

Beszélő	<16 év	16–30 év	31–45 év	46–60 év	>60 év	Összesen
Férfi	0,9%	46,1%	5,7%	3,9%	0,9%	57,5%
Nő	3,3%	27,7%	6,0%	5,1%	0,4%	42,5%

harmada (100 beszélő) beszédhangszinten kézileg szegmentált és címkézett.

SPECO – gyermekbeszéd-adatbázis.

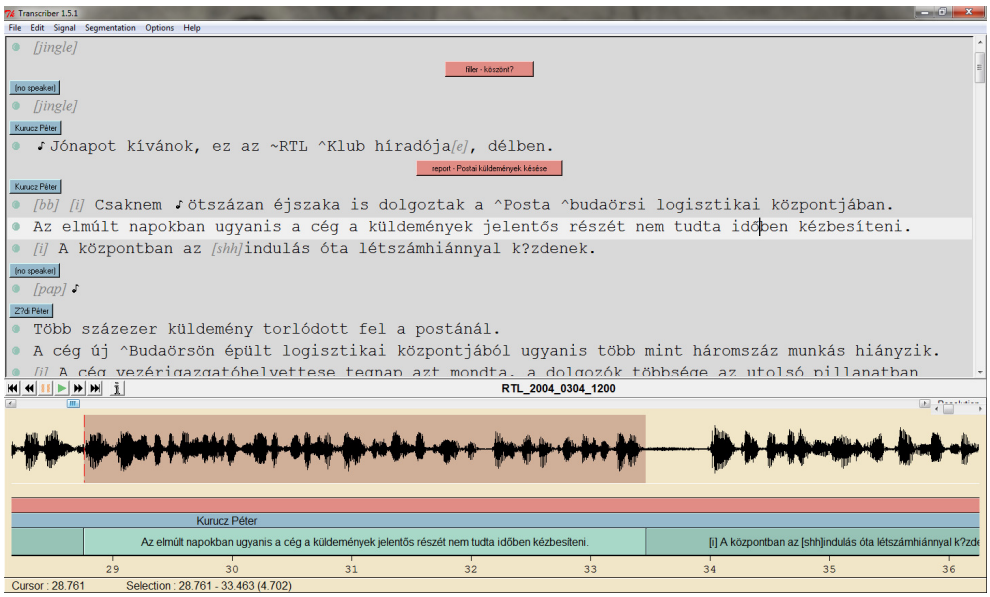
`http://alpha.tmit.bme.hu/speech/paperc013.php`

Az adatbázis csendes helyiségben 5–10 éves gyermekek által bemondott, előre meghatározott szótagokat, szavakat, mondatokat tartalmaz. A beszédadatbázis 1999-ben készült (Csatári et al. 1999). A fonetikai, beszédfelismerési kutatásokhoz biztosít megfelelő hanganyagot (hangkörnyezet, hanghelyzet, különböző szupraszegmentális jegyek stb. vizsgálata). Az adatbázis nagy részében átlagos köznyelvi olvasott gyermekbeszéd van rögzítve csendes körülmények között. Kisebb részben pösze és különböző súlyosságú hallássérült gyermekek beszédét is tartalmazza.

Híranyag-adatbázis.

A magyar nyelvű híranyag-adatbázis gyűjtése és egységes elvű feldolgozása egy nemzetközi munkacsoport, a COST278 EU projekt keretein belül jött létre (Zibert et al. 2005). Az adatbázis 3 óra 30 percnyi híranyagot tartalmaz, melyet közszolgálati és kereskedelmi adók műsoraiból vettek fel. A felvételek egy személyi számítógéphez csatlakoztatott televíziós készülék segítségével készültek (mind a hanganyagot, mind a képi anyagot is rögzítették). Az adatbázis akusztikai és nyelvi feldolgozása a hanganyag átírása, címkézése a LDC (Linguistic Data Consortium) idevonatkozó ajánlásai alapján készült. Az adatbázisban a beszélőváltások során fellépő akusztikai változásokat, a beszélő által elmondott szöveg határait, a híradások szekcióit, a híradások szignáljainak kezdetét és végét, idegen nyelvű beszédet, háttérzajt és a beszélő által keltett zajokat a címkézés mutatja. A stúdióban elhangzott beszélgetéseket, a stúdióból kommentált riportokat, illetve a műsorvezető beszéde során elhangzó hanganyagot külön címke jelöli.

Az adatbázist a többszintű, Transcriber nevű program segítségével címkézték. A program kezelői felülete a 8.8. ábrán látható. Külön sávok teszik alkalmassá a különböző események szinkronbejegyzését (Teleki–Vicsi 2006). A programról részletesen a `http://www.etca.fr/CTA/gip/Projets/Transcriber/` honlapról szerezhető be információ.

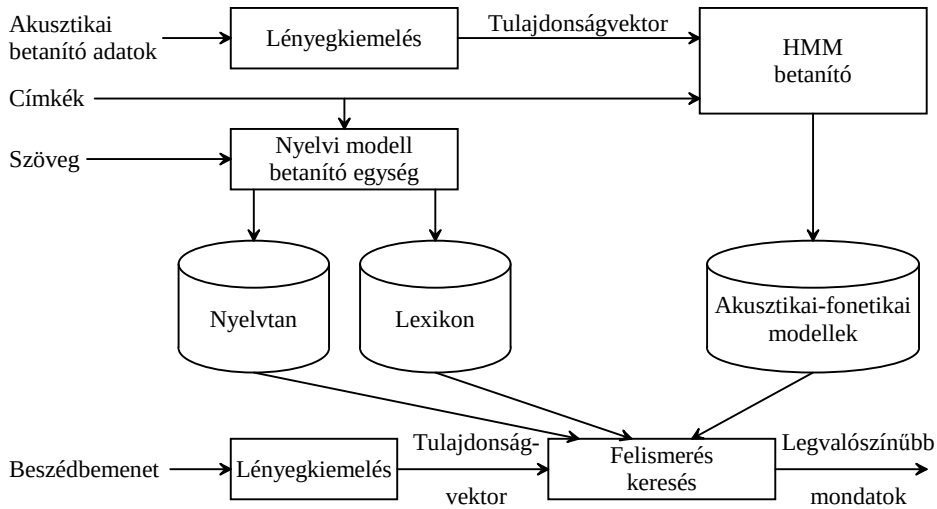


8.8. ábra. A Transcriber program kezelőfelülete

8.1.1.2. Szövegadatbázisok a nyelvi modell betanításához

A nyelvi modell feladata a gépi beszédfelismerésben az, hogy segítse az akusztikai modell kimenetén megjelenő adatsorozatot (beszédhangok vélt sorát) értelmes nyelvi tartalomhoz rendelni (például szavak, szókapcsolatok gyakoriságának definiálásával). A hozzárendelést statisztikai módszerek segítik. A nyelvi modell tanítására szöveges, írott adatbázisokat használnak. A szövegadatbázis, más néven szövegkorpusz, ténylegesen előforduló írott vagy lejegyzett beszélt nyelvi adatok gyűjteménye (például a parlamenti beszédek szöveganyaga). Az adatbázisba bekerülő szövegeket valamilyen szempont szerint válogatják és rendezik. Az ilyen adatbázis nem feltétlenül összefüggő szövegeket tartalmaz, lehetnek benne szövegtöredékek is. A szövegadatbázis nem csak tárháza a szövegeknek, hanem tartalmazza azok belső adatait is. A készítők jelölik a szerkezeti egységek határait (szó, bekezdés, mondat) és egyéb információkat is (például szófaj), a mondatokat

normalizált grafémikus formába hozzák és a fonetikai átiratot is elkészítik (lásd 4.4. fejezet). A nyelvi modelleket úgynevezett betanítóval lehet létrehozni szövegadatbázisokból. A betanító két fájlt hoz létre, a Lexikon és a Nyelvtan fájlokat. E fájlokat felismerő modulként használják a nyelvi tartalomfelismerő rendszerek, amint azt a 8.9. ábra mutatja. A nyelvi modellt olyan mondatkészlettel kell betanítani,



8.9. ábra. Nyelvi modell betanító blokkdiagramja

amelyben a mondatok jellemzően hasonlítanak azokra a mondatokra, amelyeket fel kell majd ismerni az adott alkalmazásban. A Lexikon file olyan szókészletet tartalmaz (a szó fonetikai átiratával együtt), amelyet a nyelvi betanító talált a betanító készletben. A szókészlet alkotja a felismerő rendszer úgynevezett szókincsét. A felismerő csupán a Lexikonban tárolt szavakat képes felismerni. A Nyelvtan file szósorrendi valószínűségeket tartalmaz, vagyis hogy egy szó a megelőző szavak után milyen valószínűséggel jelent meg a betanító készletben (ha statisztikai, szóalapú n-gram nyelvtanról van szó). Ha csak egy szó előfordulási valószínűségét vesszük figyelembe, akkor a nyelvi modellt „unigram”-nak nevezzük, ha szópárokét, akkor „bigram”-nak. A nyelvi modell betanításához a beszédadatbázisok lejegyzett nyelvi tartalmát szokás használni, de ezek rendszerint nem elegendőek, legfeljebb igen szűk nyelvtani keretek és szótárkészlet esetében használhatóak. Nagyszótáras felismerő rendszereknél mindig szükség van nagy szövegadatbázisok felhasználására is. Ilyen szövegadatbázisok például angol nyelven a British National Corpus, Penn Treebank Corpus, a Brown Corpus. Ezek a szövegkorpuszok XML-formátumban, az úgynevezett Corpus Encoding Standard for XML (`http://www.xces.org/`) szabvány szerint vannak szerkesztve. Ma már hatalmas méretű szövegadatbázisok állíthatók elő az internetről (például a különböző nyelvű Wikipedia-változatok

stb). Ez azt jelenti, hogy ma már a nyelvi modellek betanításához használt szövegadatbázisok függetlenek az akusztikai-fonetikai modellek betanításához használt beszédadatbázisoktól. A szavak egymásutániságára való valószínűségi becslés annál jobb, minél nagyobb méretű és változatosabb a betanításra használt szöveg. A becslés pontatlan lehet, ha a betanító készlet mérete korlátozott. A betanításra használt szövegadatbázis különböző lehet méretben, komplexitásban, a felismerő felhasználási területének témaköre szerint. Egy adott témához kötött keresőfeladatok ellátására, betanításkor a keresőfeladatra orientált szűkebb szöveganyag elegendő. Az ATIS rendszerekben (lásd például a 8. fejezetet) repülési keresőfeladatokra orientált nagy menyiségű szöveganyagot használtak fel a nyelvi modellek betanítására. Híranyag vagy parlamenti beszéd felismeréséhez már jóval nagyobb méretű, összetettebb szöveganyagra van szükség (például angol nyelvre 40 millió szavas szövegre). Ilyenkor a lexikonméret angol nyelvre már 10 000 és 100 000 elem közötti. A betanító szöveganyag összeállításánál arra kell törekedni, hogy a nyelvi modell betanító által létrehozott lexikon minél jobban közelítse a használatban elvárt szókincsméretet. A lexikon és az elvárt szókincs eltérésének minimalizálása elsődlegesen járul hozzá a felismerési rendszer globális hibaarányának minimalizálásához. A lexikális változatosság, a lexikon és az elvárt szókincs lefedettsége azonos betanító szövegméret esetén nyelvenként erősen változik. Angol nyelven lehet a legrövidebb szöveggel a maximális lefedettséget elérni. Lényegesen rosszabb a lexikális lefedettség a ragozó nyelveknél (olasz, francia), vagy pedig az összetett szavas nyelveknél (német) (Lamel–DeMori 1995), a legrosszabb a helyzet az erősen ragozott, szabad szórendű nyelveknél (magyar, finn), mivel ezeknél a lehetséges előforduló szóvariációk és szósorrendek száma a végtelenhez közelít. A nagy méretű szöveganyagokat betanításra alkalmassá kell tenni. Elsődleges feladat a szövegtisztítás. A szövegadatok különböző típusú hibákat tartalmaznak. Néhány ezek közül: tipográfiai hiba, mint amilyen az elgépelés vagy a hiányzó szóköz, egyebek pedig elsődlegesen a szövegfeldolgozás közben kerülnek be az anyagba. Néhány normalizálási lépés: a szövegfeldolgozásnál alapnak tekinthető, például a hangsúly és egyéb mellékjelek kódolása az ISO-Latin1-ben (ISO-8859-1 1998); a szöveg cikkekre, bekezdésekre és mondatokra való szeparálása; a gyakori formázási és központozási hibák korrekciója; és az egyértelmű írásjelek kidolgozása. Ezeket a normalizációs lépéseket egy alapszövegforma létrehozásához vezették be. További normalizálási eljárásokról van Eynde–Gibbon (2000) munkájában olvashatunk.

Magyar nyelvű szövegadatbázisok (korpuszok). Míg angolra 60 000 szavas lexikonnal egy általános beszédfelismeresi alkalmazás jól megoldható, magyarra hasonló lefedettséghez akár milliónál is több szót tartalmazó lexikonra lenne szükség. A nagy probléma igazán a szókapcsolatok modellezésénél jelentkezik, például egy tipikus tri-gram nyelvi modellezési megoldásnál, amikor két szó alapján következtetünk a harmadik valószínűségére. Ilyen esetben már terabájtos memóriára

volna szükség. Éppen ezért, magyar nyelvre egyenlőre csak szűk tématerületű, behatárolt nyelvtanú szövegek felismerésére születtek megoldások. Ilyenek például a meghatározott tématerületű orvosi diktálórendszerek, amelyek betanításához az adott vizsgálati területhez tartozó (hasi ultrahang, endoszkópia stb.) többezer orvosi leletből kialakított szövegadatbázist használták, vagy az időjárás-lekérdező rendszer, ahol pedig nagy menyiségű lejegyzett időjárás-jelentést használtak a nyelvi modellek betanítására. Az ez ideig publikussá vált általános magyar nyelvű szövegadatbázisok közül felsorolunk néhányat.

A Magyar Nemzeti Szövegtár (MNSZ)

`http://corpus.nytud.hu/mnsz/`

Az első olyan magyar nyelvi szövegkorpusz, amely a magyarországiak mellett a határon túli magyar nyelvváltozatokat is felöleli. Lehetőségeihez mérten reprezentatívan tartalmazza a mai magyar nyelv jellegzetes megnyilvánulásait. Az MNSZ jelenleg 187,6 millió szövegszót tartalmaz. Egyrészt öt regionális nyelvváltozatra oszlik, másrészt ezen belül öt stílusrétegből tartalmaz szövegeket. Az aktuálisan vizsgálandó alkorpuszt ezek tetszőleges variációjaként lehet kiválasztani. A határon túli nyelvváltozatokkal kiegészülve a szövegtár tehát alkalmassá vált nemcsak stílusrétegek, hanem nyelvváltozatok szerinti összehasonlító vizsgálatok elvégzésére is.

Szeged Korpusz és Treebank

`http://www.inf.u-szeged.hu/projectdirs/hlt/`

Ez idáig ez a legnagyobb kézzel annotált magyar természetes nyelvi korpusz. 1,2 millió szövegszót tartalmaz, amely 155 500 különböző szóalakot fed le, és további 250 ezer írásjelet is magában foglal. Szófajilag egyértelműsített magyar természetes nyelvi adatbázis, illetve magyar természetes nyelvi adatbázis teljes szintaktikai elemzéssel.

Magyar webadatbázis

`http://mokk.bme.hu/resources/webcorpus`

Több mint 1,48 milliárd szóval (szűretlenül, illetve 589 millió megszűrt szóval) ez jelenleg a legnagyobb magyar nyelvű korpusz, mely teljes méretében elérhető Open Content licenc alatt. A 18 millió oldal letöltésével létrejött adatbázis az irott magyar nyelvet reprezentálja. Több jelentősebb szövegadatbázis az alábbi weboldalon található: `http://hlt-platform.hu/offline-adatbazisok.html`

8.2. Beszédből készített elembázisok beszédszintézishez

Olaszy Gábor

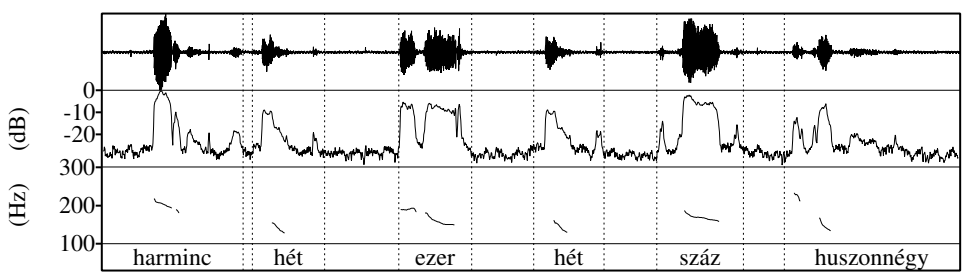
A beszéd gépi előállításához is szükség van valamilyen beszédadatbázisra, ez azonban másfajta belső szerkezettel rendelkezik, mint amiket beszédfelismerők tanítására készítenek. Lehetnek olyanok, amelyekben csak hullámformarészleteket tárolnak

(hangok, hangkapcsolatok, szórészek), de lehetnek olyanok is, amelyek folyamatos beszédet, mondatokat tartalmaznak. Az előbbi kategóriába tartozókat hangelembázisnak, az utóbbiakat beszédadatbázisnak nevezzük ebben a fejezetben. A szintézis célú adatbázisok megjelenése kissé régebbre nyúlik vissza, mint a beszédfelismerőkhöz készítetteké. Magyarországon az 1980-as évek végén már szerkesztettek hangelembázist kísérleti beszédelőállító rendszerekhez (Olaszy et al. 1986). A szintézist támogató beszédadatbázisok szerkezetét, tulajdonságaikat – hasonlóan a gépi beszédfelismeréshez – az alkalmazott módszer határozza meg. A szabályalapú beszédszintetizátorok kisebb méretű ilyen hangelembázisokból építkezhetnek, a statisztikai alapúak működéséhez nagyobb méretű beszédadatbázisokat használnak, hasonlóakat, mint a beszédfelismerők tanításához. Az adatbázis tartalmi része tehát minden esetben az alkalmazott beszédépítési technológiához kötődik. A nagy méretű beszédadatbázisok felhasználási formája kétirányú lehet: közvetlen, illetve közvetett. Az előbbinél közvetlenül a tárolt hullámformát használjuk a hangelőállításhoz, az utóbbinál pedig az adatbázis hullámformájából csak paramétereket nyerünk ki és azokat használjuk fel szintézishez (10.3.8. fejezet). A szintézishez készített adatbázisoknál kevésbé fontos a beszéd interindividuális változatosságának kérdésköre, mivel itt sokszor csak az a cél, hogy egyetlen hangon, jó minőségben szólaljon meg a szintetizátor. Az intraindividuális variáltsági tényezők is leszűkülnek (például megfázott bemondóval nem készítünk hangfelvételt). Ezek az adatbázisok a specifikus kategóriába tartoznak, hiszen nem általános felhasználásra, hanem egy meghatározott célkitűzés támogatására készülnek. Az ilyen adatbázisok egyik fontos szerkezeti jellemzője lehet, hogy a géppel előállított hang mennyire fogja hordozni a bemondó személy hangszínezetét (ez fontos lehet adott megrendelők számára, hogy például a cég hangján szólaljon meg a gépi hang is). Minél jobban meg akarjuk közelíteni egy adott személy hangszínezetét, annál nagyobb beszédadatbázisra van szükség. Hangelembázissal csak speciális tervezéssel és szűk tématerületre lehet személyhez kötött hangot előállítani (például számok felolvasására). A beszédépítéshez használt adatbázisok hullámformaelemeit a legtöbb alkalmazáshoz el kell látni zöngeszinkroncímkékkel (minden hangrezgési periódus kezdetét jelölni kell a hullámformában), valamint a hanghatárok jelzéseivel. A következőkben kétfajta hangelembázist és egy nagy méretű beszédadatbázist ismertetünk.

8.2.1. Hangelembázis számok felolvasásához

A számok gépi felolvasására számos megoldás kínálkozik. A legegyszerűbb, mikor minden számjegyet külön ejt a gép (telefonszám, mérőóra állása, bankszámla száma). Ehhez olyan elemtárat készítenek, amelyben csak 10 szót rögzítenek nullától kilencig. Az elemtár kicsi, a működtető algoritmus is egyszerű. A számok megszó-

laltatása egyenként, megfelelő hosszúságú szünetek közbeiktatásával történik. Az ilyen felolvasás a magyarban természetellenes, nem követi a kiejtési szokásokat, tehát nem célszerű alkalmazni. Sok megoldásban nem lehet számjegyenként felolvasni az üzenetben szereplő számot (pontos idő, számlaösszeg, dátum stb.). Ekkor a fejlesztő (az írott formát alapul véve) bővíti az elemtárat azokkal a számelemekkel, amelyek biztosítják, hogy a teljes számot számként lehessen kimondatni. Az ilyen elemtár 25 elemből áll, a felhasznált számelemek a következők: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, tizen, 20, huszon, 30, 40, 50, 60, 70, 80, 90, 100, ezer, millió, milliárd. A szintéziskor a megfelelő számelemek hullámformáit egymás után összefűzik, és az eredményül kapott hullámformát szólaltatja meg a gép. Ez az elembázis már nagyobb méretű, az összefűző algoritmus is bonyolultabb. Az ilyen elven összeállított elembázis csak elméletileg követi a magyar kiejtési szabályokat (ugyanis az írásképből indul ki). Gyakorlatilag bármely szám kimondatható vele, ugyanakkor közel sem ad olyan akusztikai eredményt, amilyent az elvből fakadóan elvárnánk. A 25 számelemet tartalmazó elembázissal működtetett számfelolvasó hangzása messze elmarad attól a hangzástól amit a szám természetes felolvasásából kapnánk. Ennek ellenére ilyen elemekből építkezik sok számfelolvasó a gyakorlatban (banki rendszer, telefonon felhívható információs rendszer, mérőóra-bediktáló, GPS stb.). Meghallgatva ezeket a számfelolvasást is tartalmazó beszédinformációs rendszereket, feltűnik, hogy a számokat szaggatottan, természetellenes hangsúlyozással, az összekapcsolás határpontjain megjelenő amplitúdó- és dallamív-különbségek miatt „döcögősen” ejti ki a gép (8.10. ábra). A hallgató nagyon érzi, hogy a rendszer által összeállított szám



8.10. ábra. A 37 724 szám kimondása egy hangyományos banki tájékoztató rendszerben (5,5 s). Hullámforma (fenn), hangintenzitás (középen), alapfrekvencia (dallamvonulat, lenn)

a fenti elemekből került generálásra és érzi azt is, hogy a kiejtés prozódiai szerkezete, hangsúlyozása természetessége messze elmarad a természetes ejtéstől. Ennek következménye lehet például, hogy félreértjük a számot, hosszabb számok esetén le kell jegyezni az automata által mondott számelemeket és csak utána tudjuk helyesen értelmezni. Ez a gyenge hangminőség annak is a következménye, hogy a beszédszolgáltatást megtervező szakemberek az írás szintjén gondolkoztak, nem vették figyelembe a beszédképzés szabályait, a koartikulációból adódó törvényszerűségeket,

az ezzel kapcsolatos beszédakusztikai és fonetikai tényeket. Ezek a beszédszolgáltatási megoldások nem felelnek meg a 21. század egyébként magas követelményű műszaki színvonalának. A szolgáltatás informatikai része korszerű, a hozzá tervezett beszédszintetizátor minősége ugyanakkor megkérdőjelezhető.

8.2.1.1. Jó minőségű számfelolvasó hangelembázisának tervezése

A következőkben ismertetjük, hogy milyen elv alapján kell egy korszerű, emberi ejtést hűen utánzó számfelolvasó hullámforma-elembázisát felépíteni. Az új elv lényege, hogy már az elemtár elemeinek meghatározásánál figyelembe vesszük, hogy a beszéd folyamatosan változó akusztikai jel a frekvenciaszerkezet (f), az intenzitás (i) és az időszerkezet (t) vonatkozásában. Ezt a folyamatos változást kell követni a tervezésnél, és az elembázis összeállításánál, majd a kimondandó szám hullámformajelének megvalósításánál (elemösszefűzés) is. A frekvenciaszerkezetet két részre kell bontani, nevezetesen a koartikulációból (k) adódó spektrális változásokra és a beszéddallamra (d). Ebből adódik, hogy ideális esetben az eredeti 25 számelem mindegyikét egy-egy négy paraméteres függvény szerint (k, d, i, t) kellene meghatároznunk. Ez azt jelenti többek között, hogy az elembázis nem 25, hanem jóval több hullámformaelemet fog tartalmazni. Megjegyezzük, hogy ennek az elvnek az alkalmazása a számelemekre tulajdonképpen az elemkiválasztás-alapú szövegfelolvasási technológia előfutárának tekinthető magyar vonatkozásban (lásd a 10.3.7. fejezetet). Az új elv alapján egy új számelemet a következő általános összefüggés szerint lehet származtatni a régi számelemből.

Fonetikailag illeszkedő, új számelem = eredeti számelem (k, d, i, t)

A fenti függvény szerinti modellhez az alábbi kérdésekre kell válaszolni:

Milyen koartikulációs szabályok szerint kell osztályozni a számelemeket, hogy az összekapcsolások után az akusztikai szerkezetben a lehető legkisebb torzulások keletkezzenek és a folyamatos hangzás megvalósuljon?

Milyen az általános dallammenet a kiejtett számon belül? Vannak-e hangsúlyozásra utaló F_0 kiemelkedések?

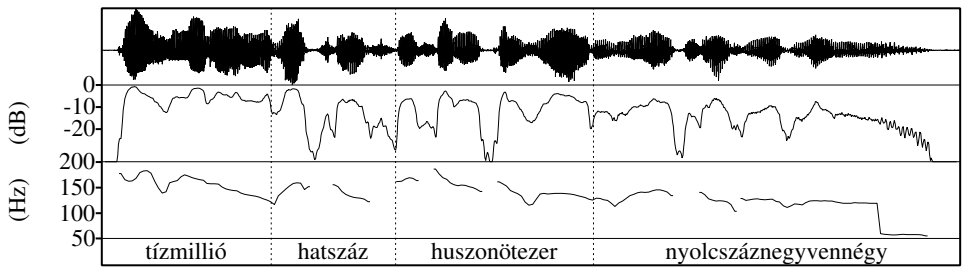
Milyen a számok kiejtésének általános intenzitásstruktúrája?

Milyen a számot felépítő számelemek időtartama a számban való elhelyezkedésük szerint?

Hány elemtípust kell megkülönböztetni, ha a számot az emberi ejtéshez közel álló ritmikával akarjuk megvalósítani?

A következőkben sorra vesszük a fenti kérdéseket. A számelemek időszerkezetére vonatkozó mérések átlagolt eredményei szerint minden számelemet legalább háromféle időtartammal: kezdő elem, belső és befejező (utolsó) kell megvalósítani,

hogy közelítsük a természetes ejtés ritmikai szerkezetét (Olaszy 1995b). A dallammenet szoros összefüggést mutat az időszerkezet hármas felosztásával. Hasonló a helyzet az intenzitás alakulásával is. Ez azt jelenti, hogy a t, i, d paramétereket nem kell külön kezelnünk, mindegyikre az előbbi hármas felosztást (kezdő, belső, utolsó) alkalmazhatjuk, a modell tehát egyszerűsödik. Az 8.11. ábrán bemutatjuk egy szám kimondásakor megvalósuló akusztikai szerkezet részleteit. Látható,



8.11. ábra. A 10 625 844 szám természetes ejtésű változatának akusztikai diagramjai (2,6 s) (hullámforma, hangintenzitás, dallammenet, szöveg)

hogy az alapfrekvencia (F_0) alapvetően ereszkedő tendenciát mutat. Erre az általános alapfrekvencia-görbére szuperponálódnak a számelemek hangsúlyozásából eredő emelkedő-eső F_0 változások a hangsúlyok helyein. A számelemek önálló ereszkedő tendenciájú alapfrekvencia-változással bírnak önmagukban is, amelynek az indulási frekvenciája mindig magasabban van, mint az előző elem befejezésekor jelen lévő F_0 (az ilyen elemeket hangsúlyozottnak tekintjük). Látható, hogy ez alól csak a helyiértéket kifejező elemek kivételek, azokban nincs hangsúlyozás, mivel ezek az elemek az őket megelőző elemhez kapcsolódnak és azzal kiejtésileg egybeolvadnak. Tehát a szám kiejtésekor az alapfrekvencia egy fűrészfogazathoz hasonló görbe szerint változik amelynek általános tendenciája: gyengén ereszkedő. A hangsúlyozás modellezése szempontjából tehát megállapíthatjuk, hogy az elméleti számelemek mindegyike tartalmaz hangsúlyozásra utaló szótagot, csak a helyiértéket kifejezők (száz, ezer, millió, milliárd) nem. Ezek után már csak a koartikulációs hatásokat (k paraméter) kell tisztáznunk ahhoz, hogy teljes legyen a természetes ejtés modellezése. Ehhez a spektrális tartalmat kell vizsgálni, mégpedig a felhasználandó építőelemek találkozási pontjain, a számelemek kezdő és utolsó hangjainak vonatkozásában. A spektrális illeszkedés modellezésére meghatároztuk a visszafelé ható (8.8. táblázat), valamint az előre ható (8.9. táblázat) hangmódosulásokat. Ha ezeket a hatásokat figyelembe vesszük, akkor az elemek összefűzésénél a csatlakozási pontokon a spektrális tartalom nagyjából egyezni fog a percepciós elvárással, és érezni fogjuk a hangzás folyamatosságát is. Koartikulációs szempontból tehát olyan elembázist kell tervezni, amelyikben az egyes számelemekből többféle akusztikai változatot is tárolunk. A természetes ejtést megközelítő modellünk tehát két szintet tartalmaz, a kez-

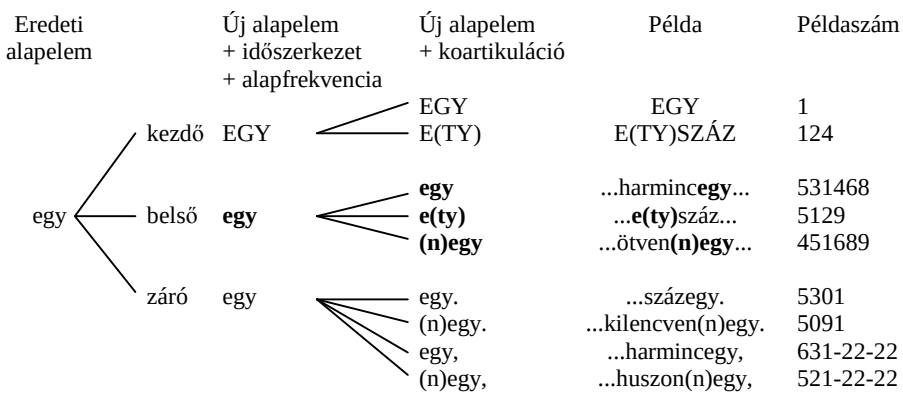
8.8. táblázat. Visszafelé ható egymásra hatás két számelem kapcsolódási határán

Az előző elem eredeti utolsó hangja (betűképpel)	Az előző elem megváltozott utolsó hangja (betűképpel)	Ha a hozzá csatlakozó elem első hangja (betűképpel)	Példa
b, d, g, v, z, zs, gy	*p, t, k, ty, f, sz, s*	*p, t, k, c, cs, sz, s, f, h*	*Négyszáz*
c	*c* zárfelpattanás nélkül	*sz*	*Kilen**cs**záz*
n	*n(k)*	*k*	*Tize**nk**ettő*
n	*n(h)*	*h*	*Huszo**nh**árom*
n	*nn*	*n*	*Ötve**nn**égy*
n	*ny*	*ny*	*Hatva**nny**olc*
n	*m*	*m, b, p*	*Ötve**nm**illió*
Magánhangzó	Magánhangzó+átmeneti szakasza	Magánhangzó	*Kett**ő**ezer*

8.9. táblázat. Egymásrahatás előrefelé két számelem kapcsolódási határán

Az előző elem utolsó hangja (betűképpel), ami előre hat a következő hangra	A csatlakozó elem eredeti első hangja	Változás a csatlakozó elem első hangjában	Példa
n, m	Magánhangzó	Nazalizálódott magánhangzó	*Ötve**ne**zer, három**e**zer*
ny, gy, ty	Magánhangzó	A magánhangzó átmeneti szakasza	*Né**gye**zer*

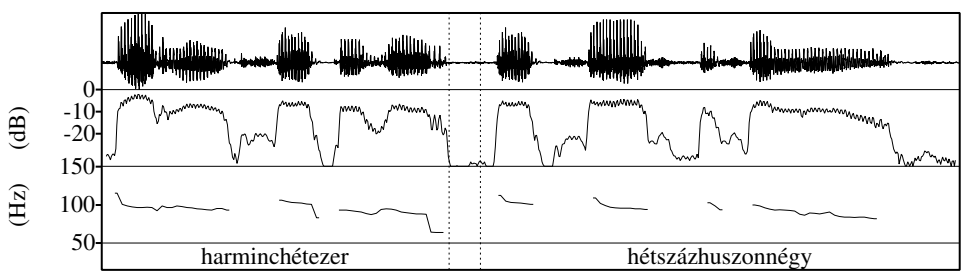
dő, belső, utolsó elemek halmazát, valamint az ezekre vetített, koartikulációt figyelembe vevő további elemeket (8.12. ábra). Fontos észrevenni, hogy csak a spektrális hatásokat tartalmazza az ábra. Idekívánkozna az „egy” hosszú változata is, melyet az egyezer kiejtéséhez használunk. Mivel ebben az esetben csak intervokális nyúlásról van szó, ezt nem kell külön elemként felvenni a hangelembázisba.



8.12. ábra. Az „egy” számelem optimális változatai (9-fajta). Ezeket kell az elembázisban eltárolni

Ha a fenti modell szerint határozzuk meg a számfelolvasó akusztikai építőelemeit, akkor sokkal több számelemvariánsunk lesz az elembázisban, mint 25, viszont

a kimondásra összeállított szám korrekt idő- és intenzitásszerkezettel fog rendelkezni, automatikusan tartalmazni fogja a szám belső hangsúlyozását megvalósító alapfrekvencia-változásokat, és a spektrális szerkezete is nagymértékben közelíteni fogja a természetes ejtését. A hangminőség javítása érdekében tehát tovább növekszik a hullámforma-elembázis, és bonyolultabb lesz az összefűzést végző algoritmus is. Ha rendszerünket pontosan illeszteni akarjuk a természetes beszédben előforduló számfelolvasásokhoz, akkor a befejező pozíciójú számelemeket további két csoportra kell osztani, teljesen befejező típusúra (vége van a számfelolvasásnak), illetve lebegőre (az elem kimondása után még következnek számelemek szünettartás után, például nyerőszámok, telefonszámok felolvasásakor). A két csoport között csupán a dallammenetben van különbség. A befejezők egy végleges, befejező formát képviselnek (eső dallam és a vége a legalacsonyabb F_0 értéken van a kimondott szám bármely részéhez viszonyítva). A lebegő típusúak ugyan számvégződési pozícióban vannak a hangsor szempontjából, mivel utánuk szünet következik, de ez nem jelenti a számfelolvasás végét Az ilyen számelemek dallammenete szinttartó, esetleg kissé emelkedő, ez adja a természetes hangzást, például telefonszámok számcsoportjainak felolvasásánál. Az elembázis optimális kihasználásához gondoskodni kell egy olyan válogató algoritmusról is, amelyik az elemtárból a legmegfelelőbb elemet választja ki az összefűzésnél. Az algoritmusnak a záró elemeknek a kiválasztását is el kell végezni. A fent ismertetett modell alkalmazásával az elemtár nagysága megtízszereződik, az összefűző algoritmus is bonyolultabb, ennek következménye viszont a rendkívül élethű hangzás. Az új elvek alapján készített számfelolvasó folyamatos kiejtésben, tökéletes prozódiával beszél (8.13. ábra). A percepciós tesztek eredményei a beszéd minőségét és természetességét illetően azt mutatják, hogy a hallgatók nem tudják eldönteni, hogy a fenti rendszer által összeállított szám egy beszédszintetizátortól származik-e, vagy egy bemondó felolvasásában hangzik el. Hasonlítsuk össze



8.13. ábra. A 37 724 szám kimondása a fonetikaimodell-alapú számfelolvasó program ejtésében

a 8.10. és a 8.13. ábrát. Látható, hogy a 25 elemes technológiával működő banki rendszer hangja szaggatott, „döcögő”, dallamában és amplitudójában természetellenes, torz. Ezzel ellentétben az új rendszer ugyanazon diagramjai folyamatosak, sok-

kal közelebb állnak az 8.11. ábrán bemutatott képekhez. A banki rendszer hangját mutató oszcillogramból látható, hogy nagyon szaggatottan hangzik fel a szám (8.10. ábra). Az elemek közé szüneteket iktattak a fejlesztők (ezzel próbálván kompenzálni, hogy az adott számelemek egymáshoz való spektrális illesztése nincs megvalósítva). A szám kimondása 5,5 s-ot vett igénybe. A középső ablakban a kimondott szám hangintenzitás-görbéjét ábrázoltuk. Látható, hogy az „ezer" és a „száz" elemek sokkal intenzívebben szólnak, mint például a „hét". Ez szintén természetellenes, mivel a számok kimondásakor a helyiértéket megtestesítő részek mindig hangsúlytalanok, tehát intenzitásuk alacsonyabb, mint a környezetüké. Az alsó ablakban a szám dallamgörbéjét ábrázoltuk. Látható, hogy a „harminc" elemet ereszkedő hanglejtéssel ejti a rendszer, mint ha befejező elem lenne, noha ez az első eleme a számnak. Az „ezer" és a „száz" elemek hangmagassága magasabb, mint a „hét" elemé, ami a természetes ejtésben sohasem fordul elő. Mindezek a hibák összeadódnak a kiejtés során, és ennek eredménye egy igen furcsa számkimondás. A fonetikai modell megoldásában (8.13. ábra) az akusztikai görbék alapján arra következtethetünk, hogy a szám kiejtése nagyon közel áll a természetes ejtéshez (annak ellenére, hogy itt is ugyanazokból az elemtípusokból raktuk össze a számot). Az oszcillogramból látható, hogy csak ott van szünet, ahol a szám kiejtési logikája megkívánja (az „ezer" után), a többi rész folyamatosan hangzik el. Ez látszik az összidőtartamon is, itt a szám kimondása mindössze 2,6 másodperc. Az intenzitásvonal és a dallamgörbe kiegyenlített és enyhe csökkenést mutat. Ez a normál beszédre jellemző képlet.

Hogyan kell elkészíteni a fonetikai modellt kiszolgáló számelembázis hullámformáit? Egy felolvasási listát kell összeállítani, amelyet egy bemondó fel fog olvasni. A lista számokat fog tartalmazni, olyanokat, amelyek mind a 25 számelem összes variánsát hordozzák a megfelelő pozíciókban a 8.12. ábra példája szerint. A lényeg, hogy az elembázis leendő számelemét a természetes ejtési környezetéből vágjuk majd ki. Ezért minden kivágandó számelemre meghatározunk egy példaszámot (vivőszámot), amit a bemondó fel fog olvasni (általában ez többjegyű, hogy a kiejtés kiegyenlített legyen nagy számok kimondása esetén is). Annyi példaszámot célszerű felolvastatni, amennyi számelemet az elembázisunk tartalmazni fog. A felolvasási listát négy részre bonthatjuk: kezdő elemek számai, belső elemek, utolsó záró elemek és utolsó felsorolásos (lebegő) elemek.

Egy ilyen gyakorlati felolvasási lista részletéből, valamint a kivágandó elemek jelöléséből mutatunk be részletet a 8.10. táblázatban.

Vannak olyan számelemek, amelyekből kevesebb változatot kell felvenni, és vannak olyanok, amelyekből többet. Amennyiben a számfelolvasót konkrét rendszerben akarjuk felhasználni, akkor a számokhoz járuló további szóelemeket (például forint, dollár) is hasonló módon kell az elembázisba beépíteni. Amennyiben telefonszámokat akarunk felolvastatni, célszerű a nulla elemet többféle variációban is felvenni figyelve a csatlakozási koartikulációs hatásokra (például nulla-nulla lehet egyetlen elem).

8.10. táblázat. Példa a számfelolvasó elembázisához felolvasandó lista néhány elemére és a kivágandó elemekre. Az egy pont befejező elemet jelöl, a vessző lebegő típusú befejezőt, a több pont azt, hogy az elem előtt, illetve után van hangsor

Kezdő elemek		Belső elemek		Utolsó elemek befejezéshez		Utolsó elemek felsoroláshoz	
Felolvasandó	Kivágandó	Felolvasandó	Kivágandó	Felolvasandó	Kivágandó	Felolvasandó	Kivágandó
1 652 844	egy...	2 651 962	...(n)egy...	321.	...(n)egy.	321-278.	...(n)egy,
egyszázkettő	e(ty)...	131 562	...(c)egy	2001.	...egy.	2001. május	...egy,
213	kettő ...	5119 (egyszáz!)	...ety...	522.	...kettő.	522-389	...kettő,
3514	három...	5 650 294	...kettő...	913.	...három.	913- 218	...három,
4 987 541	négy...	3 512 645	...kettő(e)...	3 004.	...négy.	3004, 2001	...négy,
493	néty...	1 563 546	...három...	4 055.	...(n)öt.	4055, 3200	...(n)öt,
5973	öt...	5 624 239	...négy...	4 005.	...öt.	4005, 2200	...öt,
6973	hat...	154 493	...néty...	396.	...hat.	396-987	...hat,
7973	hét...	535 599	...öt...	3 497.	...hét.	3.497, 5000	...hét,
8973	nyolc...	225 973	...(n)öt...	938.	...nyolc	938- 22-22	...nyolc,
893	nyol-(sz)...	396 973	...hat...	6 329.	...kilenc	6329, 2345	...kilenc,
9973	kilenc...	1 497 299	...hét...	16 510.	...tíz	16510, 98	...tíz,
993	kilen(sz)...	938 973	...nyolc...	920.	...húsz	920- 22 22	...húsz,
10 629	tíz...	2 829 812	...nyol(sz)...	2530.	...harminc	2530, 9800.	...harminc,
11 547	tizen...	1 329 973	...kilenc...	1540.	...negyven	1.540, 9700	...negyven,
12 123	tizen(k)...	563 933	...kilen(sz)..	1250.	...ötven	1250, 5600	...ötven,
16 975	tizen(h)...	1 910 629	...tíz...	860.	...hatvan	860-22-22	...hatvan,
14 510	tizen(n)...	6810 forint	...tísz...	5 970.	...hetven	5970, 6200	...hetven,
18 933	tizeny(ny)...	2 611 341	...tizen...	1780.	...nyolcvan	1780, 7400	...nyolcvan,
20 629	húsz...	319 639	...tizen(k)...	1990.	...kilencven	1990, 5500.	...kilencven,
21.547	huszon...	213 123	...tizen(h)...	1300.	...száz	1300, 2400	...száz,
29 639	huszon(k)...	314 642	...tizen(n)...	53 000.	...(m)ezer	53.000, 698	...(m)ezer,
26 975	huszon(h)...	3 218 978	...tizen(ny)...	5 000.	...ezer	5000, 7500.	...ezer,
24 250	huszon(n)...	920 629	...húsz...	62 000.	...(ő)ezer	62.000, 987	...(ő)ezer,
28 933	huszony(ny).	1 921 547	...huszon...	54 000.	...(gy)ezer	54.000, 598	...(gy)ezer,

Az elembázis készítésének második fázisában kivágjuk a felolvasott számból azt a részt, amelyiket számelemként akarunk majd felhasználni a szintézis során. A kivágást minden esetben fázishelyesen kell elvégezni (például a negatív irányba történő nullátmenetnél, zöngés hangoknál mindenképpen periódushatáron). A kivágott számelemet elhelyezzük az elembázisban. Ha feltöltöttük az elembázist és elkészítettük a válogató algoritmust, akkor ki lehet próbálni a számfelolvasót. Szisztematikusan végig kell próbálni az összes elem összekapcsolását (sok szintetizált számot kell meghallgatni). Ez egy előzetes hangzási teszt, nem a végleges állapot, mert általában lesznek hangzási hibák (nem sok), amelyeket korrigálni kell hangsebészeti módszerekkel, akusztikai csiszolással (lásd a következő fejezetben). Ez a munka speciális szaktudást igényel. A csiszolás folyamán amplitúdóban hozzá kell igazítani egymáshoz az összefűzendő elemeket, esetlegesen hangidőtartamokat kell nyújtani, illetve rövidíteni és a csatlakozási pontokat kell pontosan egymáshoz illeszteni. A csiszolással párhuzamosan folyamatosan újra kell hallgatni az éppen javított elemmel összeállított számot. Az így finomított elembázissal olyan hangminőséget lehet

elérni, hogy a szintetizált üzenetben nem lehet megkülönböztetni a felvett beszédet (fix üzenetrészek) és a hozzá kapcsolt számfelolvasót (változó üzenetrész). Az akusztikai csiszolás kitartást és türelmet igényel.

8.2.2. Logatomalapú, diád-, triád-hangelembázis szövegfelolvasáshoz

Az a gondolat, hogy az emberi hangból vágjanak ki hangsorépítő hullámformákat és ezeket egymásután kapcsolják össze szintézis céljából, magyar feltaláló agyában született meg 1916-ban (8.14. ábra). Ez a gondolat tetszőleges szöveg reprodukálására alkalmas beszélőgép megalkotását célozta hanghullámrészletek egymásutáni lejátszásával (viaszhengerekről). Ez volt az első beszédszintézisre irányuló szabadalom a világon (Bánó 1916). Később, az 1950-es években a számí-

Megjelent 1919. évi junius hó 21-én.

MAGYAR SZABADALMI HIVATAL.

SZABADALMI LEIRÁS

74361. szám.

IX/d. OSZTÁLY.

Tetszőleges szöveg reprodukálására alkalmas beszélőgép.

DR. BÁNÓ MIKLÓS OKL. MÉRNÖK ÉS KÖZGAZDASÁGI MÉRNÖK BUDAPESTEN.

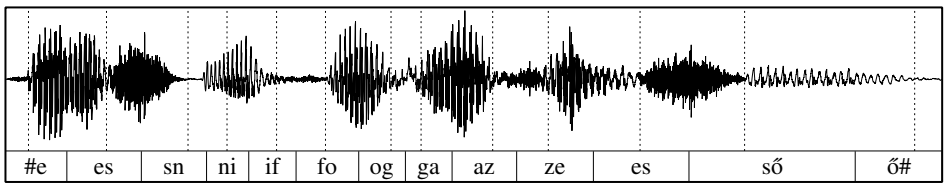
A bejelentés napja 1916. november hó 30-ika.

8.14. ábra. A hullámformaösszefűzés gondolata Bánó Miklós szabadalmában

tógép felhasználásával valósították meg ezt a technológiát (Harris 1953, Peterson et al. 1958). Egy ilyen hullámforma-elemtár összeállításánál alapvető kérdés, hogy a nyelv milyen méretű legyen az eltárolni kívánt hullámforma (hangnál rövidebb, hang, annál hosszabb, szó, szófüzér stb.). A hangalapú elemek jelentenék a legkisebb elemtárat és a legegyszerűbb összefűző algoritmust. Ezt ritkán használják, mert ezzel a módszerrel nehézkes megvalósítani a hangátmeneteket, ha viszont ezek nincsenek benne a végleges jelben, akkor az összeállított hangok sorozata érthetetlen lesz. A két elem közötti interpoláció algoritmikus megoldása sem egyszerű. Ebben a fejezetben a hangkapcsolódásokra épített megoldásokról beszélünk.

8.2.2.1. Diád-hangelembázis

A legegyszerűbb szerkezetű átmenetet tartalmazó hangsorépítő elemek az úgynevezett diádok, amelyek két félbeszédhangnyi hullámformarészletet tartalmaznak. A diádok megfelelő összefűzéséből elvileg létre lehet hozni a szintetizálandó beszédhullámot. Egy példát mutatunk be az ilyen összefűzésre a 8.15 ábrán. A diádban au-



8.15. ábra. Az *Esni fog az eső* mondat (1,13 s) szintetizált hullámformája 13 diádból összefűzve. A függőleges szaggatott vonalak mutatják a hanghatárokat a hulláformában. Az alsó részen láthatók a diád építőelemek és határaik. A # jel a hangsorkezdő és hangsorzáró diádot jelöli (az első hang kezdeti, illetve az utolsó záró szakaszát). A hangokat a betűjelükkel jelöltük

tomatikusan benne van a két hangot összekapcsoló, úgynevezett hangátmeneti rész. Ez a gondolat beszédépítési szempontból logikus, hiszen véges számú elemmel lefedhető az összes hangkapcsolódás. Ha minden beszédhang-kapcsolódást legalább egyszer tárolni akarunk a diádos elemtárban, akkor elvileg az adott nyelv beszédhangjainak a négyzetével kell számolnunk (például 40 hang esetén 1600 diád lefedi az elképzelhető összes kettős hangkapcsolatot). Érzékelhető tehát, hogy a diádok száma nem túl nagy, ezért jól kézben tartható az ilyen adattár. A diádos adatbázis jellemző szerkezeti tulajdonságai: zárt rendszerű, tudatosan megtervezett, csak egyfajta szerkezetű építőelemet tartalmaz, egységes felépítésű. Itt egy rövid időre lépjünk vissza az időben, és nézzük meg azt a fejlődést, ami elvezetett a mai, kifinomult, hangkapcsolódásokon alapuló szövegfelolvasókhoz. Az eredeti gondolat az 1970-es években született. A felmerülő kérdések a következők voltak. Az ilyen diádokból összeállított beszéd hogyan fog hangzani? Vajon hasonló hangzású lesz-e, mint a számunkra megszokott dallamos, ritmusos beszéd? Mennyire fogja megközelíteni azt? A kezdeti diádos kísérletek eredménye messze alul maradt az elvárásokon, nehezen érthető, huppogó, recsegő beszéd született. Ez abból fakadt, hogy az első ilyen elemtárak létrehozásánál 2–3 szótagú szavakat olvastattak fel (*baba, mama, látok, farmer, számol, bajos*), és ezekből vágták ki a diádokat, majd tárolták el őket egy hullámforma-elembázisban. Az ilyen diádok összekapcsolásával létrehozott beszédjel minősége nagyon rossz volt. Miért? Egyrészről azért, mert a szófelolvasásnál prozódiai elemeket is belevihet a bemondó a beszédjelbe (hangsúly, szódallam stb.), és ezek véletlenszerű összekapcsolása furcsa hangzást eredményez (az alapfrekvencia ugrálhat, a hangintenzitás szintén). Másrészről a felolvasott szavakban az egymáshoz kapcsolódó hangok egymásra hatása a szó hangsorától függ,

tehát véletlenszerű. Ebből az következett, hogy a CV1 és V1C szerkezetű diádok magánhangzó-részeiben kismértékben eltérő akusztikai szerkezetek találkoztak az összekapcsolási ponton (ezek összekapcsolása torzított jelet hoz létre a spektrális ugrás következtében). Példaként vegyük a *Maros* szó összeállítását. Ehhez hat diádra van szükségünk: -#m-, -ma-, -ar-, -ro-, -os-, -s#-. Az első kettő a *mama* hangsor elejéből kivágott diádokból állhat, a harmadik a *farmer* első szótagjából, a negyedik a *párok* második szótagjából, az ötödik és hatodik a *bajos* szó végéből kinyerhető. Belátható, hogy az így felépített magánhangzók spektrális komponensei a vágási pontokon erősen eltérhetnek egymástól, ami torzítást okoz (a -ma- diád magánhangzója után egy nazális hang volt az eredeti szóban, az -ar- diád magánhangzója előtt egy dentialveoláris réshang és így tovább). Ebből fakadóan érezhető volt, hogy a szegmentális szintű hangtest akusztikai minősége messze elmaradt az elvárttól. Mi volt ennek az oka? Az, hogy a fenti elvek alapján készített diádok alkalmazása sérti a beszédre jellemző folyamatossági kritériumot (a spektrális tartalom, az intenzitás, az alapfrekvencia viszonylag lassan és folytonosan változik, nincsenek benne hirtelen változások). Mi is történt? Szétdarabolták a beszédet kis részletekre és utána a különböző időpontokban ejtett hullámformadarabokat összillesztették, mint ha együvé tartoznának. Mindezek mellett még a megfelelő hangkörnyezet hatását sem vették figyelembe. Ez alapvetően ellentmond annak a tételnek (ami a biológiai rendszerből fakad), hogy a beszéd a pillanathoz kötött egyedi és egyéni hanghullám (különböző időpontokban ejtett ugyanazon szó hangzásában is különbség lehet). Világossá vált tehát, hogy a beszéd hullámformájából részeket kivágni és egymással összekapcsolni csak igen megfontolt kritériumok alapján lehet. A beszéd belső szerkezetének mélyebb elemzése, valamint a számítástechnika fejlődése is segített abban, hogy az előbb említett kritériumokat meghatározzák. Mire kell itt gondolni? Arra, hogy a beszédépítést ennél a technológiánál célszerű lenne legalább két lépésre szétválasztani: szegmentális és szupraszegmentális szintre. Az alapot jelentsék az összefűzött diádok, a felépítményt (a dallamot, a ritmust, a hangsúlyt) a második építési szakaszban ültessük rá az alapra, és ekkor lesz kész a beszédépítés folyamata. Az alap minőségi elkészítése a legfontosabb, hiszen ha ez zajos, kiegyenlítetlen, spektrálisan nem folytonos, akkor ezen már a prozódia ráültetése nem fog javítani. Ebben a fejezetben csak a diádok szintjével foglalkozunk, a prozódiai rész megvalósítására a szintetizátoroknál térünk ki. A diádok elkészítésére vonatkozó legfontosabb kritériumok a következők:

1. Nem szavakat kell felolvastatni, hanem értelmetlen, három szótagos hangsorokat, melyeket tudatosan tervezünk meg. Ezeket logatomnak nevezik.
2. A felolvasáskor monoton kiejtést kell megvalósítani, egyazon alaphangmagasságon kell beszélni (ez nehéz feladat).
3. A felolvasásnál a hangerőt is ugyanazon a szinten kell tartani.

4. A felolvasást tempósan kell végezni, minimális szünetet tartva két elem között (három-négy logatomot kell egy csoportban folyamatosan kiejteni, mint ha mondatot olvasnánk. Erre azért van szükség, mert csak így lehet közelíteni a természetes beszédre jellemző artikulációs sebességet (szófelolvasásnál az emberek lassabban beszélnek, mint például mondatok, szövegek felolvasásánál).

A felolvasandó anyag megtervezése. A logatomok szöveglistájának tervezésénél figyelembe kell venni az alapkoncepciót, azt, hogy a kivágandó diádok szerkezete minél közelebb álljon a szegmentális szintű követelményekhez. Továbbá biztosítani kell a lehető legjobb spektrális egyezőséget a diádok határán, hogy az összekapcsolásukkor a frekvenciaszerkezeti folytonosság megvalósuljon.

Hangtervezés. A kérdés az, hogy milyen hangokra kell diádot tervezni? A magyarra mutatunk be példát. Mivel a magyarban 65 fonéma van, a kételemű hangkapcsolatok összes kombinációjának száma igen nagy lenne (4225). Egyszerűsíteni célszerű, de úgy, hogy a fonetikai minőség ne csorbuljon. Ilyen megoldás, ha a hosszú mássalhangzókra nem tárolunk beszédmintát, azokat majd algoritmussal nyújtjuk, valamint így teszünk az [a:] hang rövid változatával is. Erre az a fonetikai tény ad felhatalmazást, hogy nincs hallható spektrális különbség az említett hangok rövid és a hosszú változata között, csak időtartamban különböznek egymástól. Mindezek tükrében alapvetően a következő diádokra lesz szükségünk: minden CV, VC, VV, CC elemre egy-egy logatom. Az elemszámok tehát 25 mássalhangzóra és 14 magánhangzóra számolva: CV-ből 2x25x14, VV-ből 14x14 és CC-ből 25x25, vagyis összesen 1521 db logatom, ennyi eleme lenne az elemtárnak. Gondolni kell arra is, hogy a hangsor első és utolsó hangjára is kell diádelemet tervezni (a diádos elemtárban a # jel jelenti az induló, illetve befejező hangelemet). Ezt minden hangra elvégezve a magánhangzók diádelemeihez 2x14 logatomot kell még hozzáadni, a mássalhangzókéhoz 2x25-öt. Ezenfelül célszerű felkészülni a variánsokra is, legalábbis azokra, amelyeknek a hangzását mindenképpen biztosítani akarjuk a hangsorban. A magyarban ilyen hang a [j] zöngétlen változata (*lopj, hívj ki*). Ez a hang csak bizonyos hangkörnyezetek esetén jön létre, tehát csak néhány logatomot kell rögzíteni ahhoz, hogy ilyen diádokat is be tudjunk tenni az elemtárba.

Milyen legyen a logatom belső felépítése? Az biztos, hogy a tervezett diádnak szerepelni kell benne. Ha monoton felolvasást kérünk, akkor elérjük, hogy az alapfrekvencia közel állandó lesz a logatomon belül, tehát nagy F_0 változásra nem kell számítani. Gondolni kell a hangsúlyozás kiküszöbölésére is. Úgy kell megtervezni a logatom belső szerkezetét, hogy a kivágandó elem (a kérdéses diád) ne az első szótagban legyen, hanem a másodikban. Ezzel azt értük el, hogy ha a bemondó véletlenül hangsúllyal ejtette a logatomot, akkor az nem azon a részen realizálódott, ami számunkra érdekes volt az elembázis szempontjából. A diádok tervezésének egyik legfontosabb kérdése, hogy a félbevágott hangokban hogyan lehet biztosítani

a spektrális egyezőséget a két félhangrész egymáshoz csatlakozó pontján? A mássalhangzóknál ez a követelmény nem annyira kritikus, mivel egyrészről sok esetben kis intenzitású hangrészre esik ez a pont, másrészről a nagyobb intenzitású réshangokban a zörej frekvenciaszerkezete nem mutat nagy eltérést a hangkörnyezet függvényében. A magánhangzóknál más a helyzet. Ezek nagy energiájú hangok, tehát a spektrális egyenetlenség (ha ilyen lesz az összekapcsolás után) zavaró lesz, és torzítani fogja a hangot. A magánhangzó belső frekvenciaszerkezetét a hozzá csatlakozó hangok megváltoztatják a koartikulációs hatások miatt, és ez hangfüggő. Minden hangkörnyezet más-más változást idéz elő. Mi lehet a megoldás? Olyan hangot kell keresni, amelyik a legkisebb hatást gyakorolja a magánhangzó frekvenciaszerkezetére és ezt kell alkalmazni a kívánt pontokon. Fonetikai mérések kimutatták, hogy ez a hang a [k] mássalhangzó (lásd az 5.2.2. fejezetben). A logatom felépítése tehát a következőképpen lehet a legoptimálisabb, mondjuk egy CV típusú diádra:

első szótag + C + V + [k] + záró magánhangzó.

Az első szótagra és a záró magánhangzóra válasszuk az [ɔ] hangot, mivel ennek az ejtése (sem különösebb ajakartikulációt sem különösebb nyelvmozgást nem igényel, csak kissé ki kell nyitni az állkapcsot). Mindezek alapján az ideális logatom képlete a CV és VC diádok elkészítésére a következő:

[ɔ]+C+V+ [k] + [ɔ] (*abáka, apáka, adáka* . . . stb.),
[ɔ]+ [k]+V+ C + [ɔ] (*akoba, akopa, adoka* . . . stb.)

Ezzel a szerkezeti struktúrával elérhetjük azt, hogy a magánhangzó közepén közel ugyanaz lesz a spektrális szerkezet, függetlenül attól, hogy a CV, illetve VC diádban van. Így tetszőleges, azonos magánhangzót tartalmazó CV és VC diádok összekapcsolhatók. Az eredmény: tiszta, torzításmentes hangzás. A VV elemekre alkalmazandó képlet:

[ɔ]+[k]+V1+V2+[k]+[ɔ] (*akoöka, akaika akaeka* stb.)

A CC elemeknél – mint korábban említettük – nem kell számolni az összekapcsolódási pontokon fellépő számottevő torzítással. A CC kapcsolatokat képviselő diádok megvalósításához célszerű hosszabb logatomokat alkalmazni (esetleg értelmes szavakat alkalmazni). Erre azért van szükség, hogy a mássalhangzó-torlódás időszerkezete képviselje a folyamatos ejtésre jellemző struktúrát). Példa ilyen CC elemek vivőhangsorára: [bd]-re *labdaszedő*, [bm]-re *lábmelegítő,* [pf]-re *népfalatozó.* A felolvasandó lista tehát a fenti elvek alapján kialakított, speciális tartalmú szöveges anyag, amely egyrészről több száz logatomot, másrészről értelmes szót tartalmaz. Ennek felolvasása némi gyakorlás után is több órás munkát igényel. A felolvasást beszédtechnológiai szakembernek célszerű felügyelni, hogy a korábban említett akusztikai követelmények teljesüljenek. Harminc percenként javasolt szünetet tartani, pihentetni a bemondót. Lehetőleg ugyanazon a napon be kell

fejezni a felolvasást, mivel a bemondó hangszíne változhat egy nap elteltével. Képzett bemondók képesek ráhangolni hangszínezetüket egy korábbi hangfelvételükhöz.

A diádok kivágása a logatomból. Mielőtt kivágjuk a diádot a logatomból (szóból), szükséges, hogy az egész hanganyagot ellássuk zöngeperiódus (marker) jelölésekkel (ezzel automatikusan bejelölődnek a zöngés/zöngétlen hangszakaszok határai is). A zöngeperiódus jelölésére két módszer is kínálkozik. Az egyik, amikor a hangszalagrezgés működését vesszük figyelembe (nyitódási pont), a másik, amikor a hangrezgés maximális amplitúdójú részét keressük meg a hullámformában. Az előbbinél a zöngemarkert a hangperiódus indulási pontjára tesszük. Ezzel a jelöléssel a fiziológiai folyamatot követjük, tehát a konkrét periódust jelöljük meg. A maximumkeresési módszerrel a zöngemarkert a maximális amplitúdónál helyezzük el, ezzel csak azt jelöljük, hogy itt zöngés periódus van jelen. A zöngemarkerek távolsága mindkét esetben a pillanatnyi alapfrekvencia értékét is magában hordozza. Mindkét eljárásnak van létjogosultsága a különböző későbbi jelmanipulációk során (lásd a 7.1.4. fejezetben). A zöngeperiódusok megjelölése fáradtságos munka, a kézi jelölési technikát egyre inkább felváltják automatizált eljárások. Az emberi munka azonban nem hagyható ki teljes mértékben (ellenőrizni és a tévesztési helyeken javítani kell a gépi jelölést). A zöngemarkerek precíz beállítása kihat a beszédszintetizátor teljes működésére. A zöngétlen szakaszokon egyenlő időosztással helyezzük el a markereket (például 5 ms-onként). A markerek a későbbi jelmanipulációt (például nyújtás) segítik. További teendő a hanghatárok kijelölése. A hanghatár jelzése egybe fog esni valamelyik markerrel. A kijelölést célszerű kézzel végezni, mivel ez adja a legpontosabb jelölést. A diád kivágását bármely olyan hangszerkesztő szoftverrel elvégezhetjük, amelyik képes megjeleníteni (akár nagyítással is) a periódushatárok jelzéseit. Az olyan elembázisnál, amelyiknél a fiziológiai folyamatot követő zöngemarkerek vannak, fontos szempont, hogy a zöngés hangok elvágása periódushatáron és nullátmenetnél történjen. A diád mindkét oldalán a hullám időfüggvényének haladási iránya (balról jobbra) a nullátmeneti pontnál pozitívból negatívba mozogjon. Így biztosíthatjuk a folyamatos csatlakozást a két diád között. Zöngétlen hangok esetében ilyen szigorú kritérium nincs. A diádhullámforma-elemek kivágása, elhelyezése a hullámforma-elembázisban precíz munkát igényel, speciális szoftvertámogatással (például meg kell jeleníteni a pillanatnyi feltöltöttségi mátrixot, azaz hogy vannak-e még hiányzó diádok a hangelemtárban).

Akusztikai csiszolás. A fentiek szerint elkészített diádos elemtár már alkalmas beszéd előállítására, de a gyakorlati tapasztalat azt mutatja, hogy további akusztikai csiszolással (hangsebészeti finomítással) lényegesen lehet javítani a hangzás tisztaságát, valamint a beszéd időszerkezetének pontosságát is. A kérdés az, hogy mely elemekben és azon belül mely pontokon kell javítani. Kiindulásképpen olyan

szoftvert kell alkalmazni, amelyikkel már el lehet végezni a szöveg-hang konverziót (a diádok összefűzését), azaz, hogy szavakat, mondatokat (folyamatos beszédre emlékeztető hullámformát) állítsunk elő. Kétfajta ellenőrzést célszerű végezni a kezdeti finomítási szakaszban: szisztematikus vizsgálatot és szövegfelolvasást. A szisztematikus ellenőrzés az akusztikai hibák feltárását biztosítja. Ez abból áll, hogy a korábban felolvasott logatomokat hangosítjuk meg, minden logatomon végigmenve. A meghallgatásos tesztek során kiderül minden hangzási probléma. Milyen hibákra lehet gondolni? Például egy zárhang zárfelpattanása nem hallatszik megfelelően, előfordul, hogy valamely hang túl intenzív, illetve túl halk a hangkörnyezetéhez képest, a hangindítás túl meredek, a magánhangzó első fele sokkal intenzívebb, mint a másik fele (mivel más-más diádból származik), bizonyos hangrészek hangzása torz. Mindezek a hibák abból adódnak, hogy a bemondónak sok órán keresztül kell beszélni, ezért fáradhat, koncentrációja hullámozhat stb. Az akusztikai csiszolás azt jelenti, hogy ezeket a hibákat kézi korrigálással javítjuk (a hanghullámelemeket szerkesztjük). A javításhoz felhasználjuk a teljes diádos elemtárat, valamint a beszédhangokra vonatkozó alapvető fonetikai ismereteket (Olaszy 2003). Ezt a műveletsort hangsebészetnek is szokták nevezni. Lássunk néhány példát. Legyen a kifogásolt hang az [f] a *falak* szóban (nem jó a hangzás). Két diád érintett a problémában, az -#f- és az -fa-. Amennyiben találunk olyan [f] hangot, amelyik hátsó nyelvállású magánhangzóhoz csatlakozik és az előbbinél jobb hangzású, akkor azt bemásolhatjuk az előbbi két diád megfelelő részébe (csak a zöngétlen réshangét), és megjavul a *falak* szó első hangja, valamint az -#f- és a -fa- diád. Ha az *eszik* szó [s] hangja kissé selypesen sikerült a bemondásnál, akkor ezen könnyen segíthetünk. Keressünk első nyelvállású magánhangzóhoz kapcsolódó [s] hangot, és ha találunk jó hangzásút, akkor azt kell bemásolni az -esz- és -szi- diádok megfelelő részeibe. Ha a *cica* szó második mássalhangzóját [t͡s] szeretnénk javítani, hasonló módon kereshetünk jobban hangzó [t͡s] hangot. Amennyiben nem találunk ilyet, akkor hangsebészeti módszerhez folyamodunk. Megvizsgáljuk az *isza, iszo, iszu* betűsor meghangosításából kapott mintákat, kiválasztjuk, hogy melyiknek hangzik legjobban a réshangja és abból elkészíthetjük az előbbi [t͡s] kiváltásra alkalmas új hangrészt. Az [s] hang első felének amplitúdóját nullára csökkentjük, a réselem második felét meghagyjuk. Ezzel előállt az új [t͡s] hang, csak az időszerkezeti értékeket kell a korábbi diádok szerint beállítani. Ezután a megfelelő hangelemeket át kell másolni a rossz -ic- és -ca- diádok megfelelő részébe (kicseréltük a két diád azon részét, amelyik a [t͡s] hanghoz tartozik). Az akusztikai csiszolás időigényes munka is lehet. Gépi módszerrel ilyen összesimítás nem képzelhető el, mert csak meghallgatással lehet eldönteni, hogy egy hang a hangsorban megfelelően hangzik-e vagy sem. Az akusztikai csiszolás előnye, hogy biztosítjuk a hangsor minden pontjára a korrekt hangzást, és ezt percepciós módszerrel tesszük. A szövegek szintetizálása jöhet ezek után. Itt többnyire időtartamkorrekciókat kell végrehajtani (túl rövid, illetve túl hosszú egy hang). Ezeket a korrekciókat a diádelemekben

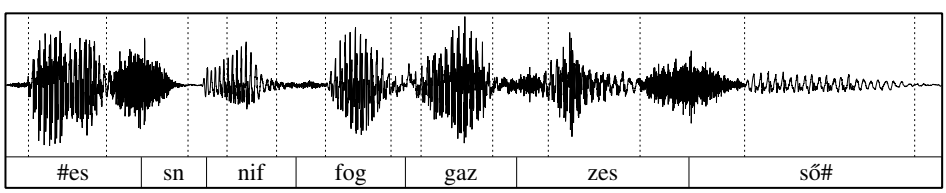
hajtjuk végre hangperiódusok kivágásával, illetve betoldásával. A két csiszolási fázis eredménye, hogy olyan diádelemtárunk, lesz amelyből szintetizált beszéd szegmentális szintű értékelése magas fokú lesz (nagyon jól érthető).

A diádos elemtár előnyei: kis tárterületet igényel, egyszerűen kezelhető, a felolvasási hiba minimálisra csökkenthető, a felmerülő hibák kis energiával javíthatók, biztos az előállított hangzás pontossága bármilyen szövegbemenetnél. Hátrányai: sok munkát igényel az elemtár előállítása, a szintetizált beszéd hangzása ugyan emberi, de nem ismerhető fel a bemondó hangszínezete. Érezhető, hogy gép beszél.

A magyarra vonatkoztatva ilyen diádos adatbázissal működik a ProfiVox beszédszintetizátor egyik alapváltozata (Olaszy et al. 2000b).

8.2.2.2. Triád a hangelembázisban

A diádos hangsorépítés torzítási hibáinak kiküszöbölésére született az a gondolat, hogy a nagy energiájú zöngés hangokat ne vágjuk el, hanem hagyjuk változatlanul, ha két mássalhangzó között vannak. Az ilyen elemeket nevezték el triádnak, ezek például két félbeszédhangot és közöttük egy magánhangzót tartalmaznak. Másfajta triád is elképzelhető. Az elemtárban tárolandó hullámformaszakasz fizikai hossza tehát megnő a diádhoz képest. Ezzel a megoldással megnő az adatbázis elemeinek száma, viszont várhatóan tovább javul a hangminőség. A VV és CC kapcsolatokat továbbra is diádokkal oldják meg (8.16. ábra). A ProfiVox rendszer továbbfejlesztett változatában CVC szerkezetű triádokat is alkalmaztak.



8.16. ábra. Az *Esni fog az eső* mondat (1,13 s) szintetizált hullámformája 6 triádból és 1 diádból összefűzve. A triád és diád építőelemek és határjelöléseik az alsó részen láthatók. A hangokat a betűjelükkel jelöltük

A felolvasandó anyag megtervezése triádos elemek készítéséhez. A triádok elkészítéséhez másfajta szöveganyagot kell összeállítani, mint amilyent a diádoknál láttunk, bár az alapelvek itt sem változnak. Szintén logatomokat kell felolvasni, szintén három szótagosakat. A logatom szerkezete változik meg a követelménynek megfelelően. Minden CVC kapcsolatra kell egy logatomot tervezni. Ennek képlete a következő: [ɔ]+ C1+V1+C2+[ɔ]. Az összes érintett hangra számolva 14x25x25 = 8750 logatommal kell számolni. A gyakorlatban kevesebb logatom is elég, mert a koartikulációs hatások miatt bizonyos hangkapcsolatok nem is tudnak létrejönni a beszédben

(lásd például az 5.19. táblázatot). A [b]+[a:]+C triádokhoz tartozó értelmetlen szavak sora így kezdődik: *ababa, abáda, abága, abágya, abápa, abáta abáka, abátya, abáma, abána, abánya* stb. Látható, hogy csak a negyedik hang változik konzekvensen, azaz minden mássalhangzó sorra kerül. A triádos lista felolvasásánál ugyanazokat az előírásokat kell követni, amelyeket a diádosnál ismertettünk.

A triádok alkalmazásának előnyei: a magánhangzók érintetlenek maradnak, tehát megtartják az eredeti bemondáskor kialakult belső szerkezetüket, a két hangátmeneti fázissal együtt. Torzítások csak a mássalhangzókban fordulhatnak elő, azoknak pedig nagy része kis energiájú, tehát a torzítás nem érezhető annyira. Javul a hangminőség, kevesebb akusztikai csiszolás kell. Hátrányok: nagyobb tárterületet igényel, nő az elemszám, több munka az elkészítése, bonyolultabb a válogató algoritmus, a hangidőtartamok kialakítása bonyolultabb, mint a tisztán diádos rendszernél.

8.2.3. Nagyméretű beszédadatbázisok szövegfelolvasókhoz

A legújabb technológiát képviselő beszédszintézis-eljárások nagy méretű beszédadatbázisokat használnak. Létjogosultságukat a technika fejlődése hozta felszínre, csökkentek a memóriakorlátok, gyorsultak a számítógépek. Ezek az adatbázisok kétféle szintézismódszerhez alkalmazhatók, közvetlenül az elemkiválasztáson alapuló gépi beszéd-előállításhoz, közvetetten pedig a gépi tanulásra épített módszerhez. Az elkészített beszédadatbázis belső szerkezeti felépítése mindkét módszert támogatja, tehát nem szükséges külön beszédadatbázist készíteni a két módszerhez. Az elemkiválaszásos technológia a beszédadatbázisban (korpusz) való közvetlen keresésen és a talált hullámformarészek közvetlen összefűzésén alapul (10.3.7. fejezet), a gépi tanulásra alapozott megközelítés a nagy beszédkorpusz(oka)t tanulásra használja fel, paramétereket nyer ki belőlük, majd egy származtatott kisebb paraméterhalmazból végzi a szintézist (10.3.8. fejezet).

Mi a korpusz? Korpuszon nagy mennyiségű adathalmazt értünk, ebben az esetben sok órányi beszédet speciális belső jelölésekkel. Az ilyen beszédadatbázis mind szerkezeti, mind tartalmi szempontból élesen különbözik az előző fejezetben tárgyaltaktól. Egyrészről nyitott, bármikor bővíthető, másrészről nincs kötöttség a tartalmi elemekre (bármilyen tartalommal bővíthető). Készítése bonyolult eszközrendszert és sok időt igényel. Egy lehetséges elkészítési módszer, hogy egy kiválasztott bemondó sok szöveget olvas fel. A hangfelvétel több hónapig is készülhet, alkalmanként 3–5 órás blokkokban. A felvétel további, belső feldolgozása szintén hónapokig tarthat. Bővítés esetén egyetlen kritériumot kell teljesíteni: ugyanazzal a bemondóval kell a hangfelvételt folytatni úgynevezett hangráhangolásos módszerrel. Ez abban áll, hogy a korábbi hangszínezet megtartása érdekében minden új felvételnél mintamondatot játszanak le a bemondó részére és megkérik, hogy hangos utánmondással hango-

lódjon rá a korábbi hangszínezetére. Minden bemondó hangjára külön korpuszt kell készíteni. Más lehetőség, hogy a korpusz hanganyaga már készen van, például korábban tárolt beszédanyagok formájában. Ekkor ezekkel a felvételekkel kell elvégezni a korpusszá váláshoz előírt munkákat.

Mit kell tartalmaznia a korpusznak? Többféle szintű és mennyiségű információt (hogy sikeresen alkalmazhassuk a fenti célokra). A korpusz tehát párhuzamosan szinkronizált blokkokból áll, ezekben tároljuk a hanghullámot (célzott címkézéssel ellátva), a szöveg ortografikus formáját, a szövegnormalizált fonemikus formát és fonetikus átiratot. Egy-egy ilyen kész korpusz 5–10 000 mondatot tartalmaz. A beszédkorpusz feldolgozási alapegysége a mondat. Minden mondatot külön azonosítóval kell ellátni. A korpuszban minden előkészítő és kereső munkafázis mondatszinten történik. A mondatok közötti esetleges összefüggéseket általában nem veszik figyelembe. A korpusz tehát egy célzottan elkészített adatbázisnak is tekinthető.

Hogyan használjuk fel az ilyen korpuszokat beszédépítéshez? Ennek megválaszolásához röviden jellemezzük az alkalmazott két technológiát. A korpuszban való, keresésen alapuló beszédszintézis két szempontból teljesen eltér a diád-triádalapú, zárt rendszerű megoldástól. Az egyik, hogy nem két lépcsőben jön létre a géppel előállított beszéd (alapjel + prozódiaráültetés), hanem egy direkt kereséssel, ahol mindkét szint jellemzőit a keresés feltételeibe építik be. A válogató algoritmus figyeli mindazokat az információkat, amelyek jellemzik a beszéd szegmentális és szuprasz egmentális szerkezetét (hangok, hangkörnyezet, hangmagasság, dallam, ritmus, hangsúlyozás stb.), és ezek alapján hozza meg a döntéseit. Az elgondolás a következő: ha a beszédadatbázis eléggé nagy, akkor minden esetben megtaláljuk benne azt a hullámformarészletet (valószínűleg elég hosszút, akár mondatot is), ami a lehető legjobban reprezentálja hangban a bemeneti szöveg aktuális részét. A másik eltérés, hogy változó hosszúságú elemekből épitjük fel a beszédet, olyanokból, amilyenek éppen a legjobban megfelelnek a szöveg adott pontjának, szakaszának. Alapvető cél ennél a technológiánál, hogy minél hosszabb hullámformarészt találjunk meg egy-egy keresésnél. Ha az esetek nagy százalékában találunk ilyen elemeket, akkor az eredeti bemondó hangját, stílusát fogja tartalmazni a felépített beszédjel, hiszen hosszabb beszédszakaszokat szinte változatlan formában adunk vissza úgy, ahogy azokat a hangfelvételnél rögzítették. Ez nagyon természetes, a bemondóra jellemző hangzást eredményez. A jó minőség záloga a nagy és jól előkészített beszédadatbázis, valamint a jó válogató algoritmus. Jelenleg ilyen beszédadatbázisok elsősorban kötött témakörhöz kapcsolódóan készülnek, mivel ezekre lehet ésszerű erőforrásokkal igényes beszédépítést megvalósítani (időjárás-jelentés, állomási utastájékoztatás, adott árucsoportokhoz tartozó árlista automatikus generálása szövegből, jegyrendelés stb.). A magyar nyelv esetében erős korlát a toldalékoló jelleg, valamint a mondatok szabad szórendje. Kötetlen szöveg felolvasására egyrészről nehéz olyan modellt megalkotni, amelyik optimális keresést biztosíthatna, másrészről nehéz olyan

beszédadatbázist létrehozni, amelyik lefedné a kötetlen bemeneti szövegben előforduló szövegelemek és azok prozódiájának nagy részét.

A gépi tanuláson alapuló beszédszintézis-technológia egy másik új irány. Ez az eljárás beszédkorpusz(ok)ból tanul, és a tanulás eredményét használja fel a későbbi beszédépítéshez (10.3.8. fejezet). A tanuláshoz több személy hangját is használhatja, annak biztosítására, hogy minél általánosabb tudásra tegyen szert az adott nyelv kiejtésével kapcsolatosan. Az ilyen technológiához készített beszédkorpuszok kialakítása sokkal több előkészítő munkát igényel, mint az előbb ismertetett, keresésen alapuló megoldásé.

8.2.3.1. A szintézis fő építőelemei

Olyan korpuszra adunk példát, amely adott témakörhöz összeválogatott szöveg felolvasásából alakul ki (tehát nincs korábbi, kész hanganyag), és a beszédszintetizátor ilyen témájú szövegekből fog majd szintetizálni a korpusz felhasználásával. A tervezés első lépésében el kell dönteni, hogy mi lesz a szintézis alapvető építőeleme. Célszerű választás: a szó. Ez azt jelenti, hogy zömmel szavakat, esetleg szófüzéreket fogunk keresni és felhasználni a beszédépítés során. Alsóbb szintű elemekről is gondoskodni kell, ha nem találunk szót. Ilyen esetben visszatérünk a diádos és triádos hangkapcsolatokhoz, ezeket keressük és ezekből állítjuk össze a hiányzó részeket, esetleg szavakat.

Prozódiai modellezés. Belső, beágyazott, komplex modellt alkalmazunk a prozódiai strukturák megvalósítására (lásd a 10.3.1.5. fejezetben). Ez a modell szoros összhangban van a későbbi szöveggyűjtéssel, és eltér a 10.3.1. fejezetben tárgyalttól. A modellben kihasználjuk a beszéd általános szerkezeti felépítését, valamint azt, hogy az ilyen szintézisrendszerekben csak kijelentő mondatokat kell szintetizálni (az eddigi gyakorlat szerint). A komplex prozódiai modell azt jelenti, hogy nincs kettéválasztva a szegmentális rész a szupraszegmentálistól, mivel az ilyen korpuszoknál cél, hogy utólagos jelfeldolgozást ne alkalmazzunk. A modell hasonló gondolatot követ, mint amit a számok felolvastatásához kidolgozott beszédadatbázisra bemutattunk (lásd a 8.2.1. fejezetben), azaz úgynevezett kezdő, belső és záró prozódiai elemekre osztjuk fel a mondatokat. Mindezt két szintre kell megtenni, szavakra és prozódiai egységekre (PRE). Ez utóbbiak szeparálására a központozási jeleket használjuk, valamint szövegelemzésből származtatott mintákat (prozódiailag egybetartozó kifejezéseket). Ez a megoldás párhuzamba állítható azzal a fonológiai dallammodellel (Varga 1994), amelyik a mondat dallamformáját karakterdallamokkal jellemzi. Az elmélet szerint minden karakterdallam hangsúlyos elemmel kezdődik. Az PRE-kre való felosztással tehát biztosíthatjuk, hogy a szintetizált formában benne lesz a mondat alapvető dallam- és hangsúlyszerkezete, valamint a

ritmikai és a hangintenzitási tulajdonságok is. Hangsúlyozzuk, hogy a modell nem valósítja meg az összes hangsúlyt a szintézis során, csak azokat, amelyek a PRE elejére jellemzők. A prozódiai modell határozza meg a szövegtervezést, annak a szövegkorpusznak az összeállítását, amelyiket fel kell majd olvasni.

Szövegtervezés. A szövegtervezés annyiban hasonló a korábbi elembázisoknál tárgyaltakhoz, hogy itt is előre meg kell határozni, hogy milyen lépésekben akarunk eljutni a bemeneti szövegből a véglegesnek tekintett beszédjelhez. A nyelvi anyag meghatározásánál számos kérdés merül fel. Milyen elvek alapján tervezzük meg a szöveget? Mit tartalmazzon, mennyi legyen az anyag? További kérdések lehetnek a konkrét hangfelvétellel kapcsolatosan is. Hogyan olvasson a bemondó? A felolvasásnál kell-e figyelni az intraindividuális változatosság minimalizálására? A szövegtervezés alapvetően két lépésből áll: szöveggyűjtés és abból a felolvasandó szöveg kialakítása.

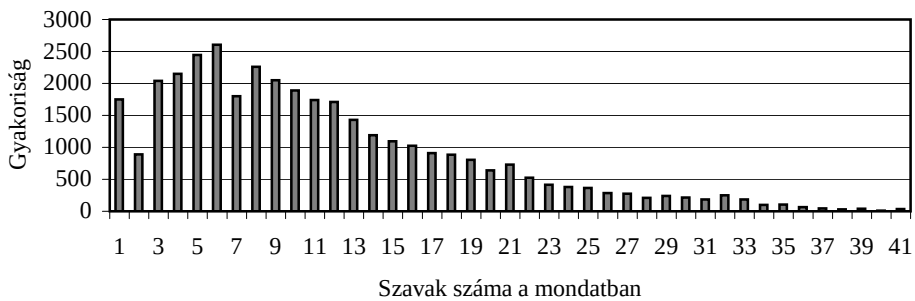
Szöveggyűjtés. Célszerű a témakört közel teljesen lefedő szövegkorpuszt gyűjteni. Ez lesz az alap, ebből alakítjuk ki a szintézishez használt szűkebb anyagot. Ha például időjárás-jelentések felolvasásához készítünk szövegkorpuszt, akkor ilyen témájú szöveganyagokat kell gyűjteni, lehetőleg mind a négy évszakból. Ha ügyfélszolgálati információkat akarunk a géppel felolvastatni, akkor akár több év ilyen anyagát célszerű összegyűjteni. Amennyiben egy termékcsaládhoz tartozó árlista felolvasása a cél, akkor az összes, ilyen céllal készült szöveget célszerű, akár visszamenőleg is összegyűjteni. Megjegyezzük, hogy vannak olyan esetek, amikor adott bemondó hangjára kell elkészíteni egy felolvasórendszert, oly módon, hogy a korábbi hangfelvételeit és azok szövegeit adják át a szintézist fejlesztő gárdának. Ilyenkor a szövegekkel csak olyan mértékig kell foglalkozni, hogy a szinkronitás biztosítva legyen a hang és a szöveg között hangszinten (lásd korábban).

A felolvasandó szöveg kialakítása. Amennyiben felolvasásból fogjuk elkészíteni a beszédadatbázist (nem kapott hangból dolgozunk), akkor az általánosan gyűjtött szöveganyagból egy szűkített változatot kell készíteni. Miért? Mert az ilyen gyűjtött szövegkorpuszok olyan nagy méretűek (50–60 000 mondat), hogy emberi bemondással fizikai képtelenség az összes mondatot elkészíteni (de nincs is rá szükség a nagy redundancia miatt). A szűkített változat elkészítését más is indokolja. Túl nagy lenne a keresési tér, és képtelenség lenne végigmenni rajta egy-egy mondat szintézisénél (nagy időkésleltetés lenne, ami információs rendszerekben nem megengedett). A nagy szövegkorpuszból tehát mindenképpen kisebbet kell csinálni. Elemezni kell a mondatokat és ki kell választani azokat, amelyek jó hatásfokkal lefedik az adott témakört, mind szegmentális, mind szupraszegmentális szinten. Ennek a szoftveres munkának a jellemző lépéseit tekintjük át a következőkben.

Szóalakok számának meghatározása. Fontos tudnunk, hogy hányfajta szóalak (lásd a 4.4. fejezetet) fordul elő a szövegben (hiszen a szintézis alapeleme a szó lesz), azoknak milyen az eloszlási statisztikája. Egy ügyfélszolgálati résztémakör például már 2000 szóalakkal jól lefedhető.

Mondatelemzés. A nagy szövegkorpuszból célszerű felmérni, hogy milyen a mondatok hossza, milyeneket kell majd a legnagyobb valószínűséggel szintetizálni. A 8.17. ábrán bemutatunk egy ilyen vizsgálati eredményt (Németh et al. 2006). Látható, hogy a szövegkorpuszban elhelyezett mondatok leggyakrabban 5–10 szót tartalmaznak, és a 20 szó felettiek előfordulása ritka.

A felolvasási szöveg meghatározása a prozódiai egységekre vonatkoztatva. A célkitűzés a következő: határozzuk meg a mondatok belső tagolási jellemzőit. Ennek jó becslése biztosítja a szintézisnél a jó prozódiai szerkezet megvalósítását (dallam, hangsúly, ritmus, intenzitás). A prozódiai egységek szintjeinek figyelembevételével a mondatokat és a szavakat sorrendbe helyezzük. A teljes fedés biztosítása a cél, tehát az, hogy olyan minimális számú mondathalmazt állítsunk össze, amelyik az összes szóalakra az összes prozódiai egységet tartalmazza legalább egyszer. Ezt nevezzük fedőhalmaznak. Az ilyen halmazlefedési problémának egy jól ismert közelítő megoldása a mohó algoritmus (van Santen–Buchbaum 1997). Minden



8.17. ábra. Mondathossz-gyakoriság a szavak számának függvényében egy szövegkorpuszban vizsgálva

lépésben egy mondatot adunk hozzá a fedőhalmazhoz. Az első kiválasztott mondat az, amelyik a legtöbb új egységet tartalmazza, utána minden lépésben azt a mondatot adjuk hozzá, amelyik a legtöbb, még lefedetlen egységet tartalmazza. Ezt addig ismételjük, amíg van lefedetlen egység. Az így kialakított szövegkorpusz kisebb lesz, mint az eredetileg gyűjtött anyag. A gyakorlati tapasztalat azt mutatja, hogy 5–10 000 mondat felolvasása még reális ráfordítással elvégezhető.

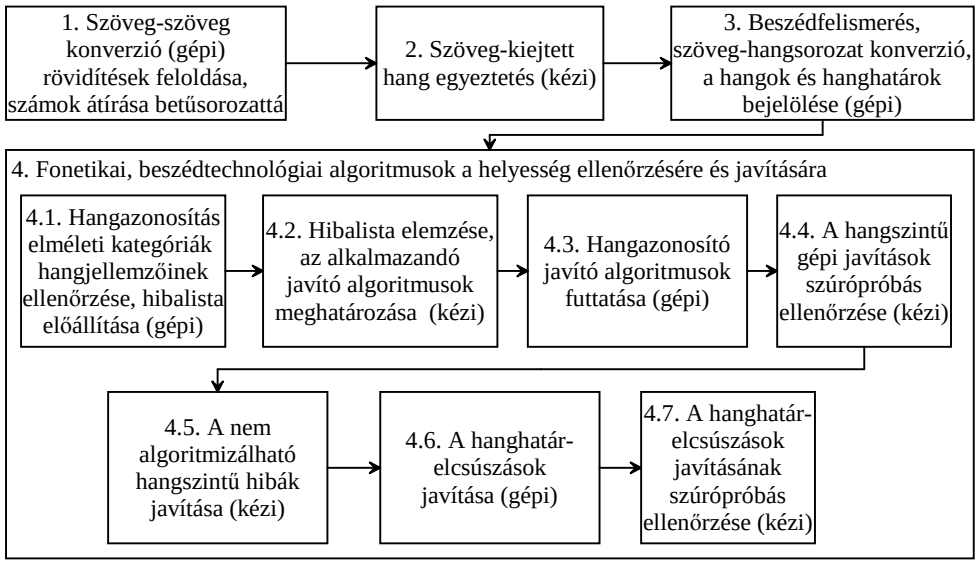
Hangátírás. A beszédkorpusz hanganyagában való keresés alapfeltétele, hogy ismertnek tekintjük a hullámformában lévő beszédhangok sorozatát. A hangátírás

két lépcsőben zajlik. Az elsőben az ortografikus formából készítünk szövegnormalizált fonemikus átírást (csak szavak fognak szerepelni a szövegben) a hanghullám tartalma alapján. A második lépcsőben ebből készül el a fonetikai hangátirat. A fonemikus átírásnál a szöveget alakítjuk a hanghoz. A felvett hangot meghallgatással összevetjük az írott szöveggel, amiből a bemondó felolvasta a szöveget. Ezt csak emberi közreműködéssel lehet megtenni, és ez a legmunkaigényesebb része az adatbázis elkészítésének. A munka lényege, hogy a szöveget fonemikus változatúvá alakítjuk, vagyis a hangfelvételen szereplő hangsorozatot vesszük alapul, ehhez alakítjuk át a szöveget (ha kell). Például a *32 C fok* szövegrészt *harminckettő celziusz fok* -ra javítjuk a szövegben, hiszen a bemondó így ejtette. A fonemikus szöveg képezi az alapját a fonetikai átírásnak, amelyet már algoritmus végez. Az elkészített fonetikai hangátirattal szemben szigorú követelmény, hogy annyi hangot kell tartalmaznia, amennyit a hullámforma. Ez biztosítja a szinkronitást a keresésnél.

8.2.3.2. A beszédadatbázis címkézése

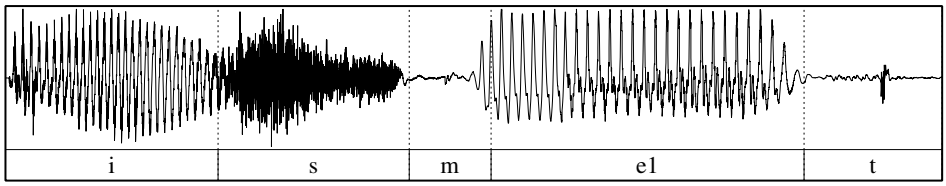
Olaszy Gábor–Bartalis Mátyás

A korpuszt mindkét technológiához fel kell címkézni, belső jelzésekkel kell ellátni. Ezeket a munkákat a beszédfelismeréshez kidolgozott gépi technológiák segíthetik. Ilyen címkék lehetnek a hanghatárjelzések, a zöngeszinkronjelek, a zöngés/zöngétlen váltási pontok; a szavak szintjén a szóhatárok, a hangsúlyos/hangsúlytalan megjelölés, a szótagok száma a szóban, a szó elhelyezkedése a szövegben, a szó határainak kategorizálása kapcsolódási vonatkozásban (átnyúlik a következő szóba); a szöveg szintjén pedig a prozódiai egységek határai és hosszuk (szószám), sorszámuk, elhelyezkedésük a mondaton belül. A beszédben talált nem beszéd elemeket is célszerű jelölni, hiszen a természetes felolvasásnál ilyenek is előfordulhatnak (csettintés, krákogás, glottalizált beszédrészlet). A nagy méretű beszédadatbázisokat nem lehet kézzel címkézni. Külön technológiát kell kialakítani a hang- és szóhatárok megállapítására. Ezen a ponton összeér a gépi beszédfelismeréshez készített beszédadatbázisok és a szintézis számára kialakítandók technológiai támogatása. Az optimális címkézés két fő lépését és azok belső műveletsorait a 8.18. ábra mutatja. A következőkben végigmegyünk az ábra egyes műveleti blokkjain. A szintézishez készülő beszédadatbázist beszédfelismerésre kidolgozott technológiával, az úgynevezett kényszerített beszédfelismeréssel címkézik (3-as blokk). Az eljárást a beszédfelismerés fejezetben ismertetjük. A kényszerített felismerés akkor sikeres, ha a hanghullám és a szöveg, valamint a fonemikus átirat közötti szoros, hangjelölési szinkronitás hibamentes (2-es blokk), azaz nincs eltérés a két hangszámban. A hanghatárcímkéket és a hangok szimbólumainak elhelyezését a hullámformával párhuzamosan a kényszerített felismerést végző algoritmus végzi (3-as blokk). Az eredmény, hogy a hullámformával párhuzamosan megkapjuk a

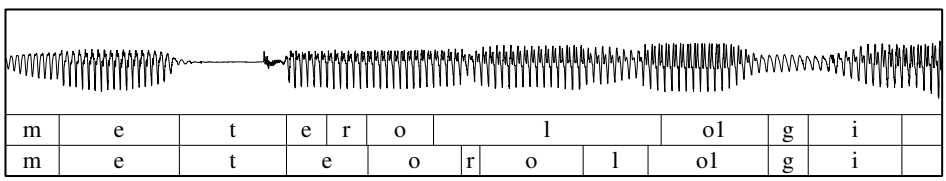


8.18. ábra. Beszédadatbázis címkézésére szolgáló technológiai műveletsor

hangsor hangjainak helyét meghatározó időadatokat és a hangok szimbólumait. A szóhatárok meghatározása már külön lépésben történik. Itt visszanyúlunk a fonemikus szövegformához, amelyben megtalálhatók még a szavakat elválasztó szóközök. A fonetikus és fonemikus változat következetes egybevetéséből meghatározhatók a szóhatárjelölések is a hullámformában. Itt ki kell térni egy problémára. A szóhatár a beszédjelben nem értelmezhető olyan egyértelműen, mint a leírt szövegben. Az egybefüggő hangfolyamban vannak olyan esetek, amikor a szóhatárok egymásba csúsznak (ugyanazon hang képviseli az előző szó utolsó hangját és a következő szó első hangját is). Például, a ... *legjobb barátom*... szókapcsolatban a kiejtésben a szóhatáron csak egyetlen [b:] hangot ejtünk, amelyik mindkét szóhoz tartozik. A probléma egyfajta megoldásáról szóltunk a 4.5. fejezetben. A gépi algoritmus hatásfoka a hanghatárok tekintetében csak 80–90%-os, a többi esetben hibázik. A hibázás oka több tényezőre vezethető vissza, ami összefügghet a beszélővel, az alkalmazott beszédfelismerővel, annak tanítási részleteivel. Ezeket a hanghatárhibákat célszerű megkeresni és javítani. Példaképpen bemutatunk néhány hibás címkézést. A 8.19. ábrán mutatott példában a koartikulációs néma fázis miatt döntött rosszul a gépi felismerő. Hibás jelölést okozhat a szinkronitásbeli eltérés is (8.20. ábra). A szinkronitási pontatlanságok esetén csak manuálisan lehet a hibát kijavítani. A szinkront helyre kell állítani és a kényszerített felismertetést újból el kell végezni. Másfajta hiba, amikor a felismerő egyszerűen rosszul dönt. Erre mutat példát a 8.21. ábra. Az optimális szintézishez az összes címkézési hibát ki kell javítani. Ez egy nagy méretű beszédadatbázisnál több tízezer javítást is jelenthet,



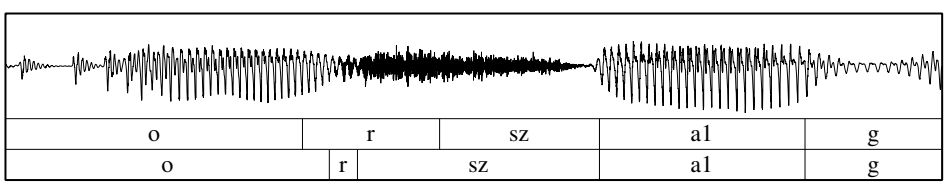
8.19. ábra. Az [m] hang jelölése az *ismét* szóban hibás, mert a gép a hangot a koartikulációs néma fázisra helyezte el. Valójában következő zöngés rész első nyolc periódusára kellett volna jelölni



8.20. ábra. Példa a szinkronitási hibából eredő gépi (felső hangsor) és a helyes kézi hanghatárjelölésre (alsó hangsor). A hiba oka, hogy a szövegben kimaradt a *meteorológia* szó harmadik magánhangzója, a fonemikus átiratban (középső jelölés) tehát csak 11 hang képviselte a szót, a hanghullámban pedig 12, hiszen a beszélő átsiklott a betűhibán és helyesen ejtette a szót

amit kézi módszerrel nem lehet elvégezni. Egy további, második lépcsős javító algoritmust célszerű kialakítani, hogy megtalálhassuk a hibás címkézéseket és kijavítsuk azokat (Olaszy–Bartalis 2008). Ezt a 8.18. ábra 4-es blokkja képviseli.

Algoritmus a címkehibák megkeresésére. A célkitűzés az, hogy a címkézés hibamentessé váljon. Ezt három további lépésben lehet elérni. Az első vizsgálat során megkeressük a gépi algoritmus által készített címkesorban azokat a helyeket, amelyek vélhetően hibásak (4.1-es blokk), a második lépésben olyan célzott algoritmusokat alkotunk, amelyekkel automatikusan ki tudjuk javítani ezek nagy részét (4.2-es blokk), a harmadik, befejező lépés a kézi javítás, amikor azokat a hibás jelöléseket vesszük szemügyre, amelyikeknél az automatikus javító nem tudott döntést hozni (4.5-ös blokk). A teljes javítóprogramnak olyannak kell lenni, hogy a végső kézi javításra csak kevés címke (pár száz) maradjon. A módszerből adódik, hogy a 4.1-es műveletsor végén ténylegesen meg lehet mérni, hogy a kényszerített felismeréssel készített, automatikus címkézés hatásfoka hány százalékos.



8.21. ábra. A gépi hangjelölés (felső hangsor) a pergő hangot rosszul jelölte. A helyes jelölés alatta látható

Modell a hibás címkék gépi megkeresésére (4.1-es blokk). Kiindulási adatként szolgálnak a gépileg elhelyezett címkék a hangsorban, valamint a fonetikai átirat. Ideális esetben a címke és a tényleges hanghatár ugyanazon a helyen van az időtengelyen (bizonyos hibahatáron belül). E két adatból fonetikai jellegű modell építhető, amely a következő elméleti paramétereket használja: gerjesztési mód, hangintenzitás, a beszédhangokra vonatkozó szerkezeti formák, specifikus hangidőtartamok a hangkörnyezet függvényében (Olaszy–Bartalis 2008). Az ellenőrzéskor a hanghullám összes beszédhangját összehasonlítjuk (a két hanghatár közötti részt) a hozzárendelt hangszimbólummal. Az összehasonlítást a hangokra összeállított fonetikai kategorizálás alapján tesszük meg (8.11. táblázat). A kategorizálásnál az

8.11. táblázat. A magyar beszéd hangelemeinek osztályozása hibás hanghatárjelzések feltárására. A hangokat a betűjelükkel jelöltük. V=magánhangzó; h*=[fi] hang; sil=szünet, illetve értelmezhetetlen hangelem

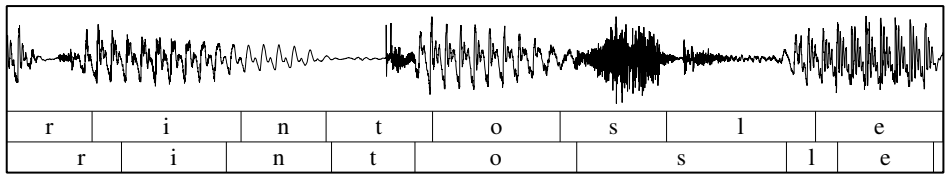
Kategória	1	2	3	4	5	6	7	8	9	10	11
Hang és tulajdonságai	V és j	b d g gy	ny	p t k ty	m n v l h*	h f	z zs	sz s	c cs	r	sil
Zöngés	X	X	X		X					X	
Zöngétlen				X		X		X	X		X
Vegyes gerj.							X				
Intenzív	X										
Közepes			X		X		X	X	X	X	
Gyenge		X		X		X					X
Egyszerű szerkezet	X				X	X	X	X			
Összetett szerkezet		X	X	X					X	X	

oszlopok azokat a hangokat reprezentálják, amelyek a megjelölt hangtulajdonságok szerint egyformának tekinthetők, illetve közel esnek egymáshoz. A táblázatban megjelölt hangtulajdonságok, valamint az időtartamok gépi eszközökkel jól mérhetők a beszédhullámon. Az algoritmus működése a következő: megkapja a vizsgálandó hang szövegből származtatott szimbólumát, és ezzel egy időben a táblázati tulajdonságok és az időtartamadatok alapján mérést végez a hullámformán, a hangot reprezentáló részen. Ha a táblázati besorolás és a talált, mért tulajdonságok legalább 75%-ban megegyeznek, akkor jónak ítéli meg a hangjelölést (ettől még kisebb hanghatárelcsúszás lehet benne). Ellenkező esetben hibát jelez. Vegyük sorra az előbbi ábrákon bemutatott hanghibák megtalálásának folyamatát. A 8.19. ábrán az [m] hangot a beszédfelismerő olyan szakaszra jelölte be, amelynek intenzitása kicsi és gerjesztése sem zöngés. Ez hibás jelölés. A 8.21. ábrán az [r] hang rossz jelölését a specifikus időtartamon alapuló mérés döntötte el (5.1.1.1. fejezet). Ebben a mássalhangzókapcsolatban a pergő hangra jellemző maximális időtartam 40 ms, a réshangra pedig 122 ms körüli. A kettő közötti arány tehát mintegy háromszoros. A beszédfelismerő által megadott jelölés ettől lényegesen eltér, ezért a vizsgáló algoritmus ezen a ponton hibát jelez. A 8.20. ábra hibás jelöléseit is a specifikus

hangidőtartamokkal való összevetés alapján lehet megtalálni. Az [l] hang időtartama sokszorosa az elvártnak. A címke- és hangellenőrző szoftver hibalistát generál, amelyben megadja a hiba hangsorbeli helyét, hangkörnyezetét és fajtáját. A listában sokféle hiba szerepelhet, számuk általában nagy. A technológiai folyamathoz tartozik, hogy ezt a listát manuálisan kell ellenőrizni, a hibákat kategóriákba kell sorolni, majd a javító algoritmusokat ezek alapján kell egyenként kialakítani. Kétfajta hibát különböztetünk meg: rossz hang, illetve hangon belüli elcsúszás.

Modell a címkézési hibák korrekciójára (4.2.–4.7. blokkok). Az eddigi hibafeltárással elértük azt az állapotot, hogy tudjuk a vélt hibák helyét. Innen léphetünk tovább a javításokra. Ennek két lépése van. Az elsőben a rossz hangjelöléseket szüntetjük meg (durva hiba). Ha egyfajta ilyen hibából sok van, akkor megpróbálunk algoritmust tervezni a javításra (4.2-es blokk). A kialakított javító algoritmusok külön programcsoportot alkotnak (4.3-as blokk), segítségükkel javítjuk a hibákat, majd ellenőrzést is végzünk (4.4-es blokk). A maradék hanghibákat kézi javítással szüntetjük meg (4.5-ös blokk). A javítás második lépcsőjében a hangon belüli hanghatárelcsúszásokat javítjuk (4.6-os blokk). Milyen hanghibákat lehet ilyen módon javítani? Például olyanokat, amelyek a zöngés-zöngétlen átmenetek határán jelentkeznek olyan formában, hogy a tényleges átmeneti ponttól (ami 10–15 ms-os sávban helyezhető el) távolabb van megjelölve a hanghatár. A 8.22. ábrán mutatunk be egy ilyen hibát. A javító algoritmus az ábrán bemutatott esetre az [oʃ] hangkapcsolatra a következő: A zöngétlen [ʃ] hang bal oldali hanghatárát jobbra eltoljuk (mivel belelógott a zöngés hangba) addig, amíg az utolsó zöngeperiódus végét meg nem találja. A zöngétlen hang jobb oldali hanghatárát pedig az őt követő zöngés hang első zöngés periódusának kezdetére állítjuk be. Az ilyen javítások helyességét szúrópróba-szerűen ellenőrizni kell. A második lépcsőben elvégzett javítások pár órás programozást és kézi javítást igényelnek, tehát gazdaságosnak mondhatók. Mindezen műveletsor után a beszédadatbázisunk szegmentális hangcímkéi 99%-os biztonsággal korrektnek mondhatók, az adatbázis alkalmas a szintézis kiszolgálására. Összesen 18-fajta fonetikai javító algoritmust dolgoztunk ki (több beszédadatbázis alapján) a hangjelölési hibák csökkentésére (4-es blokk). Ez a szám változhat a beszélő személy és a felhasznált beszédfelismerő algoritmus függvényében.

Összefoglalásként elmondhatjuk, hogy a nagy méretű beszédadatbázisok készítése sokrétű felkészültséget kíván és időigényes tevékenység. A befektetett munka eredménye, hogy jó alappal fogunk rendelkezni a keresésen alapuló beszédszintézishez, illetve a gépi tanulás is sikeresebb lesz ilyen adatbázisok használatával (például a HMM alapú beszédszintézishez).



8.22. ábra. Példa a gépi hanghatárjelölés hibájára (felső hangsorjelölés) és annak fonetika algoritmussal való javítására (alsó hangsorjelölés)

8.3. Kiejtésikivétel-szótárak

Olaszy Gábor

Az elektronikus kiejtésikivétel-szótárak elsősorban a gépi szövegfelolvasás és a beszédfelismerés számára nélkülözhetetlenek. Ilyen szótárakban adhatók meg a kiejtési szabályokkal nem leírható szövegelemek. Alapvető kérdés, hogy mi számít ilyen szövegelemnek? Itt most csak a nyelvi kategóriába sorolható elemeket vizsgáljuk (nem vizsgáljuk a karaktersorozatok egyéb fajtáit, amik nem nyelvi információt hordoznak, például webcímek). A fő kívánalom, hogy a szintetizátor a leírt szöveg minden elemét (mondjuk szavát) a kiejtésnek megfelelő formában olvassa fel, olyan kiejtéssel, amit az anyanyelvi beszélő megért. Ezt a kiejtési formát a gépi szövegfelolvasás folyamán sok esetben nem egyértelmű meghatározni (a gépeknek még nincs olyan intelligenciája, mint például egy rádióbemondónak). A jó hatásfokú gépi felolvasáshoz azt kell biztosítani, hogy az idegen eredetű szavak, személy- és cégnevek, szakkifejezések, földrajzi nevek, betűszavak kiejtésének korrekt feloldása álljon rendelkezésre a szintetizátor bemenetén (idetartozik például az ékezet nélküli szavak helyes formájának a visszaállítása is). A beszédtechnológia számára ilyen szótárak még nem készültek átfogó megoldásban, az egyes alkalmazásokhoz a fejlesztők saját kiejtésikivétel-szótárakat fejlesztenek. A beszédszintézis legnagyobb ellensége, ha olyan betűsorozat kerül a szintetizátor bemenetére, amelyiknek a hangzó formája az adott nyelven nem jelent semmit. Ilyenkor a hallgató beszédmegértési rendszere megakad, és ez befolyással van a mondat további megértésére is. Ezért a kivételszótárak a beszédszintetizáló rendszerek fontos elemei. Az ilyen szótár szerkezete függ a nyelvtől is. A magyar esetében gondolni kell arra is, hogy a toldalékos alakokat hogyan fedje le a szótár.

A kiejtésikivétel-szótárak felépítése. A szótár alapszerkezete egyszerű, megadja az ortografikus formát, majd megadja a betűsorozat kiejtési formáját (valamilyen hangreprezentációban). A magyar esetében egyetlen különleges hangjelölést kell bevezetni a korrekt átíráshoz, ez az [aː] hang rövid változata, amelyre speciális betűkaraktert kell kijelölni (ilyen lehet például az A karakter). A továbbiakban gondolni kell az olyan esetekre is, amikor a kiejtést a hangot képviselő karakterek

nem határozzák meg egyértelműen (ez tipikusan morfémahatáron fordul elő *malacság, hallászavar*. Ilyenkor két megoldás van. Vagy jelöljük a morfémahatárt, vagy olyan a hangjelölést alkalmazunk amelyik egyértelműen meghatározza a hangzást. Példákat sorolunk fel: *kg* = kilogram, *Kossuth* = kosut, *MTA* = emtéaa, *Boeing* = bőing, *Aldacton* = aldakton, *PDA* = pédéaa, *Vodafone* = vodafon, *Bush* = bus, *Maciej* = mAcsej (A = rövid á), *Tóth* = tót, *ún.* = úgynevezett, *kb.* = körülbelül, *roaming* = róming, *tone* = tón.
A kiejtésikivétel-szótáraknak figyelembe kell venni a gyakorlati élet szituációit is. Sok alkalmazásban a bemeneti szöveg formája az adatbevivőtől függ. A kiejtési szótárnak mindenfajta szövegmegjelenést kezelni kell tudni. Néhány példát mutatunk be, hogy a Web and Walk elnevezésű készülék kiejtési feloldása hogyan van megadva egy tényleges kivételszótárban (lásd a 12.3.7. fejezetben), az adatbevivők által rögzített formák alapján. A gyakorlati megoldásról a 12.3.7. fejezetben szólunk.

```
(web'n'walk) = vebnvók
(w'n'w) = vebnvók
(web`n´walk) = vebnvók
(w´n´w) = vebnvók
(wnw) = vebnvók
```

A példából látható, hogy ötféle bemenetre kell felkészülnie az átíró rendszernek. Hasonló kategóriába tartoznak az elírási hibák, illetve az ékezet nélküli formában írt szövegek (számítástechnika, mobiltelefonok). Lássunk néhány példát.

```
(hir) = hír
(0-as) = nullás
(2-as) = kettes
(autómatikusan) = automatikusan
(hivasertesito) = hívásértesítő
(szamtiltas) = számtiltás
```

Hogyan használja a gép az ilyen szótárt? A szöveg első feldolgozása során összehasonlítja a mondat minden szavát a kivételszótár elemeivel. Ha talál egyezést, akkor behelyettesíti a karaktersorba a kivételszótár által definiált betűsorozatot, ha nem, akkor továbblép a következő szóra. A fenti példákból látható, hogy az élet minden területét átszövik az olyan betűsorozatok, amelyek igénylik a kiejtés meghatározását. Ezért egyelőre még nem képzelhető el egy olyan kivételszótár elkészítése, ami általánosan lefedi az összes lehetséges esetet a magyar nyelvre (más nyelvre is ugyanez a helyzet, talán az angolra a legkevésbé). Az ilyen szótárakból komoly mennyiségre lesz szükség a következő évtizedben.

Általános tartalmú kiejtésikivétel-szótár. Olyan nyelvi elemeket tartalmaz, amelyek általános szövegekben előfordulhatnak. Ilyenek a rövidítések, a mértékegységek,

a közkeletű márka- és cégnevek, a történelmi nevek egyes csoportjai, a földrajzi elnevezések stb. Az általános tartalmú kivételszótárak elkészítése általában sikeres lehet, ha elégséges szöveganyagot tanulmányozunk át, vizsgálunk meg. A mai korban erre már vannak nyelvtechnológiai szoftverek is. Az ember mint erőforrás azonban nem hagyható ki. A kiejtést csak ember tudja eldönteni. Magyarra nincs ilyen szótár (nyilvános formában), de van már rövidítési elektronikus adattár. Az általános tartalmú kivételszótárakat a fejlesztők többnyire (valamilyen szinten) beépítik az alkalmazásokba. A felhasználó oldaláról nézve a kívánatos az lenne, ha a fejlesztők megadnák a beépített kivételszótár listáját, amiből el is lehetne dönteni, hogy az adott alkalmazásnak elégséges-e ez, vagy bővíteni kell. Ezért nehéz megtervezni a teljes körű helyes kiejtés helyességi fokát egy adott, vásárolt rendszernél. Bemutatjuk a leggyakoribb elemeket, amelyeket ilyen szótárakba be szoktak építeni. Néhány példát adunk az eddigi tapasztalatainkból.
Általános rövidítések: *stb., ún., kb., dr., gysz., egyh., trv., gk., tgk., szgk*
Mértékegységek: *kg, dkg, mg, km, m, cm, mm, l, dl, cl, ml*
Informatikai kifejezések: *www, @, SMS, e-mail, Apple, Macintosh, USB, PC, Mbyte, kbyte, ADSL, ASCI, floppy, PDA, iPhone*,
Benzinkutak: *Agip, OMV, Shell*,
Autómárkák: *Peugeot, Fiat, Renault, Chrysler, BMW, Mitsubishi, Astra*
Helynevek: *BP.* (ütközik a BP benzinkúttal),
Általános idegen szavak kiejtésének feloldásai: *crescendo, adagio, concerto, hardware, station, city, tone, make up, lotion, sun, low, airline, aerobic, AIDS*
Cégnevek: *Albacomp, Pick, Novopharma, Graphisoft, Sony, MVM, Samsung, Siemens, LG, Panasonic, Toyota, Swietelski, Allianz, OTP, ÁNTSZ, Auchan, Tesco, Cora, Astoria* stb.
Családnevek*: Beöthy, Andrássy, Aschner, Babits, Benczúr, Bernáthffy*
Közterületek rövidítéseinek feloldása: *u., sz., hsz, hrsz., ltp, em.*
Márkanevek: *Cola, KFC, Pepsi, Mac., Vodafone, T-Mobile, Maybelline, Oriflame, Chanel, Amway*
Fantázianevek: *Sun Breeze*
Köznapi műszaki elnevezések: *CD, DVD, FM, AM, URH, Hz, kHz, PC, USB, Mbyte, kbyte, ABS, GPS, air bag*

A szakmai kiejtésikivétel-szótárak. Ilyen elektronikus szótárak még csak elvétve és elszigetelten léteznek magyar nyelvre (nem nyilvánosak). Egységesítésük és rendszerezett gyűjtésük a jövő feladata. Kérdés, hogy a nyelvészet körébe tartozik-e egy ilyen munka, vagy esetleg az informatikusi társadalom feladata. A jövőkép azt mutatja, hogy a sikeres, sokoldalú gépi szövegfelolvasási technológia nem képzelhető el egy nemzeti kivételszótár(-család) megalkotása nélkül. Azt is látni kell, hogy az ilyen szótárak állandó frissítésre szorulnak, tehát a szótár folyamatos karbantartására is biztosítani kell az anyagi és az emberi kapacitást.

A szakmai kivételszótárakat minden szakmára külön kell elkészíteni. Példákat sorolunk a teljesség igénye nélkül: matematika, biológia, mérnöki tudományok, informatika, nyelvészeti, orvosi szaknyelv, gyógyászati kifejezések elnevezése, gyógyszerek nevei, mezőgazdasági témakörök stb.

A kiejtésikivétel-szótárak megvalósítási formái. Kétfajta megközelítés szokásos. Az egyik, amikor rábízzuk a felhasználóra, hogy alkossa meg a saját alkalmazásához szükséges megoldást, ezt hívják nyílt kivételszótárnak. A másik megoldás, hogy a megrendelő és a fejlesztő együttműködve alakítja ki a kiejtési kivételek jegyzékét, azt beépítik a rendszerbe. Ezt nevezik zárt rendszerű kivételszótárnak.

Nyitott kiejtési szótárak. A nyílt rendszerű szótárak alkalmazása olyan megoldási formát kíván meg a beszédszintetizátor fejlesztőitől, ami biztosítja, hogy a szintetizátor belső értelmezési szabályrendszere és a kivételszótár között egyfajta metanyelv segítségével állandó kapcsolat van a futás során. A kivételszótár egy külső fájlként képzelhető el, amit a felhasználó állandóan feltölthet új adatokkal, méreti korlát nélkül. Ez magában hordozza a hibázás lehetőségét is (megsértjük a belső kiejtési szabályokat). A hibák jelzésére egy fordítóprogram szolgálhat, a javított kiejtési szótár csak akkor fog működni, ha azt a fordító szintaktikailag hibátlannak találta. Az ilyen szótárak alkalmazásához pontos működési utasítást kell a felhasználó számára biztosítani. Nyitott rendszerű kivételszótárt tartalmaz például az ingyenesen letölthető MultiVox4 nevű magyar szövegfelolvasó is (10.3.5.1. fejezet).

Zárt rendszerű kiejtési szótárak. Az ilyen megoldásoknál csak a fejlesztő tudja bővíteni, javítani a szótár tartalmát. Ez garancia arra, hogy a beszédszintetizátor belső működésébe semmiféle zavaró tényező nem kerül be. Ilyen megoldással működik a 12.3.7. fejezetben ismertetett alkalmazás. A szótárt a fejlesztő és a felhasználó rendszeresen, közös munkával tartja karban.

A kiejtési szótárak naprakész állapota. A kiejtésikivétel-szótár (legyen az nyílt vagy zárt) állandó karbantartást kíván, tehát nem lehet lezárni a fejlesztését. Csökkenteni lehet a fenntartására szánt anyagi és emberi forrást, de megszüntetni nem. A nyelv változik, új szavak keletkeznek, mások átalakulnak, az új nyelvi elemek egy részéhez meg kell határozni a kiejtési formát is. Ezért az ilyen szótárak napi (heti, havi) karbantartást igényelnek. Várhatóan fognak születni olyan szoftverek, amelyek meg fogják tudni jósolni a kiejtésileg problémásnak vélhető nyelvi elemeket, és ezek megkönnyíthetik a szótárt karbantartó személy munkáját is.

Példák magyar kivételszótárakra. Az utóbbi évtizedekben célzott alkalmazásokhoz fejlesztettek kiejtésikivétel-szótárakat. Tudomásunk szerint a ProfiVox

szövegfelolvasó technológiához készítették a legtöbb kiejtésikivétel-szótárt magyar nyelvre (10.3.6. fejezet). Ezek a következők: általános rövidítések (több száz), idegen szavak kiejtési formái (több ezer), cégnevek (több ezer), családnevek (több ezer), magyar keresztnevek (több száz), magyar települések nevei és gyógyszerek, orvosi latin kifejezések kiejtése (több ezer).

Néhány szót a feldolgozásokról. A cég nevének kiejtése sokszor nem egyértelmű. Sokszor maga a cég sem tudja eldönteni, hogy mi a hivatalos kiejtése a nevének (AM-IT Rt., CGNU Mébit stb.). A BME TMIT beszédtechnológiai laboratóriumában több százezer cég-, család- és keresztnév feldolgozását végezték el kiejtési szempontból. Speciális szoftvereket fejlesztettek a szelektálásokra. A fejlesztés célirányos volt, hiszen egy általános név- és címfelolvasó műszaki megoldáshoz kellett olyan szoftvert készíteni, amelyik Magyarország összes telefon-előfizetőjének a nevét és címét korrektül fel tudja olvasni (12.3.4. fejezet). A magyar települések nevei nagyrészt helyesen kiolvashatók az általános betű-hang átalakítási szabályokkal, de van néhány olyan eset is, amikor a kiejtést konkrétan meg kell adni. Ezeket az eseteket a fenti címfelolvasó-fejlesztésben feldolgozták és a működő rendszerben alkalmazzák. Néhány példa: *Vácszentlászló, Váchartyán, Kiskunlacháza, Ferencszállás, Zselicszentpál, Babarcszőlős, Bácsszőlős*. Más alkalmazáshoz készült a gyógyszerek neveit és az orvosi latin kifejezéseket definiáló szótár. Egy telefonos beszéddialógussal működő gyógyszerinformációs rendszer beszédfelismerő és beszédszintetizátor moduljában mintegy 5000 gyógyszer nevét és azonfelül a gyógyszer-tájékoztatókban található idegen eredetű (főleg latin) nevek feloldását tartalmazó kivételszótár működik. Feljesztését az Országos Gyógyszerészeti Intézet segítségével végezték (12.3.5. fejezet). A megoldás kiejtési szabályokat is tartalmaz, valamint tételes kivételszótárt is. A szabályokból adunk ízelítőt a 8.12. táblázatban. Végül megjegyezzük, hogy el-

8.12. táblázat. Példák a gyógyszernevek kiejtésének meghatározását segítő szabályokból

betű = hang	Feltétel	Példa
c = c	Ha a c után elöl képzett magánhangzó áll	CEDAX, CEFALEKOL, CEFAM, CEFANORM, CEFAZOLIN, DALACIN
c = k	Ha c után hátul képzett magánhangzó vagy bármilyen mássalhangzó áll, vagy a c szó végén van	CORVATON, COSOPT, OCUMETER PLUS, COTRIPHARM, COVEREX, COVEREX KOMB, COVIOGAL, COZAAR, AEROBEC, ANTIHEMOPHILIC FACTOR, BACTROBAN, BACTLOFEN
ph = f	Kivétel, ha összetett szóban a tagok határán található	LYMPHOGLOBULIN, COTRIPHARM
y = i		LYMPHOGLOBULIN, MEROMYCIN kivétel: Bayer
s = z		MAGNESIUM, METEOSPASMYL
s = sz		MAGNEVIST, MERISTIN, METEOSPASMYL
w = v		METHOTREXAT-EBEWE
oe = ö		OESOPHAGUS

készült a magyar szóalakok általános kiejtését megadó elektronikus kiejtési szótár is a BME TMIT fejlesztésében, amely 1,5 millió szóalak kiejtését adja meg többfé-

le hangszimbólummal (karakteres formában). Ez az adattár az általános használaton felül kiszolgálhat majd későbbi statisztikai fejlesztéseket is a magyar kiejtéssel kapcsolatban (8.4.3. fejezet).

8.4. Oktatási, kutatási célú internetes adatbázisok

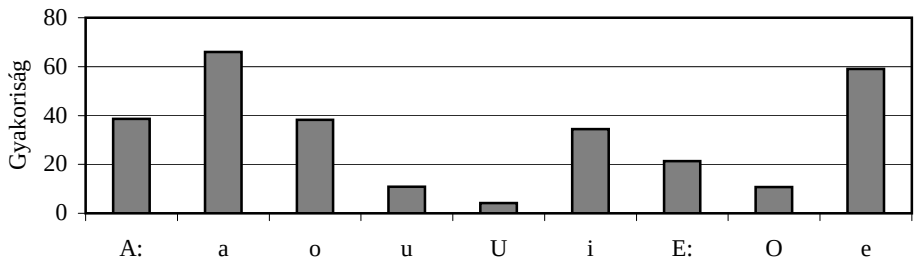
Ebben a fejezetben olyan adatbázisokat, adattárakat mutatunk be, amelyek nem konkrét fejlesztésekhez készültek, hanem a kutatást és az oktatást hivatottak általánosságban szolgálni. Többségük nyilvános, ami azt jelenti, hogy közvetlenül az interneten is használhatók, illetve kérés esetén kutatáshoz hozzáférhetők. Ezeknek az adattáraknak fontos szerepe van a felsőoktatásban, hiszen segítségükkel feladatok írhatók ki, mérések tervezhetők, gyakorlati ismeretek átadhatók.

8.4.1. A magyar hangkapcsolódások akusztikai bemutatása szavakban

Olaszy Gábor–Abari Kálmán

A magyar beszéd akusztikai szerkezetének első nyilvános, internetes bemutatására 2006-ra készült el egy olyan beszédadatbázis, amely szisztematikus anyaggyűjtés eredménye (`http://fonetika.nytud.hu/cccc`), és egy férfi és egy női bemondó hangján tartalmazott mintaszavakat. Ennek szűkebb célja az volt, hogy az oktatás és kutatás számára adjon támpontot a beszéd szegmentális elemeinek a vizsgálatára. Az adatbázis tartalmát használni lehet konkrét mérésekhez, és általános magyarázatokhoz is. Ezt a beszédadatbázist csak a mássalhangzó-kapcsolódások bemutatására készítették, de a mintaszavak egyéb hangkapcsolatai is tanulmányozhatók. Összesen 897-féle mássalhangzó-kapcsolódásra tartalmaz mintaszavakat (Olaszy–Bartalis 2008). A vizsgálható mássalhangzó-kapcsolatok részletezve a következők: minden CC kapcsolat: 373-féle; CCC: 445-féle; CCCC: 74-féle és öt eleműre 5 mintaszó. Az adatbázis szavainak akusztikai szerkezete képi formában tanulmányozható és szöveges magyarázat is tartozik minden mássalhangzó-kapcsolódáshoz (a mintaszó egésze, illetve kijelölt részei hangban is meghallgathatók). Később elkészült ennek az adatbázisnak a bővített változata (Abari–Olaszy 2007), amely már lefedte az összes hangkapcsolódási alapformát (`http://fonetika.nytud.hu/cvvc`). Ez utóbbi tartalmazza az összes CV, VC kapcsolatot, az egy- és többelemű magán- és mássalhangzó-kapcsolatokat, továbbá példákat tartalmaz a hiátustöltés jelenségére is. Ez az adatbázis 1400 különböző mintaszót tartalmaz. Mindkét adatbázisban ugyanaz a szó férfi és női ejtésben is megtalálható. A mintaszavakat szólistaként olvastattuk fel (minimális intonációval). A bemondók átlagos artikulációs sebessége 10,5 hang/s ebben az adatbázisban. Minden mintaszó önmagában hordoz több

hangkapcsolatot is. Így alakult ki a lefedettségi adatsor (lásd később), vagyis az, hogy egyes hangkapcsolatokra több mintaszóban is van példa, másokra viszont csak egyben. Egy-egy szó a hangsorából adódó hangoknak és azok összekapcsolódásainak akusztikai bemutatására szolgál. A hangkapcsolatok 9 magánhangzó és 25 mássalhangzó (plusz a hiátustöltés, amit j+ hangszimbólummal jelöltünk) kombinációit jelentik. A szisztematikusan felépített hangkapcsolati adatsort nem terjesztettük ki a hosszú hangokra, azok csak esetlegesen fordulnak elő. Az ismertetett két beszédadatbázis minden mintaszóra a következő adatokat tartalmazza: a szó szöveges (karakteres) alakja, a szót alkotó hangsor hangjainak szimbólumai (E1-hangjelöléssel, például *gyáros* = GA:roS), a hangidőtartamok ms-ban, a szó hullámformája, a hangokhoz tartozó hanghatárok jelei minden akusztikai diagramon, valamint a szó spektrogramja és intenzitásgörbéje. Ezek az adatok lekérdezésekkel válnak láthatóvá. Az átlagos lefedettségi adatokat a CV és VC kapcsolatok magánhangzóira a 8.23. ábra mutatja. Az adatbázisban tehát a legtöbbet



8.23. ábra. A CV és VC kapcsolódások átlagos darabszáma a cvvc jelű beszédadatbázisban a magánhangzókra vonatkoztatva

az [ɔ] és [ɛ] magánhangzóval megvalósuló CV, illetve VC kapcsolat fordul elő.

Az adatbázis szolgáltatásai. Az adatbázis használatát egy erre a célra fejlesztett, felhasználóbarát kereső és megjelenítő program könnyíti. A keresővel kikereshetők a kívánt hangkapcsolatokat tartalmazó mintaszavak, a megjelenítővel a szó hullámformája és az akusztikai képek tehetők láthatóvá (spektrogram, intenzitásmenet, hanghatárok). A szót felépítő beszédhangok szimbólumai az akusztikai diagramok alatt láthatók a szóhatárjelzések között. Az akusztikai diagramokról fizikai adatokat is megkaphatunk (formánsfrekvenciák, intenzitásértékek, hangidőtartamok stb.). Ilyen módon más vizsgálatokhoz is gyűjthetünk anyagot. Két szó akusztikai diagramjai is összehasonlíthatók (egymás alatt jelennek meg a képernyőn), ilyenkor vizuálisan vizsgálhatók az egyezések, különbségek. Ezzel az adattárral egységes hangzó és képi anyag áll a kutató/oktató rendelkezésére akár célzott kutatásokhoz, akár képi információkra támaszkodó oktatáshoz (spektrogram, hullámforma, hangátmenetek stb.) is.

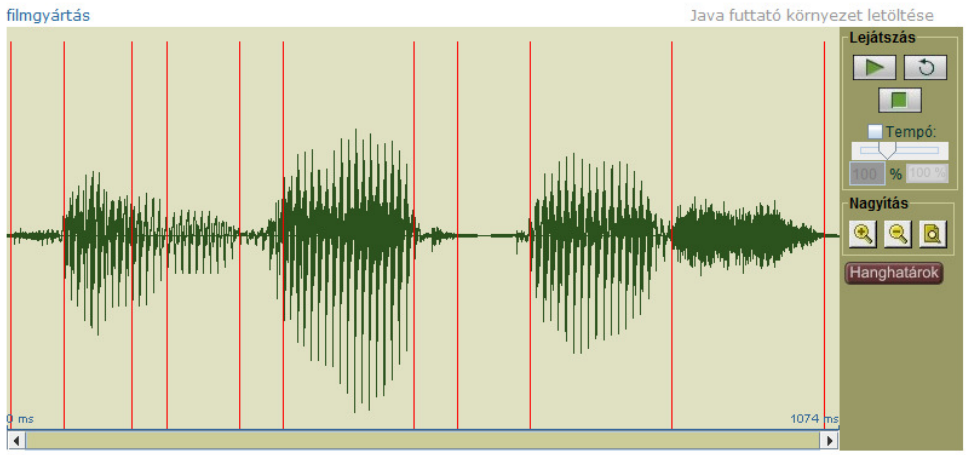
Keresés adott hangkapcsolatra. A megjeleníteni és vizsgálni kívánt hangkapcsolatot hangalapú kereséssel, illetve szövegforma-megadással találhatjuk meg az adatbázisban. A keresés szűkítését, bővítését a következő karakterek segítik: a kezdő, illetve hangsorzáró pozíciót a kettős kereszttel adhatjuk meg (#), a csillag (*) több tetszőleges elemet, a kérdőjel (?) egyetlen tetszőleges elemet jelent.

Ha a betűalapú keresési ablakba a #rek* szekvenciát adjuk meg, akkor minden olyan szót megkapunk, ami *rek* betűkkel kezdődik (*rekord, reklám*). A keresett betűkapcsolat állhat több betűből is. Például a *gyako* betűsorozat megadásakor minden olyan szót megkeres a program, amelyikben ez a karaktersorozat ilyen sorrendben előfordul. Így akár szótagokat is kereshetünk.

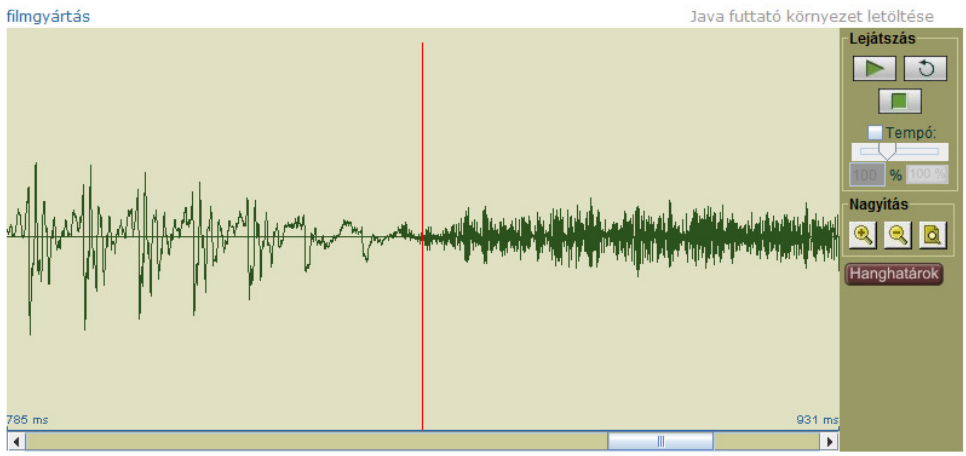
Hangalapú lekérdezéshez egy hangkapcsolatsorozat hangjait kell megadni a kereső ablakban. A hangmegadás könnyítéséhez segítséget nyújt a megjeleníthető hangkódtábla (IPA-jelekkel), amiből kattintással kiválaszthatjuk a kívánt hangokat (azok bekerülnek a kereső ablakba). A keresés eredménye egy szólista, amely ábécésorrendben jelenik meg.

Az akusztikai bemutatás módja. A hangkapcsolódások akusztikai szerkezetét a koartikulációs sajátosságok szabják meg. Így az akusztikai diagramok tanulmányozásával következtethetünk az artikulációs mozgásokra és a gerjesztésváltásra is. Az akusztikai diagramok bemutatása két szinten kérhető az adatbázisból. Az első szinten a részletes hangrezgés tanulmányozható; a hanghatárok megjeleníthetők, a hangsor, illetve a kijelölt része meghallgatható, a másodikon az akusztikai részletekről kaphatunk információt a megjelenített hangspektrogram, az intenzitásgörbe, az időfüggvény tanulmányozásával. Részletes hangrezgés megjelenítéséhez a rezgésképet úgy kaphatjuk meg, hogy a szólista kiválasztott szavánál a [Részl.] feliratra kattintunk, a hangrezgés megjelenik egy külön ablakban és interaktív módon tanulmányozható (8.24. ábra). Az itt alkalmazott feldolgozóprogram lehetőséget ad a felhasználónak, hogy széthúzza az időtengelyt, belelásson a hangok rezgésformájának legapróbb részleteibe (8.25. ábra), a kijelölt részeket meghallgassa, folyamatosan ismétlődő meghallgatást kérjen stb. A hanghatárok megjelenítése a hullámformán segítheti az artikulációs mozzanatok akusztikai következményeinek tanulmányozását. A mintaszó lejátszási sebességének változtatására is lehetőség van (Tempó felirat), valamint a kijelölt hangrész (például egy magánhangzó-periódus) folytonos, ismételt lejátszására is a visszafelé görbülő nyílra kattintva. Ezzel hangzó formában is tanulmányozható a zöngés hangok átmeneti fázisának periódusonként változó akusztikai tartalma, vagyis az akusztikai vetület fázisainak meghangosítása érhető el.

Akusztikai részletek. Az akusztikai diagramok megtekintéséhez az „Akusztikai részletek vizsgálata” feliratú ablakba kell „áttenni” a vizsgálni kívánt szót (a



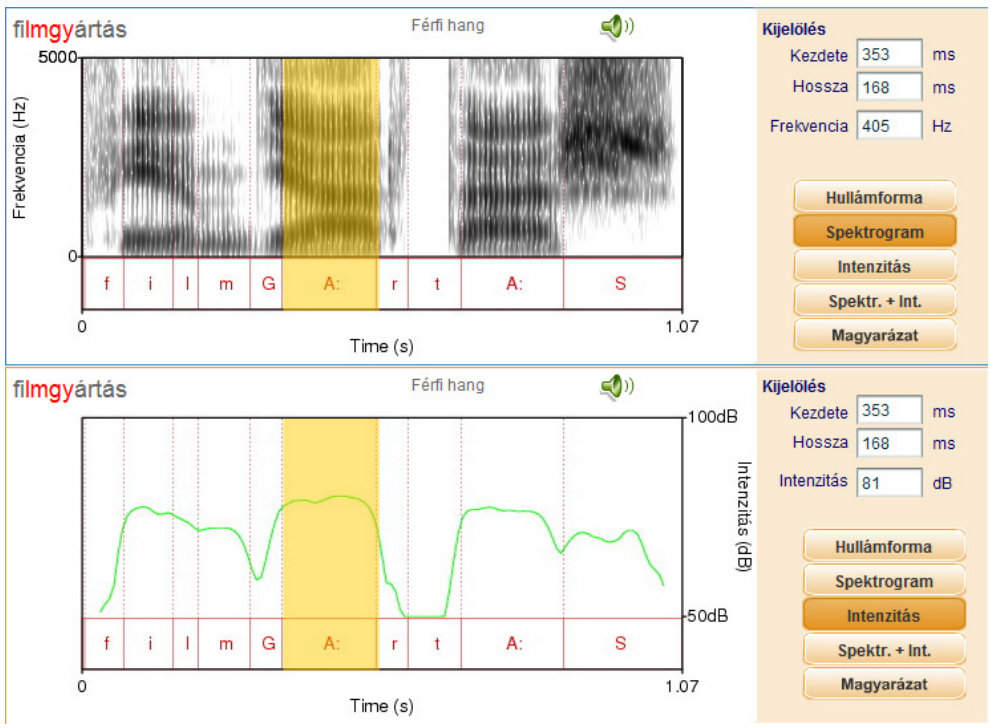
8.24. ábra. A *filmgyártás* mintaszó (1070 ms) rezgésformája a megjelenített hanghatárokkal



8.25. ábra. A *filmgyártás* mintaszó utolsó két hangjának összekapcsolódási pontja nagyított képen. A hanghatárt a függőleges vonal jelöli

[Vizsg1.], illetve a [Vizsg2.] feliratokra való kattintással). Két szót kell elhelyezni ebben az ablakban (egymás alatt jelennek meg). Ezután négyfajta diagram hívható elő. Alaphelyzetben a hullámforma jelenik meg. A Spektrogram gombra kattintva a program a szó frekvenciaszerkezetét mutatja meg. Az intenzitás is megjeleníthető, ilyenkor a mintaszó hangjainak erőssége válik láthatóvá az idő függvényében. A negyedik diagram a spektrogramot és az intenzitást mutatja ugyanazon a képen. A CCCC adatbázis használatakor a CC-kapcsolatok esetében magyarázatot is kaphatunk az adott hangkapcsolatról a Magyarázat gombra kattintva. A Hangszóró

ikonra kattintva a mintaszó meg is hallgatható. Az 8.26. ábrán a *filmgyártás* mintaszó akusztikai diagramjaiból bemutatjuk a Spektrogram és az Intenzitás ablakok által megjelenített képet. A spektrális szerkezet tanulmányozásánál fontos lehet a frekvenciaértékek mérése is. A rendszer erre is lehetőséget ad. A Spektrogram ablak képén az egérrel kiválasztjuk azt a pontot, amelynek a frekvenciáját meg akarjuk mérni (például egy magánhangzó formánsértékét), és az eredmény automatikusan kiíródik a Frekvencia feliratú ablakban. A hangintenzitások részletei hasonlóan mérhetők. A mért dB-érték az Intenzitás feliratú mezőben jelenik meg.



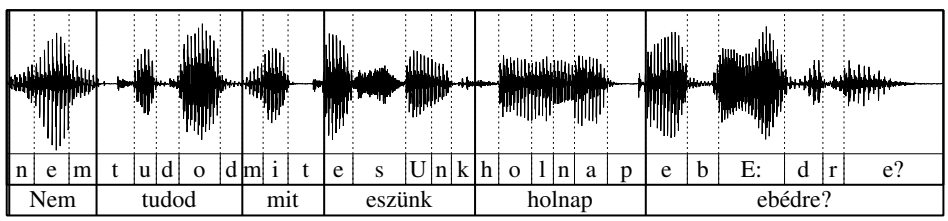
8.26. ábra. A *filmgyártás* mintaszó hangspektrogramja (fent) és intenzitásgörbéje (lent). A vízszintes tengelyen az idő olvasható le másodpercben

8.4.2. Mondatfajták beszédadatbázis

Olaszy Gábor

Ez a magyar beszédadatbázis felolvasott mondatokat tartalmaz férfi és női hangon kutatási célra. Felhasználásával bárki hozzájuthat olyan alapvető szegmentális és szupraszegmentális beszédakusztikai adatokhoz a magyar beszéddel kapcsolatosan,

amelyekhez eddig csak kutatólaboratóriumokból lehetett hozzáférni. Ez az első nyilvános, elektronikus, magyar nyelvű, folyamatos beszédet tartalmazó beszédadatbázis, és mondatkorpusz néven hivatkozunk rá. Elérése: `http://magyarbeszed.tmit.bme.hu`. Ez a mondatkorpusz egy férfi és egy női bemondó felolvasásában készült. Mintegy 500 mondatot tartalmaz beszélőnként. Mindkét beszélő ugyanazt a szöveganyagot olvasta fel, amely különböző hosszúságú és modalitású mondatokból áll, tartalmaz dialógusokat is. A hangállományt 22 kHz mintavételi frekvenciával és 16 bites lineáris kvantálással rögzítették. Minden mondat külön fájlban van tárolva. Ezenfelül két címkefájl is tartozik minden mondathoz. Az egyik tartalmazza a hangsor hangszimbólumait és a hanghatárcímkéket, a másik a szóhatárokat. A hanghatárok kijelölése kézi módszerrel történt. Úgy láttuk, hogy a címkézés pontossága érdekében érdemes interaktív vizuális-auditív ellenőrzést is felhasználni, hiszen egy ilyen precízen előkészített hanganyag sokáig szolgálhatja a kutatást, oktatást, fejlesztést. A címkézésre példát mutat a 8.27. ábra. A hangsor megadásakor külön kettőspont



8.27. ábra. Példa egy hangidőtartam-mérésre előkészített mondat időfüggvényére (2 s) a bejelölt szó- és hanghatárokkal

(:) jel szolgál a fonológiai hanghosszúsági kategória jelzésére. A hangok jelölése az E1-hangszimbólumokkal történtek (lásd a 4.2. fejezetben).

8.4.3. Elektronikus kiejtési szótár IPA-jelekkel és hangidőtartamokkal

Olaszy Gábor–Abari Kálmán

A kiejtési szótárak jól használhatók a kutatásban, az oktatásban, a nyelvtanításban, a gyakorlati alkalmazásokban és még számos területen. Lényegük, hogy az írott betűforma mellé megadják az adott lexikai elem kiejtési formáját is írott formában, úgynevezett hangszimbólumok alkalmazásával. Ebben a fejezetben az első magyar elektronikus kiejtési szótárt ismertetjük, amelyből magyar szavak és szóalakok kiejtési hangalakját lehet lekérdezni (Abari et al. 2006). A szótár 1,5 millió szóalakot tartalmaz.

A hagyományos kiejtési szótárak könyv alakban jelentek meg eddig (Fekete 1992, Tóthfalusi 2006). Ezekben a szótárakban a szerzők természetesen nem arra töreked-

tek, hogy a magyar szóállomány kiejtési formáit (a lehető legtöbb szóra) megadják, hiszen a terjedelmi korlátok ezt nem tették lehetővé. Inkább a különleges kifejezések, az idegen szavak kiejtését tették közzé. A feldolgozott elemek száma ezekben a szótárakban 10 000, illetve 40 000 egység.

Más nagyságrendek jellemzik az elektronikus formában megvalósított ilyen szótárt, amely tágítja a felhasználási lehetőségeket és a használat terét is, mivel többen férhetnek hozzá és lényegesen több lexikai elemből válogathatnak. Egy ilyen elektronikus kiejtési szótár nagy nemzetközi érdeklődésre is számíthat, ugyanis a más nyelvet beszélő megkeresheti benne bármely magyar szó kiejtési formáját. Az itt ismertetett elektronikus kiejtési szótár tervezésekor lényeges szempont volt a számítógépes támogatás maximális kihasználása. Ez sok különbséget jelent egy hagyományos szótárhoz képest. Az egyik ilyen a szóállomány kialakítása. A szóban forgó elektronikus kiejtési szótár nemcsak szótöveket tartalmaz, hanem azok ragozott, toldalékolt formáit is, mivel folyó szövegből kerültek be a listába. Ezért ebben a kiejtési szótárban a szótárelemeket *szóalak*nak hívjuk. A szóalak pontos definíciója a következő: olyan betűkből álló lexikai egység egy szövegben, amelyiket nem betűkarakterek határolnak (zömmel szóközök). Az így definiált betűkarakter-sorozat betűtartalma minden egyes szóalaknál legalább egy betűkarakterrel eltér a szótár más szóalakjától. Belátható, hogy elegendően nagy szövegkorpusz esetén az ilyen szóalakállomány jól lefedi a magyar nyelv leggyakrabban használt szavainak szóállományát, tehát általános kiejtési szótárként használható. A szótár minden lexikai elemének a kiejtését nemzetközi fonetikai hangjelekkel (IPA) adjuk meg írott formában. Ez lehetővé teszi, hogy anyanyelvtől függetlenül bárki értelmezze a kérdéses szó kiejtését. Külön lexikai csoportot alkotnak a leggyakoribb magyar családnevek, valamint az összes magyar település neve. A szótár másik különlegessége a hangos szótárrész, 60 000 szóalak hangban is meghallgatható. Ez segíti a felhasználót a tényleges kiejtés észlelésében, a hangidőtartamok érzékelésében, a szó ritmusának elsajátításában. A hangalakokat beszédszintetizátor állítja elő. A 8.28. ábrán példát mutatunk a szótár által kiadott találati listára egy négykarakteres lekérdezésre.

A szótár fejlesztésének menete. Első lépésben különböző szóalakokat gyűjtöttünk, nyolcvan millió szavas elektronikusan rögzített és hozzáférhető korpuszból (Németh–Zainkó 2002). A gépi gyűjtést többször kellett szűrni. Első szűrés: kivettük a listából az értelmetlen elemeket és a nem magyar szavakat. A tisztított lista képezte az alapot a hangátírásra. Második lépésben futtattuk a ProfiVox hangátíróját, ezzel megszülettek a hangszimbólumok minden szóalakra, amelyeket egyedi azonosító sorszámmal láttunk el. Ezek után ez a sorszám határozta meg a szót. A harmadik lépésben többlépcsős szűrést iktattunk be, ennek érdekében szétszedtük a szótárat, azaz gépi módszerrel kiválogattuk azokat a szóalakokat, amelyek hasonulásokat tartalmazhattak. Ezt betűkombinációkból alkotott kategóriák segítségével végeztük. Így több rövidebb szóalaklistát kaptunk (10–30 000 elem/lista). A következő

8.28. ábra. Az elektronikus kiejtési szótár válasza a lekérdezett **tech** betűsorozatra. A lista első 10 eleme látható

betűkombinációkat határoztuk meg kategóriaként: *tj, lj, tsz, tssz, ts, tss, cs, csz, ccs, ch* stb. Mindegyik kategóriára külön lista készült, amit emberi ellenőrzéssel javítottunk. Minden szóalakhoz el kellett dönteni, hogy jó-e a hozzá megadott hangátírási forma. Hiba esetén javítani kellett a hangkódokban. Ugyanilyen módon határoztuk meg a hiátustöltés jelölési pontjait. A negyedik lépésben összeszereltük a teljes szótárt (2009 nyarán), és ettől kezdve a teljes szóalakállományt vizsgáltuk különböző lekérdezésekkel. A feltárt hibákat gépi-kézi módszerrel javítottuk. Ebben a fázisban kerültek terítékre az idegen szavak, valamint minden, még észlelt rossz átírás. Ezekből villantunk fel néhányat: *ruggyanta, lánggyújtó, meggyé, meggyág, meggyepál, aljnövényzet, kontrolljoga, juhágazat, jazzszakszofon, szeizmikus, Wurlitzer, Yamaha, Braille, joystick, Disney, concerto, copyright*. Összességében ebben a fázisban 24 000 szóalakot kellett javítani a legkülönbözőbb átírási hibák miatt.

A kiejtési szótár 2009 végére került olyan állapotba, hogy tesztelésre érett legyen. A tesztelés 6 hónapig tartott. A szótár hibaaránya kevesebb, mint 1%. Ezért referencia kiejtési szótárnak is használható (például statisztikai elemzéshez, gépi tanuláshoz stb.).

Betű-hang átalakítási szabályok. A hangtani szabályok a magyarra a nyelvészeti szakirodalomban jól definiáltak, bár leginkább leíró formában hozzáférhetők (Jászó 1991). A gépi támogatáshoz használt ProfiVox hangátírót alapvetően nem kiejtési szótár elkészítésére tervezték, hanem szöveg-beszéd átalakításra egy adott technológiához, azonban a benne lévő szabályrendszer jó hatásfokkal lefedte a magyar kiejtési szabályokat. Ez a szoftver 85%-os pontosságúnak tekinthető, a maradék 15%-ot főleg a szabállyal nem kezelhető hasonulások és speciális kiejtési

formák jelentették (hiátustöltés, nazalizáció, fonémavariánsok, egyedi esetek). Ezeket kézi módszerrel építettük be a szótárba. A szótár átírási helyességét iteratív munkafázisokban gépi-kézi ellenőrzéssel vizsgáltuk. Az átíráshoz használt hangok a következők: alapállásban a magyar fonémák és azok néhány variánsa. A magánhangzók közül jelöltük az [aː] hang rövid variánsát (*sztrájk*= [strajk]), a mássalhangzó-variánsokból a palatális zöngés közelítő hang zöngétlen változatát (*lépj* = [leːpç]), a [h] veláris változatát (*almanach* = [ɔlmɔnɔx]) és annak palatoveláris formáját (*pechtől* = [pɛxtøːl]). Külön hangjel jelzi a hiátustöltés létrejöttét (*stúdió* = [ʃtuːdi^{j}oː]). A [^{j}] jelöléssel egyrészről magát a hiátustöltés tényét kívánjuk érzékeltetni, másrészről azt, hogy a hiátustöltő hang az esetek nagy többségében akusztikailag különbözik a fonémaértékű [j]-től. A hiátustöltésről részletesebben az 5.2.1.1. fejezetben szólunk.

A magyarban az esetek nagy többségében az írott betű megfelel a kiejtett fonémának (*ablak* = [ɔblɔk]), azaz helyesírásunk jó hatásfokkal fonetikusnak mondható. Vannak esetek amikor ez nem teljesül. Például a *c* és *s* betűkapcsolat funkciója kettős. Jelenthet egyetlen betűt (*kacsa*) vagy két különállót (*malacság* = [mɔlɔt͡s ʃaːg]). Ezeknek az eseteknek a feldolgozása külön lépésben történt. Idetartoznak azok az esetek is, amikor zöngétlenedés után a mássalhangzó-kapcsolatban rövidülés jön létre (*mondta* = [montɔ]; *küldte* = [kyltɛ]). Az *ipszilon* betűnek megfelelő kiejtés önálló hangként is megjelenhet idegen eredetű szavakban (*yen*=[jɛn], *hobby*=[hobːi], *boyszolgálat*=[bojsolgaːlɔt].

A hangidőtartamok tekintetében két további alsóbb szinten módosulhat a kapott hangsorozat: hosszan írjuk, röviden mondjuk *jobbra* = [jobrɔ]. Kiejtésváltozat lehet még a (*vállalat* = [vaːlɔlɔt], *kommunikál* = [komunikaːl], *mennyország* = [mɛɲorsaːg]. Ennek ellenkezője is előfordul, amikor röviden írjuk, hosszan ejtjük a szó valamelyik hangját (*lesz* = [lɛsː], *NATO* = [naːtoː]). Az ilyen kiejtésformákat is tartalmazza a szótár.

Posztlexikális szabályok. A magyar hangátírás legproblematikusabb része a hasonulások korrekt kezelése. A szótár készítése során azokkal foglalkoztunk, amelyeket a helyesírásunk nem jelöl. Ezek a szabályok a mássalhangzókat érintik, vagyis a részleges hasonulásból a zöngésedés, illetve zöngétlenedés, a képzés helye szerinti hasonulásból az [n] hasonulása [m], illetve [ŋ] hanggá (*színpad* = [simpɔd], *ponty* = [poŋc]). A teljes hasonulások, illetve az összeolvadás tekintetében sok esetben kétfajta átírás jellemezheti a betűsorozatot, a hasonult és a nem hasonult. Például összetett szavak határán a kiejtés közeledik az írásképhez, míg más esetben ugyanaz a betűkapcsolat hasonulással ejtendő (*feljavít* = [fɛljɔviːt]), de *teljes* = [tɛjːɛʃ]; *kétsávú* = [keːtʃaːvuː], de *kétséges* = [keːt͡ʃːeːgɛʃ], *átjavít* = [aːtjɔviːt], *látja* = [laːcːɔ].

A két betűvel jelzett fonémák és környezetük vizsgálatára nehéz egyértelmű szabályokat kialakítani. Ezeknél az eseteknél sok esetben csak ember tudja meghatározni a kiejtést. Külön vizsgáltuk azokat az eseteket, amikor az írásképben kettős

(hármas) betű jelöl egy-egy fonémát, és az egyes betűalakok külön-külön hangként is ejthetők. Ezekből mutatunk be példákat.

Az *s+z (s+s+z)* betűkapcsolatok:
szempont = [sɛmpont], *verszárlat* = [vɛrʒzaːrlɔt], *társszerző* = [taːrʃsɛrzøː],
szemeteszsák = [sɛmɛtɛʒːaːk].
A *z+s (z+z+s)* betűkapcsolatok:
zsák = [ʒaːk], *vízsodrás*= [viːʃːodraːʃ], *pénzzseni* = [peːnzʒɛni],
színészstáb = [sineːsʃtaːb], *rozzsal* = [roʒːɔl].
A *c+s (c+c+s)* betűkombinációk:
csoda = [t͡ʃodɔ], *ínyencség* = [iːɲɛnt͡sʃeːg],
láncszem = [laːnt͡sːsɛm], *tánccsoport* = [taːnt͡st͡ʃoport], *loccsan* = [lot͡ʃːɔn].
A *c+z* betűkapcsolat:
tánczene = [taːnd͡zzɛnɛ], *Rákóczi* = [raːkoːt͡si], *arczsába* = [ɔrd͡zʒaːbɔ].
A *c+h* betűkapcsolat:
táncház= [taːnh͡aːz], *Chile* = [t͡ʃiːlɛ], *scherzo*= [skɛrt͡soː],
cherry= [ʃɛri], *achát* = [ɔxaːt], *stichje* = [ʃtixjɛ]).
Más kombinációk is előfordulnak:
ánizsszag = [aːniʃsɔg], *törzsszám* = [tørʃsaːm], *bokszzsargon* = [bogzʒɔrgon]
varázszsámoly = [vɔraːzʒaːmoj].

A szótár végleges kialakítása során közel 150 000 szóalak kiejtése került kézi ellenőrzés formájában meghatározásra. Ez a kiejtési szótár alapvetően a magyar szóalakokra adja meg a kiejtési formát, erre is készült. Azonban a gépi összeállításból következik, hogy idegen eredetű szavak is szerepelnek a hatalmas szólistában. Ezeket és más, a korábbi szabályokkal nem kezelhető elemeket is tartalmazza a szótár kivétellistája. Ez közel 15 000 elem. Ezek között vannak idegen nevek, cégnevek, rövidítések és minden olyan kiejtési forma, amelyik az előző két modullal nem fedhető le (*city* = [siti], *plasa* = [plaːzɔ], *Peugeot* = [pøʒoː], *MTA* = [ɛmteːɔ], *adagio* = [adad͡ʒːoː]).

A kivételszótár elkülönített részét képezik a magyar családnevek kiejtési meghatározásai is, valamint külön csoportot alkotnak a magyar helységnevek (Páty = [paːc], Kistarcsa=[kiʃtɔrt͡ʃ ɔ] is.

A szótár a következő honlapon érhető el: `http://magyarbeszed.tmit.bme.hu`

8.4.4. A magyar formánsadatbázis

Olaszy Gábor

Ebben a fejezetben az első magyar formánsadatbázist mutatjuk be, amely referenciaként szolgálhat beszédkutatási és oktatási célok támogatására (Olaszy et al. 2009).

Referenciának azért tekinthető, mert a benne lévő adatokat kézi és gépi módszerrel hozták létre, és helyességüket többszöri vizsgálattal ellenőrizték. A magánhangzókat tárgyaló fejezetben ebből az adatbázisból közlünk lekérdezéseket bizonyos tendenciák, jelenségek bemutatására. A formánsadatbázis egy férfi és egy női bemondó hangjából kinyert adatokat tartalmazza. Mindketten ugyanazt a szólistát olvasták fel.

Az a gondolat, hogy formánsmérések eredményeit adatbázisba tömörítsék, nem régi. A Microsoft kutatólaboratóriumában (Deng et al. 2006) készítettek minősítési és beszédkutatási feladatok támogatására formánsadatbázist a TIMIT beszédadatbázis kiválasztott mondataira angol nyelvre. Egy speciális vizuális-grafikus editort fejlesztettek ki, melynek segítségével szakértők rajzolták be a mondatok minden hangjára a formánsmeneteket. Ezek jelentik abban az adatbázisban a referenciaadatokat, vagyis a formánsadatbázist. Más ilyen adatbázisról nincs tudomásunk. A magyar fejlesztés célja az volt, hogy egységes és nyilvános adathalmazt hozzunk létre a magyar magánhangzókat és hangkapcsolódásokat jellemző formánsértékekre, formánsmozgásokra. A magyar magánhangzók formánsértékeinek kutatása régi keletű (Tarnóczy 1941, Magdics 1965, Bolla 1978, Olaszy 1985, Gósy 2004b). Mi a különbség a referenciaformáns-adatbázis és a korábbi formánsmérésekből közölt adatok között? Annyi, hogy a publikációkban csak egy-egy célzott mérés végeredményét adják közre a kutatók, a kiindulási adatok nem hozzáférhetők. A referenciaformánsadatbázisban viszont a formánsok frekvenciaadatai tételesen szerepelnek, és ezekből lehet tetszés szerinti kutatást végezni. Mindehhez az is hozzátartozik, hogy a referencia-formánsadatbázis számadatai mellett rendelkezésre áll a mérések beszédanyaga is, annak fonetikai átiratával és hanghatárjelöléseivel egyetemben. Mindez nyilvános, így nem csak a végeredmény, de a mérési folyamat bármely eleme is minden kutató részére hozzáférhető, tetszőleges mérések tervezhetők, az eljárások reprodukálhatók. Ez a referencia-formánsadatbázis csak viszonylag kis méretű korpuszra készült el, de meggyőződésünk, hogy jól reprezentálja a magyar beszédet. Az elkészítéséhez kidolgozott technológia lehetővé teszi, hogy tömegesen lehessen ilyen adathalmazokat készíteni a jövőben. A következőkben ismertetjük az adatbázis létrehozatalának folyamatát és a munka végeredményét.

A fejlesztési munka alapvető célja az volt, hogy referencia-formánsadatbázist hozzunk létre egy nyilvános beszédkorpuszra alapozva. Az ilyen adathalmaz többféle munkában nyújthat segítséget. Egyrészről lekérdezhetők formánsadatmezők, átlagok, hangfüggő mérések végezhetők, hangkörnyezeti hatások tanulmányozhatók stb. Másrészről arra is alkalmas, hogy független minősítő méréseket hajtsunk végre bármely gépi formánselemző hatásfokának vizsgálatára.

A kísérletek nyelvi anyagául egy nyilvános, nagy pontosságú belső adatokkal rendelkező magyar beszédadatbázist használtunk (Abari–Olaszy 2007), amely a `http://fonetika.nytud.hu/cvvc` honlapon található (más, nyilvános magyar beszédadatbázisról nem tudunk, amelyen bárki ellenőrző méréseket végezhet). Ezt az adatbázist részletesen ismertettük a 8.4.1. fejezetben. A módszer tekintetében más

utat követtünk, mint az angol nyelvre kidolgozott formánsadatbázis fejlesztői (Deng et al. 2006). Egyrészről nem folyamatos formánsmeneteket rögzítettünk, hanem előre meghatározott pontokra adtuk meg a formánsértékeket, csak magánhangzókat jellemeztünk és a formánsadatokhoz való hozzájutást is más módszerrel végeztük. További különbség, hogy nem mondatokra, hanem szavakra épül a vizsgálati anyag. Ugyanakkor követtük Lee és munkatársai módszerét, akik kimutatták, hogy könnyítheti a gépi formánsmérő algoritmus dolgát, ha ismerjük a hanghullám tartalmának fonetikai átírását, azaz a hangokat, amelyekben a formánsokat meg kell határozni (Lee et al. 1999).

A célkitűzés végrehajtására dolgoztuk ki azt a kétlépcsős módszert, amely a jövőben az ilyen munkát segítheti. Az új eljárás lehetővé teszi, hogy a magánhangzókban létrejövő formánsmozgások jellemzésére nagy pontossággal lehessen formánsadatokat meghatározni bármely formánselemző gépi algoritmus felhasználásával. A kifejlesztett módszer alapvetően kétlépcsős és nyelvfüggetlen.

Az első lépésben a Praat általános gépi formánsmérő programmal (Boersma–Weenink 2009) elvégeztük a méréseket. Ezeket nyers adatoknak tekintettük. A második lépésben egyszerű matematikai eszközökkel ellenőriztük ezt az adathalmazt, és a fonetikai elvárások figyelembevételével jelöltük a mérési hibából eredő rossz formánsértékeket, ezeket vizuálisan ellenőriztük, és javítottuk. A tömeges formánsmérésekhez készítettünk kiegészítő algoritmust és szkriptet, amely elvégezte a szóállomány hullámformáinak automatikus feldolgozását. A gyári alapbeállításokon nem változtattunk, a mérési sáv maximuma 5500 Hz volt.

A formánsmérés tere. A mérést közel 50 percnyi beszédanyagon (szólista) végeztük. A formánsmérés terét a mért hangok típusai, a hangon belüli mérési pontok, valamint a mért formánsok száma határozta meg. A felolvasott szavak összes magánhangzójában végeztünk mérést. Figyelembe vettük a magánhangzót megelőző és követő hangot is, hogy a későbbiekben környezetfüggő lekérdezést is lehessen kérni. Összesen 9975 olyan hanghármas szerepelt a beszédadatbázisban, amelyek középső hangja magánhangzó (zömmel CVC kapcsolatok). A magánhangzó 25, 50 és 75%-os pontjain határoztuk meg a formánsok értékét. A 3 mérési ponton való adatgyűjtéssel az volt a célunk, hogy a későbbiekben vizsgálni lehessen a hangkörnyezet hatását is, vagyis a formánsmozgások jellemző tendenciáit a magánhangzón belül. A szókezdő és szózáró magánhangzókat is vizsgáltuk, a # jelet használva a hanghármas definiálásánál. A szókezdő és szózáró magánhangzókban csak két mérési ponton rögzítettük az eredményeket (50, 75%, illetve 25, 50%), mivel a kísérletek során kiderült, hogy az induló, illetve lecsengő hangokban a kezdeti, illetve a végső szakaszokban az elemző sokszor hibás értéket mért. Ezeken a helyeken a beszéd spektrális tartalma bizonytalannak tűnt. A hangok jelölésére az E1-szimbólumokat használtuk (lásd a 4.2. fejezetet). A mérések során 29 926 mérési ponton rögzítettünk adatot. Minden mérési ponton négy formáns adatait mértük

meg. Összesítve tehát 119 704 formánsadat szerepelt a Praat program által előállított nyers adathalmazban. Fontos célkitűzés volt, hogy az adatbázis pontossága minél magasabb százalékértékkel legyen jellemezhető. A pontatlanságnak több oka is lehet (pontatlan előkészítés, pontatlan mérés, emberi tévedés, rossz ejtés stb.). Esetünkben a fonetikai elvárásnak nem megfelelő mérési adatokat tekintettük pontatlannak (például a spektrogramon látható a harmadik formáns vonulata, a program mégsem ad adatot erre a formánsra, ugyanakkor más elemző esetleg jelezné ezt).

Manuális hibafeltárás és javítás. Az automatikus elemző által produkált nagy mennyiségű formánsadattal kapcsolatosan tudni lehetett, hogy vannak köztük hibás értékek is. A kérdés az, hogy milyen módszerrel tudjuk megtalálni a hibás számértékeket az adathalmazban, valamint hogy mi alapján javítunk. A hibakeresést több fokozatban, manuálisan végeztük. Első lépésként formánsonként emelkedő sorrendbe rendeztük az adatokat. Tájékozódásul felhasználtuk a magyar magánhangzókra vonatkozó korábbi kutatások adatait (Olaszy 1989a, Gósy 2004b), és ezek tükrében meghatároztuk a várható és a mért érték közötti különbségből azt (megfelelő küszöbértékek alapján), hogy elfogadható vagy hibás-e az adat. Az esetleges ellenőrzésekhez és a javításokhoz a formánsadatokat a kísérletek nyelvi anyagát adó honlapról megjelenített hangspektrogramok vizuális elemzéséből vettük (átírtuk a mért rossz értéket a vizuális mérésből kapottra). Az ilyen hibák kijavítása után a kivonásos elvet alkalmazva különbségi adatsorokat képeztünk, amelyekben az x = F2−F1, y = F3−F2 és z = F4−F3 értékeket szerepeltettük sorba rendezve. A formánsok távolsága ugyanis megadott határokon belül mozog minden magánhangzónál. A különbségképzés legkirívóbb eredményeire lássunk néhány példát (ezeket a hibás értékeket a formánselemző program tévesztései eredményezték). Az x = F2−F1 legkisebb értéke 14 Hz volt, és 100 Hz alatti különbség 25 esetben fordult elő. Az y = F3−F2 különbségnél 7 Hz volt a legkisebb távolság, és 127 esetben volt 100 Hz alatt a két különbség. A z = F4−F3 mérésből 8 Hz volt a legkisebb, és 58 esetben 100 Hz alatti volt a különbség a két formáns között. Az ilyen hibajelenségekből azt következtetjük, hogy a formánselemző nincs felkészítve a formánstávolságok érzékelésére. Elvárható lenne, hogy 10–20 Hz távolságra ne jelöljön két külön formánst, hiszen ez fizikai képtelenség. A fent említett hibás mérési értékeket a vizuális feldolgozási módszerrel javítottuk. A manuális hibafeltárás harmadik lépésében a három mérési pont formánsértékeit hasonlítottuk össze a magánhangzókon belül. Felhasználtuk azt a fonetikai ismeretet, hogy egy magánhangzón belül az egyes formánsok mozgástere behatárolható a hangkörnyezet ismeretében és tipizálható is. Így megállapíthatók a gyanús elemek. Ezeket manuálisan ellenőriztük és javítottuk. A hibakeresés utolsó fázisában eloszlásokat rajzoltunk az x = F1 és y = F2, majd az x = F3 és y = F4 síkokra a mért 14-féle magánhangzóra, külön a férfi és külön a női ejtésre (összesen 28 eloszlási kép). A kiugró értékeket ellenőriztük és javítottuk. A manuális ellenőrzés során mindösszesen 10 307

formánsadatot kellett módosítanunk. Ennek a munkának a végén érkeztünk el ahhoz a ponthoz, amikor azt mondhattuk, hogy a formánsadatbázisunkban csak minimális számú hiba maradt, tehát referencia-adatbázisnak tekinthető. Egy példát mutatunk a 8.13. táblázatban a formánsadatbázis belső szerkezetéről.

A mérési, javítási eredmények összefoglalása, értékelése. A nyers adatbázis összes magánhangzójának 25,1%-ában kellett kézi javítást végezni (egy vagy több formáns adatát kellett korrigálni), összesen 2669 hangban. A férfi ejtésű magánhangzók 88,8%-ában voltak fonetikailag elfogadható adatok, a nőiben 76,4%-ban. Ez a két adat a Praat program mérési pontosságát is tükrözi (ebben a mérésben). Általános tendenciának látszik, hogy a férfihangban lényegesen jobb a formánselemző mérési pontossága, mint a nőiben. Ez valószínűleg abból is fakad, hogy a mélyebb alaphang felharmonikusai sűrűbbek, mint a magasabbé. A sűrűbb felhangok jobban kitöltik a formánsfrekvencia körüli frekvenciateret, mint a ritkábbak. Az értékelésnél fontos a formánsokra lebontott részletezett adatok eloszlása is, hiszen az első két formánsnak fontosabb szerepe van a hang jellemzésében, mint a többinek (8.14. táblázat).

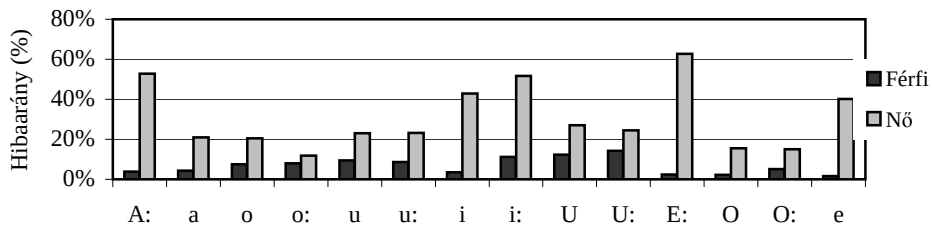
8.13. táblázat. Az *ácsjelvény* szó három magánhangzójában mért formánsadatok megadása az adatbázisban a férfi (f) és a női (n) bemondó hangjára

	Szó	A hang sorszáma	Előző hang	Mért hang	Következő hang	Pont	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
f	A:CjelvE:N	1	#	A:	C	50%	723	1478	2561	3529
f	A:CjelvE:N	1	#	A:	C	75%	567	1654	2586	3593
f	A:CjelvE:N	4	j	e	l	25%	490	1845	2466	3506
f	A:CjelvE:N	4	j	e	l	50%	576	1683	2452	3566
f	A:CjelvE:N	4	j	e	l	75%	585	1597	2481	3596
f	A:CjelvE:N	7	v	E:	N	25%	382	1965	2472	3508
f	A:CjelvE:N	7	v	E:	N	50%	392	2027	2560	3550
f	A:CjelvE:N	7	v	E:	N	75%	361	2105	2644	3522
n	A:CjelvE:N	1	#	A:	C	50%	896	1645	3100	4280
n	A:CjelvE:N	1	#	A:	C	75%	862	1859	3030	4392
n	A:CjelvE:N	4	j	e	l	25%	600	2063	2876	4322
n	A:CjelvE:N	4	j	e	l	50%	720	1888	2785	4344
n	A:CjelvE:N	4	j	e	l	75%	714	1850	2859	4321
n	A:CjelvE:N	7	v	E:	N	25%	458	2546	2929	4266
n	A:CjelvE:N	7	v	E:	N	50%	483	2549	2954	4242
n	A:CjelvE:N	7	v	E:	N	75%	896	1645	3100	4280

8.14. táblázat. A hibásnak talált formánsadatok részletezése (db)

	F1 javítása	F2 javítása	F3 javítása	F4 javítása	Összes javítás
Férfi	177	215	235	289	916
Nő	185	1516	3104	4586	9391

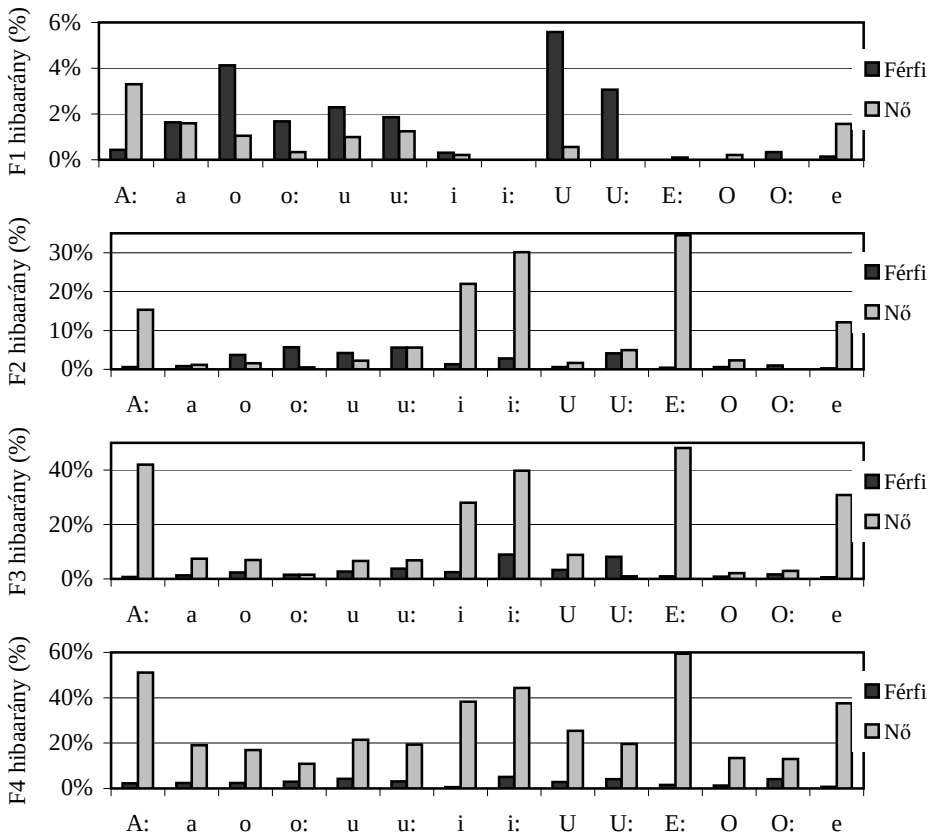
A részletezett adatok azt mutatják, hogy minél magasabb formánsról van szó, annál bizonytalanabb a gépi formánsmeghatározás. Az F1-et nagy pontossággal meg tudja mérni a Praat elemző, ha mégis téveszt, akkor egyforma gyakorisággal fordul elő nemtől függetlenül. A férfi hangban minden formáns tévesztési aránya közel egyenlő. A női hangnál az F2 esetében a hibás adatok száma jelentősen megugrik, az F3-nál megduplázódik, és tovább növekszik az F4 esetében is. A hibás formánsmérések hangfüggőek. Azoknál a hangoknál, amelyekben a formánsok egymástól való távolsága minden formáns esetén viszonylag nagy, a tévesztések száma kicsi ([ø øː]). Ez minden hibaarány-grafikonon jól látható. A mért hangok számának arányában feltüntetett javítások eloszlását a 8.29. ábrán mutatjuk be. A férfi hangnál a



8.29. ábra. Az automatikus formánselemző program tévesztései hangonként és nemenként lebontva. A hibaarány azt mutatja, hogy az előforduló hangokból mennyiben kellett valamelyik formáns értékét korrigálni

legnagyobb arányban az [i y yː]-nél tévesztett az automatikus elemző, a női hangnál az [aː i iː eː ɛ]-ben kellett leginkább javítani. A formánsokra és hangokra lebontott hibaelőfordulások adják a legrészletesebb adatokat az automatikus formánselemző működéséről. A 8.30. ábrán bemutatjuk formánsonként és hangonként a hibás mérések eloszlását.

A gyakorlati hibaelemzésből végül két általános tapasztalatot emelünk ki. Az egyik, hogy a Praat sok esetben úgynevezett elcsúszási hibát mutatott, ami azt jelentette, hogy például egy hamis formánst észlelt az F2-nek, így a valódi F2-t az F3-ra, a valódi F3-at pedig az F4-re tette. A másik megfigyelés, hogy ha nazális környezetben van a V, akkor az elemző nehezen tud jó adatokat kinyerni, mert a formánsok jobban elmosódnak a nazalizálódott magánhangzóban (például az [o] hang a *demonstráció* szóban). A manuális javítás tapasztalatai alapján olyan algoritmust lehet kidolgozni, amelyik eredményesen fogja detektálni a hibásnak tekinthető formánsméréseket nagy adathalmazokban, ezzel megnyílik az út a tömeges formánsgyűjtésre, azok pontosítására, tehát hiteles formánsadathalmazok kialakítására. A formánsadatbázis segíti a kutatást és az oktatást (feladatok adhatók, mérések végezhetők, tendenciák bemutathatók, vizsgálhatók stb.).



8.30. ábra. Az automatikus formánselemző program tévesztései hangonként lebontva az F1, F2, F3 és F4 esetében

8.5. Spontánbeszéd-adatbázisok

Olaszy Gábor

A spontánbeszéd-adatbázisok nem felolvasott beszédet, hanem spontán, a pillanatnyi beszédtervezés alapján létrejött hangzó anyagot tartalmaznak. Ezek a hanggyűjtemények állnak a legközelebb az emberi kommunikáció szemtől szemben lefolytatott dialógus beszédformájához. Elkészítésük sokkal problematikusabb, mint a felolvasott adatbázisoké (Gósy 2008a), a hangfelvételtől kezdve az annotálásig, a fonetikus átírásig, az esetleges címkézésig. Magyarországon már a 80-as években elkezdődtek ilyen gyűjtések az MTA Nyelvtudományi Intézetében. A Budapesti Szociolingvisztikai Interjú projekt keretében 250 beszélővel rögzítettek többek között oldott stílusú riport beszélgetést is (Kontra 1988, Váradi 2003). Az anyagot nyelvészeti vizsgálatokhoz használták (Kontra 1988). Más céllal készült hangfelvételek rögzített anyagai szintén tekinthetők egyfajta spontán beszédadatbázis-kezdeménynek, bár

nem ilyen céllal rögzítették őket. Ilyen például a Hegedűs-archívum (Horváth 2008). Szorosabban vett beszédkutatási, illetve beszédtechnológiai céllal tervezett magyar spontánbeszéd-adatbázisok csak a 21. században kezdtek el kialakulni.

Egyik reprezentánsa ezeknek a munkáknak az MTA Nyelvtudományi Intézetében 2007 őszén elkezdett gyűjtés (Gósy 2008a), amely BEA (BEszélt nyelvi Adatbázis) néven ismert. Ez egy fonetikailag megtervezett, többcélú beszédkorpuszgyűjtemény, amely szolgálhatja a modern beszédkutatást, a nyelvészetet, valamint a beszédtechnológiát is. A több száz beszélővel számoló adatbázis kialakítása több évet vesz igénybe. 2010-re 100 beszélő anyagát készítették el. A BEA nem csak spontán beszédet tartalmaz, hanem – a későbbi összehasonlíthatóság kedvéért – utánmondott mondatokat és olvasott beszédet is. A spontán beszéd része három részből áll: a) narratíva – a beszélő a munkájáról és saját életéről mesél, b) véleménykifejtés különböző témakörökben, c) tartalomösszegzés egy meghallgatott szöveg alapján. Általában – ahogy azt a korábbi beszédadatbázisoknál is láttuk – minden beszédadatbázis tartalmazza az elhangzottak szöveges lejegyzését is. A BEA esetében a lejegyzők magyar helyesírással, központozás nélkül írják le az elhangzottakat. A lejegyzés formájára és szabályaira a tervezők külön szabályrendszert dolgoztak ki. Mondathatárokat nem jelölnek, nagybetűt csak a tulajdonnevek írásakor használnak. Az egyéb hangadások jelölésére a felkiáltójelet illesztik be a betűsorba. Például a *ki!mentünk!* lejegyzés azt jelenti, hogy egy nevetés volt a *ki* után, illetve hogy a szó végén volt egy torokköszörülés. A csillag karaktert használják az értelmezhetetlen elemek jelölésére. A megakadásjelenségeket vastagított betűkkel jelölik. Szögletes zárójelek közé írva jelölik az elharapott szót betűkarakterekkel (nem hang) (például: [tát] *tehát*, a szünet beiktatását a négyzet jellel érzékeltetik. A tapasztalatok szerint a hangzó anyag leírása nagy türelmet, kiváló fonetikai ismereteket, átírási gyakorlatot és koncentrációt kíván. Egy 10 perces monológ leírása egy gyakorlott átírónak mintegy 1 órájába kerül, és további 30 percet kell szánni az ellenőrzésre. Egyetlen beszélő teljes felvételének leírása és ellenőrzése mintegy 12 órát vesz igénybe. A fonetikai átírás nem szerepel a kitűzött célok között. A BEA többcélú beszédadatbázison már 2008-ban számos kutatás indult el. Ilyenek: a beszédtervezés és -kivitelezés sajátosságai (Gósy–Horváth 2010), a megakadásjelenségek vizsgálata, az ismétlések, újraindítások milyen ejtési összefüggéseket mutatnak, a téves kezdéseket és szótalálásokat is vizsgálják. Az akusztikai vizsgálatok is megindultak főleg a formánsstruktúra (Horváth–Gráczi 2010), illetve a hangidőtartamok területén (Gósy–Beke 2010).

9. fejezet
A beszéd gépi észlelése és felismerése

Szaszák György–Mihajlik Péter–Fegyó Tibor

Az ember régi vágya, hogy az általa konstruált gépekkel, berendezésekkel emberi nyelven tudjon kommunikálni. A természetes nyelvű ember-gép dialógusnak a beszédmegértésre irányuló elemét nevezzük gépi beszédészlelésnek. A beszédészlelés terminus gyakorlatilag mindent magába foglal abból, amit az ember a beszédpercepció során megállapíthat a másik ember beszédéből: a tartalmon túl a beszélő attitűdje, érzelmei (például könnyed vagy dühös), fizikai állapota (például náthás, rekedt) is kitűnnek, csakúgy, mint az, hogy az anyanyelvén beszél-e, járatos-e a témában, milyen társadalmi rétegből származhat stb. Ez új tudományterület, és a 20. század közepétől van jelen a kutatásban. Ennek fő oka, hogy a számítógép megjelenéséig nem volt mód az emberi hang beható analízisére, elemekre bontására, részenkénti tanulmányozására. Ugyanakkor éppen a számítógépek azok, amelyek programnyelvei a természetes nyelvi gondolkodáshoz képest teljesen más logikán alapulnak, így a géppel való természetes nyelvű kommunikáció lehetőségének hiánya hasonló problematikát vet fel, mint két különböző nyelvet beszélő ember kommunikációja: tolmácsra van szükségük, vagy meg kell tanulniuk a másik nyelvén. Eleddig az ember kényszerült az általa alkotott gép nyelvét megtanulni. A természetes nyelven történő kommunikációval mindez szükségtelen lehetne.

A gépi beszédészlelést megvalósító alkalmazások tipikusan a gépi beszédfelismerők, amelyek puszta beszéd-szöveg átalakítást végeznek anélkül, hogy a beszédben hordozott jelentést megérteni képesek lennének. Az első gépi beszédfelismerő alkalmazás angolul kimondott számjegyeket tudott azonosítani (Davis et al. 1952). Mintegy 10 évnyi kutatás volt szükséges ahhoz, hogy a számjegyeken túl matematikai műveleteket hangvezérléssel tudjon elvégezni a számítógép. 1964-ben mutatta be az IBM a Shoebox rendszert a Seattle-ben rendezett világkiállításon. Ez a rendszer 16 szót tudott felismerni. Ezután rohamos fejlődésnek indult a gépi beszédfelismerés, köszönhetően a számítógépek gyors fejlődésének, a memóriakorlátok csökkenésének és a különböző matematikai modellek számítógépes megvalósításának. Az elmúlt 50 év során azonban az is világossá vált, hogy egyelőre a gép még napjaink technikai

színvonalán sem versenyezhet az emberi beszédmegértési mechanizmussal. Ezért a különböző alkalmazásokhoz célorientált felismerő algoritmusokat fejlesztenek, az általános megoldástól még távol vagyunk. Különös nehézséget jelent a környezeti zajok szétválasztása a beszédjeltől, amit az emberi beszédértési rendszer könnyedén meg tud tenni (hamar adaptálódik a zajos környezethez). Az emberi beszédpercepciót vizsgáló kutatások eredményei arra engednek következtetni, hogy az ember a beszédfelismerés két fő komponensét használja: az akusztikai jel feldolgozását és párhuzamos társítását az agyban lévő nyelvi bázissal. A mai beszédfelismerő rendszerek már mindkét elemet használják (egyszerű modellek formájában) a sikeres felismerés eléréséhez. A gépi beszédfelismerés a társadalom minden területén használható. Főbb területei a 21. század elején a következők: távközlési alkalmazások, automatikus keresés hangzó anyagokat tartalmazó adatbázisokban, diktáló rendszerek (hangvezérelt és beszédalapú szövegszerkesztők), biztonságtechnikai hanglenyomatok (adathozzáféréshez), parancsszavas vezérlések (gépipar, járművek, mobiltelefonok, televíziók) stb.

A gépi beszédészlelés tehát átfogó beszédfeldolgozási témakör. Legismertebb célja a beszéd tartalmának automatikus meghatározása, azaz a beszéd lejegyzése (gépi beszédfelismerés, azaz a beszéd-szöveg átalakítás), illetve szóban történő utasítások végrehajtása. Mindez kiegészülhet az ember és a gép közötti természetes nyelvű dialógus megvalósításával (ember-gép interfész). Napjainkra a beszédfelismerés e történetileg elsőként felmerült feladatkörénél már többet várunk el a géptől: nemcsak a beszéd írott formájúvá alakítása lehet érdekes, hanem a beszélő személy felismerése, azonosítása, továbbá a beszélő hangulatának, érzelmeinek, egészségi állapotának automatikus felismerése is stb. A lehetséges feladatokat a következőkben rövidesen bemutatjuk. Általánosságban azonban még meg kell jegyeznünk, hogy napjainkban a beszédfelismerés (speech recognition) felől az érdeklődés egyre inkább eltolódik a beszédértés (speech understanding) irányába: azaz már nem pusztán a beszéd szöveggé alakítása a cél, hanem a beszéd által átvitt információ jelentésének is minél teljesebb körű dekódolása, azaz a beszéd megértése.

Mivel történetileg a gépi beszédészlelés sokáig egyet jelentett a gépi beszédfelismeréssel, illetve mivel a gépi beszédészlelést célzó alkalmazások közös gyökerének a beszédfelismerés tekinthető, a fejezetben elsősorban a gépi beszédfelismeréssel, azaz a beszéd-szöveg átalakítással foglalkozunk, annál is inkább, mert napjainkban a gépi beszédészlelési feladatok közül szinte egyedüliként a beszéd-szöveg átalakítás, illetve korlátozottan a beszélőfelismerés tekinthetők olyan technológiáknak, amelyek korántsem teljesen kifejlesztettek és megvalósítottak ugyan, mégis már a mindennapi életben, illetve kereskedelmi forgalomban is találkozhatunk velük. A gépi beszédészlelés egyéb területei napjainkban még szinte kizárólag kutatási fázisban vannak.

A gépi beszédészlelés, illetve ennek részeként a gépi beszédfelismerés alapvetően az akusztikai beszédjelre támaszkodik, de létezik olyan megközelítés is, amely egyéb kommunikációs csatornán – leginkább vizuálisan – közvetített információt is

figyelembe vesz, ez a *multimodális* beszédfelismerés. Ennek során leggyakrabban a beszédhangok artikulációját kísérő szájmozgást és a beszélő gesztusait (arcjátékát, de akár testtartását, teljes testével kifejezett gesztusait) is figyelemmel kísérik. Ekkor a beszéd felismerése és minél mélyebb megértése a két információforrás (auditív és vizuális) integrálása alapján történik (lásd a 9.12. fejezetben).

9.1. Gépi beszédészlelési feladatok

Amint a bevezetőben már szó volt róla, a beszéd gépi észlelése meglehetősen tág témakör. A következőkben röviden áttekintjük, milyen feladatok tartoznak a ebbe a témakörbe. A felsorolás korántsem teljes, mert hely hiányában lehetetlen volna valamennyi alkalmazási területet felsorolni.

A legalapvetőbb feladat a *beszéd-szöveg átalakítás*, azaz a beszéd tartalmának felismerése és lejegyzése – ez a történetileg kialakult gépi beszédfelismerés (speech recognition). Számos további, a gépi beszédészleléshez sorolható technológia ezen alapul vagy ebből fejlődött ki. Éppen ezért a későbbiekben ezt tárgyaljuk a legmélyebben.

A *kulcsszókeresés* akkor lehet hasznos, ha adatbázisokban beszédalapú keresést szeretnénk megoldani, ekkor a keresést nem írott állományokban, hanem hangállományokban végezzük a megadott szöveges vagy bemondott kulcsszó alapján.

Beszélőfelismerés esetén egyik lehetséges célunk a beszélő személy igazolása (speaker verification), és ezáltal valamely rendszerhez való hozzáférési jogosultság vizsgálata. A felhasználó megadja az azonosítóját (bemondja), a rendszer az aktuális bemondást egyetlen vagy több referenciamintával veti össze, az eredmény pedig elfogadás vagy elutasítás. A másik lehetséges, bár ritkább felhasználási terület a beszélő személy azonosítása egy előre definiált halmazból (speaker identification) (Gordos–Takács 1983). Ekkor több beszélő hangja közül választja ki a rendszer az aktuálisan beszélőt, az eredmény pedig a beszélő megnevezése, vagy annak jelzése, hogy a beszélő hangmintája nem található meg a referenciahalmazban. A beszélőfelismerés történhet szövegfüggő vagy szövegfüggetlen úton (Furui 1996). Az előbbi esetben a beszélő igazolása vagy azonosítása meghatározott, és a beszélő által előzetesen ismert beszédelemek alapján történik. A módszer nagy hátránya, hogy a beszélőtől felvétel útján előzetesen rögzített bemondás alapján visszaélésre ad lehetőséget, így a szövegfüggetlen módszer tekinthető biztonságosnak: ekkor a beszélő azonosítása egy előzetesen nem ismert, az azonosítás során számára a helyszínen képernyőn megadott szöveg bemondása alapján valósul meg. A beszélőfelismerésre használhatók a beszédfelismerésre kidolgozott algoritmusok, esetenként megfelelő módosításokkal (Furui 1996), ezeket a vonatkozó fejezetekben részletesen ismertetjük.

A *beszéddetekció* (Voice Activity Detection (VAD) vagy Speech/Non-speech Detection) szinte minden beszédészlelő alkalmazás elengedhetetlen része (Tucker 1992). Tudnunk kell ugyanis, hogy mikor beszél a felhasználó és mikor nem, hiszen az utóbbi esetben felesleges például a felismerő rendszernek működnie. A csendes beszédszünetek jelzésén kívül szükség lehet a környezeti zajok, sőt a zene beszédtől való elkülönítésére is. Tipikusan ilyen problémával találkozhatunk a híranyagot tartalmazó hangos adatbázisokban, ahol a kulcsszókeresés vagy az automatikus feliratozás előfeltétele lehet, hogy a beszédet elkülönítsük a háttérbeszédtől, illetve a háttérzenétől (Vandecatseye et al. 2004). A beszéddetekcióhoz kapcsolhatók a beszéd szakaszolását megvalósító alkalmazások is (prozódiai frázisokra történő szegmentálás, beszélőváltás-detektálás stb.).

Az *érzelmi töltet felismerése* viszonylag fiatal ága a gépi beszédészlelés tudományának (Sebe et al. 2005). Egyelőre az érzelmek durvább osztályozása lehet reális célkitűzés, általában 6–8 úgynevezett alapérzelmet szokás elkülöníteni. A számos felhasználási lehetőségen túl az érzelmek változásának követése sokat segíthet a dialógusok dinamikus felépítésében, a beszélő érzelmeire adekvát gépi válasz kiválasztásában, ily módon az ember-gép kommunikáció teljesebbé tételében.

*Nyelvfelismerés*re van szükség többnyelvű beszédfelismerő rendszerekben, amelyekben első lépésként a munkanyelv kiválasztását kell automatikusan megoldani.

9.1.1. A gépi beszédfelismerők osztályozása

A továbbiakban a gépi beszédfelismerőkkel foglalkozunk, amelyeket számos szempont alapján tovább osztályozhatjuk a következők szerint.

Beszédmód. Megkülönböztethetünk *izolált szavas*, *kapcsoltszavas* és *folyamatos* beszédfelismerőket. Az izolált szavas beszédfelismerő szavak felismerésére alkalmas, használatakor a felhasználónak a szavak között szünetet kell tartania. A kapcsoltszavas rendszerben egymás után kiejthetünk bizonyos szavakat elválasztó szünet nélkül, míg a folyamatos beszédfelismerő (Jelinek 1976) képes kezelni a folyamatos beszédet, így a legközelebb áll a természetes nyelvhasználathoz. Az izolált szavas felismerők egy lehetséges felhasználási területe dialógusokban képzelhető el, számjegyek, megerősítő válaszok, nevek stb. felismerése, a kapcsolt szavas felismerők akár teljes dátumokat vagy címeket is képesek kezelni, míg a diktáló rendszerekben a folyamatos felismerés a követelmény. Izolált szavas beszédfelismerő működik a 12.3.5. és a 12.3.8. fejezetekben ismertetett információs rendszerben.

Beszélőre adaptáltság. A beszélőre történő adaptáltság szempontjából a beszédfelismerő lehet *beszélőadaptált*, ekkor a felhasználó hangjára rátanul a rendszer, a felismerés során tehát személyfüggő tudást is használ. A beszélő adaptációja a felhasználó aktív közreműködésével történik. Lényegében ezzel a felismerés pontosabb lesz

(Padmanadham et al. 1998). Diktáló rendszerekben gyakorlatilag elengedhetetlen ez a megoldás, azonban például nyilvános információlekérdező rendszerekben, ahol sok ember használja, kivitelezése nem megoldható. Általános felhasználásra alkalmazzák a *beszélőfüggetlen* beszédfelismerőt, amely nem adaptált ugyan, kialakítása azonban olyan, hogy ennek ellenére optimális működést lehet elérni. Ilyen felismerőt ott használnak, ahol nincs lehetőség az adaptációra (sokfelhasználós rendszer (például a lakosság), rövid idejű használat).

Akusztikai környezet. A beszédfelismerők működése szempontjából rendkívül fontos a környezet figyelembevétele. *Csendes környezet*ben jó jel-zaj viszonyt ($> 30dB$) tudunk biztosítani, ezért a felismerés pontosabb. *Zajos környezet*ben speciális algoritmusokkal szükséges a beszédfelismerő zajtűrését, robusztusságát javítani (Acero–Stern 1990, Stern et al. 1992), a felismerés hatékonysága azonban várhatóan így is romlik a csendes környezettel összehasonlítva, ugyanis műszaki megoldással ma még csak az emberi percepciónál jóval korlátozottabban tudjuk a zajelnyomást megvalósítani. A *telefonbeszéd* kezelése gépi beszédfelismerővel a sávkorlátozott jelleg és az egyre gyakrabban alkalmazott tömörítő kódolás miatt megkülönböztetett és fontos felhasználási terület, mivel fontos szerepe van a telefonos információs rendszerekben, melyeket milliók használnak.

Szótárméret. A felismerőrendszer szótára az ember szókincsével hozható párhuzamba. Nagyobb szótárméret esetén nő az erőforrásigény. A szótár méretét erősen befolyásolja (korlátozza) egyrészt a valós idejű működés követelménye, másrészt az egyes szavak közötti akusztikai hasonlóság, továbbá az akusztikai környezet, a beszélőadaptáció léte vagy nem léte stb. Ennek alapján *kisszótáras* (néhányszor 10–100 szó), *középszótára*s (néhányszor 100–1000 szó) és *nagyszótáras* (néhányszor 1000–10 000 szó) felismerőkről beszélhetünk. Szótár megadása nélkül nem lehetséges korszerű beszédfelismerő rendszert készíteni, a szótárnak azonban nem feltétlenül kell teljes szóalakokat tartalmaznia. Toldalékoló nyelvek esetén – ilyen a magyar mellett az észt, a török és a finn nyelv is – morfémaszerű szótárelemek, illetve speciális szótárszimbólumok használatával érték el a legjobb felismerési eredményeket nagyszótáras, folyamatos beszédfelismerésben (Creutz et al. 2007, Arisoy et al. 2009, Tarján–Mihajlik 2010).

Üzemmód. A gépi beszédfelismerőket alapvetően kétféle üzemmódban használhatjuk: *parancsmód*ban valamilyen eszköz, számítógép vezérlése oldható meg beszédinterfészen keresztül, *diktáló üzemmód*ban pedig szövegszerkesztés jellegű munkához kapunk támogatást. Vegyük észre, hogy az előbbi tipikusan izolált szavas, az utóbbi pedig folyamatos felismerőt igényel. A dialógusalapú üzemmód jóval intelligensebb rendszert, a beszédet nem csak átalakító, de azt mélységében értelmező és megértő rendszert feltételez, ezért napjainkban ebben a körben az emberhez mérve csak szerény tudású vagy kísérleti alkalmazásokkal találkozhatunk.

Az üzemmód szerinti másik lehetséges csoportosítás az *online*, illetve *offline* működés. Ha a felismerő a felismerést a beszéd elhangzásával párhuzamosan végzi, ak-

kor online működésről beszélhetünk. Ha előre rögzített beszédet visszajátszva dolgozunk fel, akkor offline használjuk a felismerőt. Miután a beszédfelismerés igen műveletigényes, gyakorta problémaként jelentkezik a valós időben történő üzemelés iránti igény. Online működésnek gyakorlatilag csak akkor van értelme, ha a valós idejű futás nagyjából biztosított.

9.2. A beszéd gépi felismerésének alapjai

A fejezet bevezetőjében már említettük, hogy napjaink technikai színvonalán és jelenlegi tudásunk alapján a gépi beszédértés csak nagyon korlátozottan valósítható meg. A gépi beszédfelismerés kezdeteire jellemző „vak" beszéd-karaktersorozat átalakítás manapság már támaszkodik a nyelvi elemzésre is. Egyre inkább világossá válik, hogy a puszta beszéd-karaktersorozat átalakítás sem oldható meg a szövegfeldolgozás szintaktikai-szemantikai összefüggéseinek vizsgálata nélkül. E két szint összekapcsolásával működnek a mai felismerők. A beszédfelismerés műszaki megközelítésében elsősorban azt használják fel, amit az emberi beszédfeldolgozásról (beszédpercepció) jelenleg tudunk, de a beszédképzési ismeretekre is támaszkodnak. Alapkoncepciójukban a gépi beszédfelismerést végző eszközök egyrészről az emberi beszédpercepció hallásmechanizmusának tulajdonságait próbálják követni, másrészről nyelvi elemzéseket is használnak. A beszédfelismerők a beszéd valamilyen egységét – leggyakrabban a szót – veszik alapul, és a felismerés során valamiféle háttértárból (tudásbázis) nyert információk alapján végzik a beszéd osztályozására visszavezetett felismerést. Az osztályozás alapja valamilyen referenciamintával való összehasonlítás, illetve az ahhoz való akusztikai hasonlóság mérése.

A gépi beszédfelismerésben az alapvető célkitűzés géppel felismerni minden beszélő beszédét. Ehhez három alapvető problémát kell megoldani.

Egyrészt kezdenünk kell valamit a beszéd nagyfokú *akusztikai* (spektrális) *változatosság*ával. Mit jelent ez? A változatos akusztikai paraméterekkel rendelkező közegben, változatos akusztikai paraméterekkel jellemezhető beszélőktől elhangzó beszédet szeretnénk felismerni. Ezt a problematikát a mindennapi életünkben kevéssé érzékeljük, mivel az emberi percepció feldolgozási folyamatai lehetővé teszik, hogy ezek az akusztikai változékonyságok (akár zavaró jelek is) számunkra szinte észrevétlenek maradjanak.

Másrészt kezelnünk kell a beszéd nagyfokú *időbeli változatosság*át is. Az emberi beszédértést egyáltalán nem zavarják beszédsebességbeli eltérések (lassan, gyorsan beszélünk). Ezzel szemben az egyes beszédhangok időtartamának közlésen belüli és közlésről közlésre is változó értékei a műszaki gyakorlatban komoly kihívás elé állítják a gépi beszédfelismeréssel próbálkozót.

A harmadik megoldandó probléma annak eldöntése, hogy a beszéd akusztikailag igen gazdag, változatos, de redundáns paramétereiből melyek azok, amelyek a közölt információ velejét, lényegét hordozzák. Ehhez a beszéd matematikai eszközökkel kiterjesztett absztrakt modellezésére van szükség. Az ilyen modellek a beszédből kiemelt lényeges (invariáns) paraméterekre támaszkodva alkalmasak lehetnek a gépi beszédfelismerés valamilyen fokú elvégzésére.

Az előbbiekben vázolt három problémakör elvi megoldási lehetőségeit próbáljuk meg visszafelé haladva áttekinteni! A korábbi fejezetekben volt már szó a *lényegkiemelés*ről (7.1.6. fejezet), melynek során a beszéd akusztikailag releváns, ugyanakkor tömör reprezentálhatóságot adó paramétereit nyerik ki a beszédjelből. Leginkább a rövid idejű burkolóspektrum érzeti transzfomációján alapuló eljárások terjedtek el a beszédfelismerésben. A legkedveltebb a könnyen implementálható mel-frekvenciás kepsztrális együtthatók (MFCC) számításán alapuló eljárás. Érdemes még megemlíteni a lineáris predikciót alkalmazó, de az emberi beszédészlelés kifinomultabb modelljét használó perceptuális lineáris predikciót (Perceptual Linear Prediction, PLP) is, amely különösen zajban hetékonyabb, mint az MFCC (Hermansky 1990).

Az időbeli változatosság kezelésére egyrészt a beszédhangok további, kisebb részekre való felbontása, másrészt a dinamikus programozáson (Bellman 1957) alapuló mintaillesztési eljárások adnak lehetőséget, ezeket a következőkben részletesen ismertetni fogjuk. Alapvető megközeltés a *keretképzés*, ami a beszéd időbeli változatosságának rugalmas, dinamikus kezelésében hasznosnak bizonyul. Beszédfelismeréskor a tudásbázisban tárolt referenciamintákhoz időben illesztjük a felismerni kívánt bementi beszédmintát, majd valamilyen mérőszámmal kifejezzük a referencia és a bemenetről származó minta közötti hasonlóságot (hasonlósági mérték).

A beszélőn belüli és beszélők közötti akusztikai változatosság talán a legnehezebben kezelhető az említett három probléma közül. A probléma megoldása részben statisztikai eszköztárral, részben kompromisszumokkal történik. Statisztikai eszköztárral, a lényegkiemelt paramétereknek egyes beszédelemekre (például beszédhangokra) jellemző eloszlását kellő számú minta alapján jól meg lehet becsülni. Mint látni fogjuk, a statisztikai alapú beszédfelismerőkben éppen ezen alapul az egyes beszédelemek, leggyakrabban a beszédhangok modellezése. Fontos látnunk, hogy a lényegkiemelt paraméterek szórása igen jelentős, az egyes modellezendő beszédelemekre jellemző értéktartományaik pedig gyakran át is fedik egymást (lásd például az egyes magánhagzók formánsainak jellemző, egymást a határokon átfedő eloszlását az 5.1.2. fejezetben). Emiatt a jelentős mértékű szórás miatt kényszerülünk különböző kompromisszumokra, éppen a szórás kordában tartására: a beszéd akusztikai változatosságát okozó paraméterek behatárolásával erre van is némi lehetőségünk. A legtipikusabb behatároló kompromisszum az akusztikai környezet, amelyben a beszéd elhangzik. Napjainkban külön beszédfelismerők (legalábbis külön tudásbázissal rendelkezők) készülnek például irodai vagy otthoni, azaz csendes környezetben történő alkalmazásra, vagy például telefonos rendszerekben történő használatra. Az előb-

biek esetében további előny, hogy a felhasználók köre is jól behatárolható, illetve lehetőség van a felhasználó hangjához való alkalmazkodásra (beszélőadaptáció) is, ezáltal az interindividuális variancia kiküszöbölhető. Cserébe jóval nagyobb elemtárral (szótárral) rendelkezhet a felismerő, hiszen az egyes elemek közötti akusztikai határok nem mosódnak annyira össze. Ezzel szemben a telefonon keresztül működő beszédfelismerőben nem nagyon lehet megkötésünk a beszélő személyére vonatkozóan, így a magas interindividuális variancia az elemtár méretének jelentős csökkentésére kényszerít bennünket (ehhez még hozzáadódik, hogy a telefon átviteli sávja keskenyebb az optimálisnál).

A gépi beszédfelismerés megközelítése kapcsán összegzésként tehát az alábbiakat mondhatjuk: a beszédet valamiféle referenciamintával történő összehasonlítás során, az akusztikai hasonlóság mérésével ismerjük fel. Ehhez a felismerendő beszédmintát a referenciamintához kell illesztenünk úgy, hogy az egymásnak megfelelő részek időbeli, szakaszos, a beszédjel egészét tekintve nemlineáris nyújtással-zsugorítással egymás mellé kerüljenek. Fejlettebb megoldások esetén a beszédből fix időközönként kereteket képzünk, ezek a keretek egy-egy rövid idejű, fix időeltolással, tehát előre rögzítetten kiválasztott beszédjelrészlet akusztikailag lényeges (diszkriminatív) információtartalmát sűrítik magukba. A keretekre bontott lényegkiemelt beszédjel algoritmikusan igen hatékonyan illeszthető a szintén keretekkel reprezentált referenciához. A referencia készítéséhez a keretekben hordozott paramétereknek a modellezni kívánt beszédelemre vonatkozó jellemző eloszlásait becsüljük meg mintasokaság alapján (ehhez kellenek a beszédadatbázisok), majd azokból referenciamodellt alkotunk (gyakran gépi tanulás útján). A nyelvi modell alkalmazása jelentősen javítja a felismerés pontosságát (ahogy az embernél is). A gyakorlati alkalmazásokban a szóhibaarány a mérvadó. Ez több tényezőtől is függ, ezek közül a legmeghatározóbb a felismerendő egységek közötti akusztikai és nyelvi különbözőség. Tehát – a közhiedelemmel ellentétben – a szótár mérete önmagában nem determinálja a felismerés pontosságát: kevés számú, de egymáshoz nagyon hasonló rövid felismerési egység esetén nagyobb lehet a hiba, mint nagy számú, de egymástól akusztikailag és nyelvileg jól elkülönülő hosszú felismerési elem esetén. A környezet zajossága és a beszéd gondos vagy laza artikulációja, azaz formáltsága viszont minden esetben jelentős meghatározó tényező.

9.3. Lényegkiemelési eljárások

A gépi beszédfelismerő rendszerek legalsó szintje a jelfeldolgozó egység. Különböző rendszerek különböző megközelítéseket használnak, de közös bennük, hogy a beszédjelből a felismerés számára fontos, számítógéppel feldolgozható adatokat állítanak elő. A beszédkutatók jelentős eredményeket értek el az emberi beszédkeltő

és beszédfelfogó szervek működésének feltárásában, és elkészítették ezek matematikai modelljeit. Ezeken a modelleken alapulnak a jelfeldolgozás alapvető lépései. A modellezés nem tökéletes, és általában a gyakorlat dönti el, hogy melyik megoldás vezet a legjobb eredményekhez az adott feladat szempontjából. Egy beszédhangmodelleket tároló folyamatos beszédfelismerő felépítése a 9.8. ábrán látható. A jelfeldolgozó, vagy más néven előfeldolgozó egység tehát a beszédjelből transzformációk sorával a számítógép által értelmezhető adatsorozatot állít elő, amelyet *jellemzők*nek vagy *tulajdonságvektor*oknak hívunk, mivel valamilyen tulajdonságát jellemzik a beszédjelnek. A kiindulási alap a digitalizált beszédjel, és ebből állíthatók elő különböző jellemzők, mind az idő-, mind a frekvenciatartományban. Ilyen jellemzők az időtartományban például a rövid idejű energia, illetve amplitúdó burkoló, a nullátmenetek száma stb. A frekvenciatartományban általánosan elterjedten használják a rövid idejű spektrumot, valamint az erre épülő további feldolgozási algoritmusokat, amelyeket a könyv korábbi szakaszaiban részletesen bemutattunk (például MFCC). A beszédjel feldolgozása mellett a lényegkiemelésbe sorolható feladat a környezeti zajok mérése és jellemzőinek meghatározása is. Zajadaptációra, illetve -kompenzációra van szükség – egymikrofonos esetben az előbbi, több mikrofon, illetve mikrofonsorok esetén az utóbbi alkalmazható. Minden valós feladatnál komoly kihívást jelent a környezeti zajok kezelése, kezdve a számítógépek ventilátorának a zúgásától, egészen az elhaladó mentőautó szirénájáig. A tudomány mai állása szerint a fül harmonikus rezgéselemzést végez, amit kísérletek is bizonyítanak. Ezen álláspont szerint kézenfekvő jelfeldolgozási módszer az összetett rezgések elemzésére kidolgozott Fourier-féle elemzés. A beszédfelismerő rendszerek általában további transzformációkat (akár párhuzamosan többet is) hajtanak végre az FFT-vel kinyert spektrális paramétereken. Ezek közül a feladat jellege, kísérleti optimalizálás vagy szakirodalmi adatok alapján lehet választani. Az alkalmazott technikák általában a szűrősoros elemzés, kepsztrumelemzés (e kettő kombinációja adja az MFCC együtthatókat), illetve a lineáris prediktor együtthatóival történő reprezentáció. Mind a szűrősorokkal, mind a kepsztrumparaméterekkel jobb felismerési eredmények érhetőek el, mint önmagában az FFT-együtthatókkal, továbbá ezek az eljárások az egy kerethez tartozó jellemzők számát, azaz a tulajdonságvektor dimenzióját is csökkentik (a jellemzőket dekorrelálják), így a további feldolgozás sebességét növelik. A lényegkiemelés eredménye tehát a beszéd tömör, tulajdonságvektorokba sűrített reprezentációja. A tulajdonság- vagy más néven jellemzővektorok a rövid idejű, gördülő elemzés révén fix időközönként követik egymást. A tulajdonság-, illetve jellemzővektor elnevezés mellett a *keret* elnevezés is használatos. A keretek közötti időeltolás tipikusan 10 ms, vagy ehhez közeli érték, de semmiképpen sem több, mint az elemzőablak hosszának 50%-a. Jelenleg nem ismerünk olyan jellemzőhalmazt, amely egyértelműen képviseli a beszédhangokat. Sőt, mint említettük, az egyes beszédhangok különböző reprezentációihoz tartozó tulajdonságvektorok által alkotott térrészek nem diszjunktak az eddig ismertetett paraméterek esetén sem. Az emberi

fül sem tökéletes: rövid – néhány hangból álló – egységeket nehezen, néha hibásan értünk, mégsem jelent problémát a folyamatos szöveg megértése, mivel az agyban működő komplex nyelvi feldolgozás korrigálja a hibákat. A lényegkiemelés során a célunk tehát olyan jellemzők keresése, amelyek minél hatékonyabb mintaillesztést tesznek lehetővé. A korábban bemutatott spektrális jellemzők mellett többek között használják még az alapfrekvenciát, a formánsfrekvenciákat, a beszéd/nem beszéd, illetve a zöngés/zöngétlen paramétereket is.

9.3.1. Normálás

A beszédfelismerés során lényegében osztályozást végzünk, így hasznosnak bizonyul, ha a különböző beszédelemek (leginkább a beszédhangok) tulajdonságvektorai egymástól minél jobban elkülönülnek, az azonos osztályba tartozók pedig minél hasonlóbbak, „tipizáltak". Ezt elsősorban a későbbi modellezés során tudjuk megvalósítani, vannak azonban olyan paraméterek is, amelyek az osztályozás szempontjából nem relevánsak, ezeket eltávolítva lényegében a felesleges, jelfeldolgozási értelemben vett zajtól tisztítjuk meg a beszédjelet. Ezt a célt szolgálja a normálás, melynek során az osztályozás szempontjából irreleváns, de az akusztikai varianciát feleslegesen növelő paraméterváltozásokat kiszűrjük. Az ember megérti, ha valaki hangosabban vagy halkabban beszél, sőt még azt is, ha énekel. Így például a hangerő, helyesebben annak fizikai megfelelője, a jel energiája egységes tartományba transzformálható (például erősítéssel, maximális szintig történő kivezérléssel). Fontos tudnunk azonban, hogy a normálás az egyes beszédhangok közötti, specifikusan jellemző energiaszintbeli különségeket nem érintheti. A normálást tehát – a feldolgozás időléptékeit alapul véve – hosszú beszédszakaszra kell végrehajtanunk (jellemzően két levegővétel, vagy hosszabb szünet közé eső prozódiai frázisokra). A normáláshoz tehát a beszédjel egy-egy kiválasztott átlagos tulajdonságát megmérjük egy hosszabb időablakban, és az így kapott számértéket a tulajdonságtól függően levonva, vagy a számértékkel osztva előállítjuk a normált jelet. Néhány általánosan használt normálási eljárás (Mihajlik 2010):

Egyenszint-kompenzáció: A beszédjel egyenszintjét (a jel átlagát) ki kell vonni a jelértékekből.

Jelszintnormálás: úgy növeljük a jel amplitúdóját, hogy a legnagyobb érték a teljes kivezérlést adja. Csökkentésről itt nem lehet szó, mivel akkor túlvezérelt jelet kellene lehalkítani. Ezzel nagyjából egyformán hangosra állítjuk a feldolgozandó jeleket. Rövid idejű, de nagy amplitúdójú zajok esetén előfordul, hogy azokhoz igazítják a maximumot, ekkor a beszédjel továbbra is halk marad.

Kepsztrum-átlagkivonás: A logaritmikus frekvenciatartományban a lineáris torzítás eltolásként, additív tagként jelentkezik a hosszú idejű átlagban. Ennek kompen-

zálására a hosszú idejű átlagot kivonjuk az egyes log-spektrum elemekből. Ez jól használható a lineáris torzítás kompenzálására (amelyet például a különböző mikrofonok okoznak). Ha alkalmazzuk, akkor a numerikus pontatlanságok miatt szélsőséges esetekben romolhat a felismerés pontossága, azonban általában hatékonyan használható.

9.3.2. A tulajdonságvektorok előállítása

A tulajdonságvektorok előállításához a beszédjelet felbontjuk periodikusan egyenlő méretű (tipikusan 20–30 ms), átlapolódó darabokra (ablakolás). Minden ablakra kiszámítjuk a tulajdonságvektort, ami az előző fejezetben leírt tulajdonságok alapján jellemzi a beszédjel ablakolt részét. A vektorban az egyes tulajdonságok (paraméterek) jellemző számértékei foglalnak helyet egymás alatt. Ez a vektorsorozat képezi a mintaillesztő egység bemenetét. A paraméterek kiválasztása függ a feldolgozó rendszer képességeitől (sávszélesség, memória, sebesség), valamint a feladattól is. Minél komplexebb a feladat, annál pontosabb és komplexebb paraméterek számítására van szükség. Egy beszéddetektor esetén akár egyetlen elem, az energia is elegendő lehet, de felismerőkben tipikus a 12 elemű kepsztrum, vagy mel-kepsztrum vektor (lásd a 7.1.6. fejezetben).

A paramétervektorok mellett általában megadjuk azok változását is, ezeket nevezzük delta, illetve delta-delta paramétereknek. Ezek az aktuális ablak előtti, illetve utáni néhány (2–3) vektor lineáris kombinációjaként számíthatóak, lényegében a deriválás numerikus megvalósítását végző algoritmusokkal egyező módon. A delta paraméter tehát a jellemző változásának mértékét, míg a delta-delta paraméter a változás sebességét adja vissza. A paramétervektorok előállítása során az egymást követő ablakokat (kereteket) függetleneknek tekintjük, ami nyilván helytelen feltételezés, hiszen a beszédjel erősen redundáns, de jelenleg nem ismert ezt hatékonyan kezelő módszer. A delta paraméterek számítása során azonban képletesen figyelembe vesszük ezt az összefüggést, hiszen több egymást követő keretből (szegmens) jön létre az eredmény.

A mai hagyományos és célprocesszorok alkalmazásával nem jelent problémát a fent említett vektorok előállítása, ehhez a valós idő töredéke, néhány százaléka szükséges csupán. A feldolgozási idő nagyobb részét a következő szint, a mintaillesztés veszi el. Ezért is szükséges minél kompaktabb paraméterhalmaz előállítása. Amennyiben 20 ms-os ablakkal és 22 kHz-es mintavételi frekvenciával dolgozunk, akkor egyetlen ablakban 440 beszédminta van. Ezzel szemben a tulajdonságvektor a fentiek szerint mintegy 36 elemet tartalmaz, ami jelentős adatmennyiség-csökkenést jelent. A 36 elemet az eredetileg meglévő 12 együttható első és második deriváltjaival kiegészítve kaptuk.

9.4. Mintaillesztési eljárások

A beszédfelismerőkben a lényegkiemelést végző előfeldolgozó modult a mintaillesztő követi. A mintaillesztő feladata a bemenetként kapott tulajdonságvektor-sorozat leképzése felismert szimbólumok – nyelvi elemek – sorozatává, melyek további processzálásra az utófeldolgozóba kerül(het)nek. Fontos látni, hogy a bemeneti jel még akusztikus információkat tartalmazó állandó ütemezésű jelfolyam, míg a kimenet a már sokkal absztraktabb nyelvi egységek sorozata. Vagyis a kritikus lépés, az akusztikus/nyelvi konverzió – azaz felismerés – itt történik meg. A mintaillesztő tehát maga a szűkebb értelemben vett beszédfelismerő. Két alapvető mintaillesztési technika alakult ki: a *sablon (template) bázisú mintaillesztés* (például dinamikus idővetemítés alapú eljárások), valamint a *statisztikai* mintaillesztési eljárások (például rejtett Markov-modell, illetve mesterséges neurális háló segítségével). Az első módszer történetileg előbb alakult ki, azonban a korszerű felismerők túlnyomó többsége a második típusú mintaillesztést alkalmazza, a gyakorlatban a leghasználhatóbbnak a tisztán rejtett Markov-modelles felismerők bizonyultak, így a mintaillesztés áttekintése során a hangsúlyt ez utóbbira helyezzük. Minden mintaillesztési eljárás azon alapul, hogy a felismerni kívánt beszédelemekről (például szavak), vagy beszédelemek láncolatairól (beszédhangok és a belőlük felépülő szavak) referenciamodelleket alkotunk (tanítás), azokat a rendszerben tároljuk. A mintaillesztés során megkeressük a felismerő bemenetéről érkező minta és valamennyi referenciaminta közötti legjobb illeszkedést. Ez alapján számítjuk ki valamennyi referenciamintára a hasonlóság mértékét az adott referencia és a bemeneti beszédminta között. Bonyolultabb alkalmazásokban (folyamatos beszédfelismerők) a referenciák maguk is bonyolult hálózatokat képezhetnek, amelyek mentén – bizonyos szabályszerűségek tiszteletben tartása mellett – szabadon vándorolva keressük a legjobb illeszkedést.

9.4.1. Sablonbázisú mintaillesztés

A sablonalapú mintaillesztés során előre tárolt, a felismerendő beszédegységhez tartozó mintákhoz hasonlítja a rendszer a mért tulajdonságvektor-sorozat egyes elemeit. A sablonalapú mintaillesztést kisszótáras, izolált szavas felismerőkben használják. A felismerő ekkor tehát sablonokkal dolgozik. Hogyan készül a sablon? Hangfelvételeket készítünk a felismerni kívánt beszédelemekről (leginkább szavakat, esetleg rövid mondatokat), azokon végrehajtjuk a lényegkiemelést, majd az előállított tulajdonságvektorokhoz hozzárendeljük a szó- vagy mondatcímkéket, és ezeket mint referenciákat sablonként eltároljuk. A beszédfelismerő számára legalább annyi sablont kell biztosítani ahányféle beszédelemet (szót stb.) fel akarunk majd ismertetni. A felismerés során ezeknek a sablonoknak a segítségével történik meg a bemene-

ti jellemezővektor-sorozat leképezése szimbólumokra (azaz lényegében szavakra). A beszéd nagyfokú akusztikai változatossága miatt célszerű, ha az egyes referenciaminták egy vagy több embertől gyűjtött minták átlagolásával állnak elő. A referenciaminták mint jellemzővektor-sorozatok hossza változó. Nyilván egy szóhoz tartozó referencia több vektort tartalmaz, mint egy hanghoz tartozó. Egy-egy szimbólumhoz akár több különböző referencia is megadható az esetleges nagy különbségek esetén, amennyiben a felismerő algoritmus nem képes ezt kezelni. Például ugyanazt a hangot néha hosszabban máskor rövidebben ejtjük, ezt akár különböző referenciával is ábrázolhatjuk, ugyanígy oldható meg a nyelvjárások kezelése is. Felismeréskor a mintaillesztő a beérkező vektorsorozatot sorban összehasonlítja az összes tárolt referenciával, és mindegyikre meghatározza a hasonlóság mértékét. Ehhez még két dolog szükséges: egyrészt meg kell adnunk a hasonlóság mértékének számítására szolgáló algoritmust, másrészt biztosítanunk kell, hogy a beérkező minta, valamint a tárolt referencia között lévő időtartambeli különbségek – ésszerű időtartambeli ingadozásokat lehetővé téve – ne befolyásolhassák számottevően a hasonlóság mérését, ezeket az ingadozásokat (időbeli változatosság) ugyanis az emberi beszédértés is igen széles tartományban tolerálja. A hasonlóság mérőszáma a jellemzővektorok közötti *távolság*ként határozható meg. A vektoronkénti távolságszámításnak különböző módjai vannak. Alapvető és egyben a legelterjedtebb módszer ha a térben lévő legközelebbi elemre döntünk a d_ε euklideszi távolság alapján:

$$d_\varepsilon(\mathbf{r},\mathbf{o}) = \sqrt{\sum_{n=1}^{N}(r_n - o_n)^2}, \tag{9.1}$$

mely az N-dimenziós $\mathbf{r}$ és $\mathbf{o}$ tetszőleges vektorok között értelmezett és koordinátánként számítható a fenti összefüggés szerint. Mivel a távolságot vektorról vektorra számítjuk, a vektorsorozatokból álló referenciaminták és a beérkező egyetlen vektorsorozat közötti D összegzett távolságok meghatározásához a vektoronkénti távolságokat összeadhatjuk, majd normáljuk a vektorok számával elosztva annak érdekében, hogy az eltérő hosszú referencia-vektorsorozatoktól számított összegzett távolság a vektorsorozat hosszától független legyen. Az $\mathbf{R} = \mathbf{r_0}, \mathbf{r_1}, ...\mathbf{r_M}$ referenciavektorsorozatra a beérkező $\mathbf{O} = \mathbf{o_0}, \mathbf{o_1}, ...\mathbf{o_M}$ vektorsorozat összegzett távolsága így:

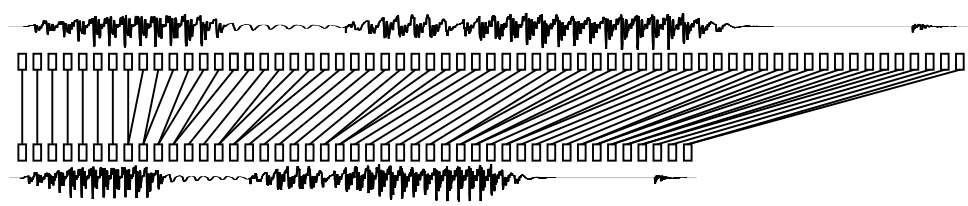
$$D(\mathbf{R},\mathbf{O}) = \frac{1}{M}\sum_{i=1}^{M} d_\varepsilon(\mathbf{r_i},\mathbf{o_i}), \tag{9.2}$$

ha feltételezzük, hogy mindkét vektorsorozat M vektorból áll és az artikulációs tempó állandó. Ugyanakkor meg kell még oldanunk azt a problémát, hogy mi történjék, ha a beérkező tulajdonságvektor-sorozat elemszáma nem egyezik valamelyik referenciaminta jellemzővektorainak elemszámával (az azonos elemszám igen valószí-

nűtlen, hiszen közelítőleg 10 ms-os időfelbontásról van szó). Erre szolgál a *dinamikus idővetemítés*.

9.4.1.1. A dinamikus idővetemítés

Tudjuk tehát, hogy két vektorsorozat egymáshoz viszonyított távolságát szeretnénk meghatározni, ugyanis ennek alapján tudjuk eldönteni, hogy melyik sablon illeszkedik legjobban a bemenetre érkező, felismerendő beszédmintára. Adott tehát egy felismerendő beérkező vektorsorozat és sok tárolt referenciavektor-sorozat (sablon). Meg kell mondani, hogy melyikhez illeszthető legjobban a felismerendő sorozat. Nagy valószínűséggel a referenciák hossza különböző, és a beérkező sorozat hossza is különbözhet ezektől, mivel ugyanazt a szót különböző sebességgel mondhatjuk ki. A beérkező és a referenciavektorokat azonban valahogy párba kell állítani a távolság méréséhez. Erre a legegyszerűbb, bár önmagában nem használható módszer a lineáris idővetemítés. Egyszerűen valamelyik időtengelyt a másikhoz nyújtjuk lineárisan, és ahol szükséges, a köztes vektorokat ismétléssel vagy interpolációval beszúrjuk (lásd 9.1. ábra). Az így párba állított vektorok távolságát összegezzük, és a legkisebb távolságra, azaz a „leghasonlóbb" referenciára döntünk. Ezen algoritmus alkalmazhatóságának az szab határt, hogy a szavakat nem arányosan mondjuk gyorsabban, vagy lassabban, hanem a szavakon belül is lehetnek sebességkülönbségek. Ennek kiküszöbölésére használható a dinamikus idővetemítés (Dinamic Time Warping, DTW). Ez egy nemlineáris transzformációt hajt végre az időtengelyeken, és ezek mentén hasonlítja össze a vektorsorozatokat. Legyen a referencia vektorsorozat



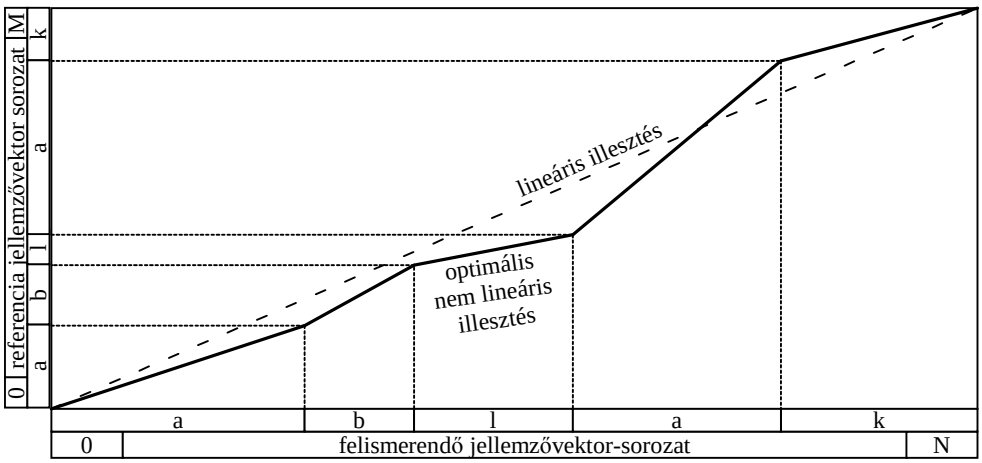
9.1. ábra. A DTW működése: a tulajdonságvektorok párba állítása hasonlóság alapján

$\mathbf{R} = r[0], r[1], ..., r[S]$, a beérkező vektorsorozat: $\mathbf{O} = o[0], o[1], ..., o[T]$, és a vetemítőfüggvények az $F_r(i)$, és az $F_o(i)$ függvények; ezek közül keressük az optimálisat. Így a távolság:

$$D(\mathbf{R}, \mathbf{O}) = \min_F \sum_i d_\varepsilon \left(r[F_r(i)], o[F_o(i)] \right). \tag{9.3}$$

A vetemítő út az $(r[0], t[0])$ pontból indul, és $(r[S], o[T])$ pontba érkezik; monoton nő, hisz nem beszélünk visszafelé; nem lehet akármekkorát ugrani, hisz nem beszélhetünk tetszőlegesen lassan, vagy gyorsan. A konkrét implementációban lokális felté-

teleket lehet megadni a továbblépésre, így az F függvény által megadott út nem lehet tetszőleges (9.2. ábra). A fenti 9.3 összefüggésben a vektorsorozat elemszámára tör-



9.2. ábra. A DTW vetemítőfüggvénye (vetemítő út)

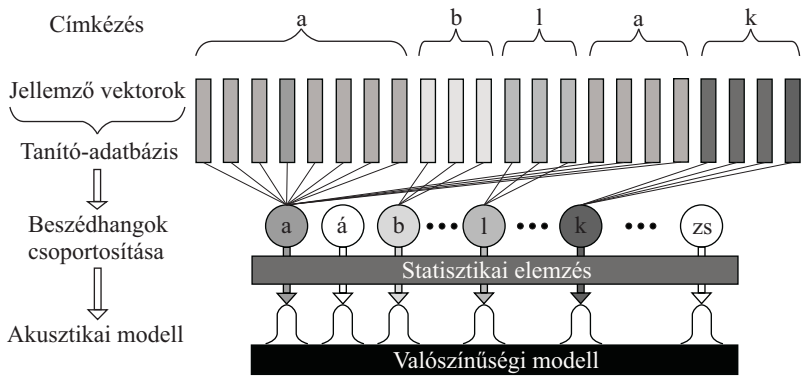
ténő normálást elhagytuk és formalizált jelöléseket alkalmaztunk a jobb áttekinthetőség érdekében. A DTW nagy előnye, hogy igen gyorsan – akár egyetlen mintával is – hatékonyan *tanítható*. Robusztusabbá tehető a felismerő több tanítóminta alkalmazásával, ezek a vetemítő út mentén átlagolhatóak. Amennyiben elegendő mennyiségű tanítóminta áll rendelkezésre, akkor már általában statisztikus módszereket alkalmaznak, például a rejtett Markov-modelleken alapuló felismerőt, mely gyakorlatilag magában foglalja a DTW algoritmust is.

A sablonbázisú, dinamikus idővetemítéssel végrehajtott mintaillesztés kiválóan alkalmazható kisszótáras beszédfelismerésben, vagy a felhasználó által bemondott és rögzített szavak, rövid kifejezések felismerését igénylő kisebb alkalmazásokban (például mobiltelefonok). Kis erőforrásigényű, valós időben működő algoritmus. Hátrányai a szótár bővülésével kerülnek előtérbe: a sablonbázisú mintaillesztést alkalmazó beszédfelismerők legfeljebb egy-kétszáz szavas szótárral használhatók, e feletti méret esetén a tévesztések száma már zavaróan magas. Hátrányt jelent, hogy a kisebb mértékű szótárbővítés is csak újabb referenciák létrehozásával lehetséges, ehhez gyakran beszédfelvételeket is kell készíteni. Jóllehet az alkalmazhatósági korlátok eleve lehetetlenné teszik a sablonbázisú mintaillesztés nagyszótáras rendszerben való alkalmazását, de nagy elemszám esetén a háttértárigény is jelentősen növekedne a redundáns reprezentáció miatt: a jellemzővektorok ugyan a lehető legnagyobb fokban tömörítettek, de túl sűrűn követik egymást (egy-egy beszédhang hossza 100 ms nagyságrendjében is lehet, ez beszédhangonként több, egymáshoz igen hasonló vektort jelent), ugyanakkor ritkításuk szinte megoldhatatlan, mert a nyers beszédjelből

semmilyen támpontunk nincs arra vonatkozóan, hogy mely részekből kellene a jellemzőket kinyernünk és melyek azok, ahol ritkábban is elégséges a reprezentálásuk. További probléma, hogy az egyes beszélők ejtéséből adódó akusztikai különbségeket a sablonbázisú rendszer csak igen korlátozottan képes figyelembe venni.

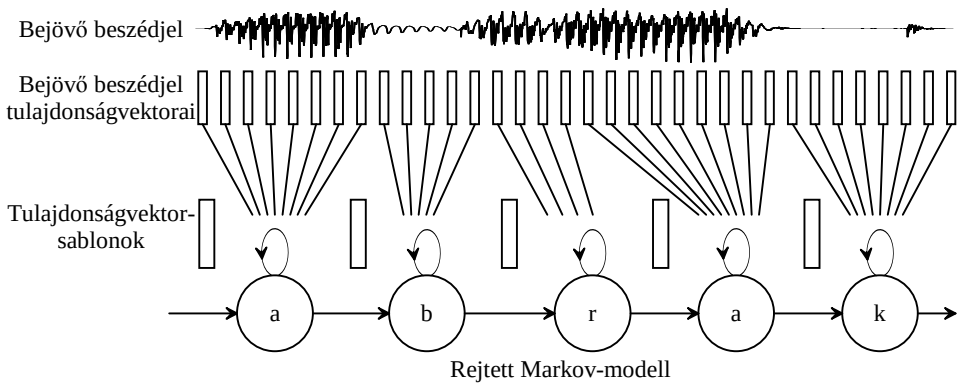
9.4.2. Statisztikai mintaillesztési módszerek

A sablonbázisú mintaillesztés hátrányait küszöbölik ki a mintaillesztést statisztikai alapokon végző beszédfelismerők. Ezeknek az eljárásoknak a közös tulajdonsága, hogy valamilyen akusztikai modell(eke)t használnak, melyekben az akusztikus információ tárolása és felhasználása hatékonyabban valósul meg, mint a sablonalapú módszernél. Ennek ára a modellparaméterek becsléséhez szükséges nagyobb beszédadatbázis a tanításhoz, illetve a megnövekedett számítási idő. Fontos megjegyezni, hogy a statisztikus jelző nem magára a mintaillesztési folyamatra vonatkozik, hanem a modellparaméterek kiszámítására, becslési módjára. A statisztikus mintaillesztés esetén a referencia jóval absztraktabbá válik. Bár maga a sablon is egyfajta absztrakt reprezentáció, statisztikai eszközökkel azonban lehetőség nyílik még magasabb szintű absztrakció megvalósítására. Az akusztikailag hasonló, vagy egy fonémához tartozó, de annak különféle realizációjából származó beszédrészeket egyetlen abszrakt – elemi akusztikus – modellként kezeljük. Például a [t] hang zárfelpattanásaiból vagy egy nazalizált [o] hang lecsengéseiből származó tulajdonságvektorokat összefogjuk és statisztikailag leírjuk. Ekkor egyetlen elemi akusztikai modellel jellemezhetjük a különféle beszélőktől és környezetből származó hangrészletet. Ha a koartikuláció jelenségétől eltekintünk, egy elemi modell egy komplett beszédhangot is leírhat. A szó egésze ekkor nyilvánvalóan elemi modellek sorozataként áll elő. A legegyszerűbb eset tehát, ha a szavakat felépítő egyes beszédhangokat feleltetjük meg egy-egy elemi modellnek, ekkor az egyes beszédhangokra jellemző tulajdonságvektorokból kell valahogyan a modellek lényegét meghatározni. Ehhez hívjuk segítségül a statisztika eszköztárát. Mielőtt ennek módját röviden ismertetnénk, megjegyezzük, hogy a mintaillesztés ezek után a dinamikus idővetemítésben megismertek szellemében történik, a különbség csak annyi lesz, hogy a beérkező vektorsorozat elemeit nem másik vektorsorozathoz, hanem elemi modellek sorozatához illesztjük. A elemi modellek közötti átjárást pedig nyelvi szabályok határozzák meg (például csak értelmes szavakat, szókapcsolatokat rakhatunk össze). Az egyes beszédelemek absztrakt modelljeként használt elemi modellek kialakítása a következők szerint történhet: alkalmasan kiválasztott beszédadatbázisból minden egyes beszédelemre (beszédhangra) mintákat válogatunk ki ügyelve arra, hogy az egyes minták az adott beszédelem valamennyi olyan akusztikai változatát tartalmazzák, amelyre a felismerés során számítanunk kell (így lesz pontos az eloszlásbecslés). Ezeknek a mintáknak kiszámítjuk a



9.3. ábra. Statisztikus mintaillesztés véges automatával (rejtett Markov-modellel)

tulajdonságvektorait, majd megbecsüljük a tulajdonságvektorok sokdimenziós sűrűségfüggvényeit a teljes paramétertérben. Tulajdonképpen ezek a sűrűségfüggvények lesznek az elemi akusztikai modellek (9.4. ábra). Az eloszlásbecsléshez tudnunk kellene, hogy milyen eloszlás sűrűségfüggvényét keressük. A gyakorlatban legikább az olyan jellegű eloszlások váltak be erre a feladatra, amelyek normális eloszlást, vagy normális eloszlások lineáris kombinációiból előállítható eloszlást adnak meg (angol nevük GMM, azaz Gaussian Mixture Model), mivel ezek univerzális függvényapproximátoroknak felelnek meg. A feladat tehát visszavezethető arra, hogy az N-dimenziós paramétertérben adott mintahalmazból N-dimenziós normális eloszlásokat becsüljünk meg. A statisztikából tudjuk, hogy a normális eloszlás paramétereinek torzítatlan és konzisztens becslését kapjuk, ha az eloszlás várható értékét a mintaátlag-statisztikával, a szórását pedig a tapasztalati szórás statisztikával becsüljük meg. (Ezek egyben maximum likelihood becslések is.) Ezután már csak egyetlen probléma van a modellünkkel: nincs semmilyen információnk arról, hogy az egyes elemi modellek milyen sokáig (hány keretnyi ideig) tartanak. Bár az egy elemi modellhez tartozó vektorok megszámlálása és az átlag eltárolása kézenfekvőnek tűnik, előnyösebb, ha az állapothossz-információt átmeneti valószínűségek bevezetésével reprezentáljuk. Alapesetben csak a következő elemi modellbe jutás, illetve helyben maradás valószínűségét érdemes 0-nál nagyobbra választani (minden ütemben lépni kell egyet, legfeljebb az addigi modellben maradunk a lépéskor). Ekkor világos, hogy a hosszabb beszédelemeknél saját modellbe való jutás (azaz helyben maradás) valószínűsége nagy lesz, a következőbe jutásé kisebb. Rövidebb beszédelemeknél a különbség kisebb lesz ezen valószínűségek között, tehát a két átmeneti valószínűség aránya jelzi a beszédelem hosszát. Itt megállhatunk, hiszen eljutottunk a beszédfelismerésben oly sikerrel alkalmazott folytonos megfigyelési-sűrűségfüggvényű rejtett Markov-modellig. A következőkben egy egzaktabb matematikai definíciót adunk meg, majd a mintaillesztési-felismerési folyamatot mutatjuk be.



9.4. ábra. A rejtett Markov-modell által végzett, absztrakt dinamikus idővetemítés a mintaillesztésben

Megjegyzés:

- Mivel a mintaillesztés – legyen akár sablonbázisú, akár statisztikai – dinamikus idővetemítéssel történik, a beszédfelismerésben a rövid és hosszú beszédhangok megkülönböztetése háttérbe szorul, hiszen a dinamikus vetemítés eleve a beszédhangok hosszában meglévő különbségek kiküszöbölésére hivatott, ez járulékosan rontja a rövid és hosszú beszédhangok megkülönböztetésének lehetőségét is. Emiatt a mintaillesztésben gyakran nem is tesznek különbséget a rövid és hosszú beszédhangok között, ez ugyanakkor azt is jelenti, hogy például a *kasza* és *kassza* szavak a mintaillesztés alapján gyakorlatilag megkülönböztethetetlenek. A kettő között mégis tudunk majd különbséget tenni, mégpedig a kontextus segítségével, erről részletesebben a későbbiekben lesz szó az úgynevezett nyelvi modell kapcsán. (A kontextus az emberi beszédértésben is szerepet kap, gondoljunk például arra, hogy a többjelentésű és azonos alakú szavak esetén az aktuális szövegbeli jelentés a kontextusba helyezve többnyire már egyértelmű.)

9.5. A beszéd-szöveg átalakítás alapjai

A gépi beszédészlelés megvalósítása történetileg a beszéd-szöveg átalakítás igényével kezdődött, de mint azt a felismerési feladatok ismeretésénél láttuk (9.1. fejezet), ma már tágabb értelemben értelmezzük és egyéb felismerési feladatok társulnak hozzá, egyre inkább a komplex gépi beszédértés irányába tolódva. A következőkben részletesen bemutatjuk a beszéd-szöveg átalakítás menetét, hiszen ez a feladat a tágabb területet felölelő gépi beszédészlelésben továbbra is kulcsfontosságú, illetve a nem közvetlenül beszéd-szöveg átalakítást igénylő, de a gépi beszédészlelés feladat-

körébe sorolható feladatok eszköztárának is a beszéd-szöveg átalakításra kidolgozott modellezési és algoritmikus eljárások képezik az alapját.

9.5.1. A beszédfelismerési feladat matematikai megfogalmazása

Beszédfelismerés során kiindulási alapunk a beszédjel, a keresett felismerési eredmény pedig az a karaktersorozat, amely az elhangzó beszédnek (leginkább) megfelelhet. Tekintsük kiindulásként az izolált szavas felismerés esetét, amikor a beszédjel ismeretében keressük a felismert szót ($\hat{w}$). A beszédjelen lényegkiemelést hajtunk végre, az így előálló tulajdonságvektor-sorozatot szokás *megfigyelés*nek (observation) nevezni, jelöljük ezt most $\mathbf{O}$-val: $\mathbf{O} = \mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_T}$. Mint már láttuk, a felismerést úgy végezhetjük el, hogy minden szóba jövő, azaz a felismerő által ismert, és emiatt a felismerő szótárában szereplő szóra (mint referenciára) kiszámítjuk a hasonlóság mértékét és a beérkező mintára leghasonlóbb szóra döntünk. A matematika nyelvén mindezt úgy is megfogalmazhatjuk, hogy azt a szót keressük, amelyik esetében a felismerő bemenetén megfigyelt tulajdonságvektor-sorozat realizálódására a legnagyobb esélyünk van, azaz a $P(w_i|O)$ valószínűségeket keressük minden w_i szóra. Az az i, amelyre ez az érték maximális, a felismert szó ($\hat{w}$):

$$\hat{w} = \arg\max_i P(w_i|\mathbf{O}). \tag{9.4}$$

Ha tehát ismerjük a $P(w_i|O)$ valószínűségeket, akkor a felismerési feladatot el tudjuk végezni. Bayes tételét alkalmazva a keresett feltételes valószínűségekre a következőt írhatjuk:

$$\hat{w} = \arg\max_i P(w_i|\mathbf{O}) = \arg\max_i P(\mathbf{O}|w_i) \cdot \frac{P(w_i)}{P(\mathbf{O})} = \arg\max_i P(\mathbf{O}|w_i) \cdot P(w_i), \tag{9.5}$$

azaz ha ki tudjuk számítani $P(\mathbf{O}|w_i)$ és $P(w_i)$ értékeket, meg tudjuk oldani a feladatot. Az előbbi (9.5) összefüggésben az utolsó lépésben $P(\mathbf{O})$-t azért hagyhattuk el a nevezőből, mert az az i szerinti maximalizálást nem befolyásolja. Amire tehát szükségünk van, azok olyan modellek, amelyek lehetővé teszik $P(\mathbf{O}|w_i)$ és $P(w_i)$ értékeinek kiszámítását. Az előbbit lehetővé tévő modell neve *akusztikai modell*, utóbbi valószínűségeket pedig a *nyelvi modell* alapján számolhatjuk. Az akusztikai modell feladata tehát az, hogy egy adott megfigyelésre megadja valamennyi, a felismerő szótárában lévő szó előfordulási valószínűségét (azaz lényegében egy olyan értéket, amelynek alapján a szótár szavait mint felismerési hipotéziseket rangsorolni lehet). A nyelvi modell pedig azt adja meg, hogy az adott szó előfordulása a beszédben milyen valószínű (gyakori szóra nagyobb, ritkára kevesebb ez az érték). Mint látni fogjuk, ezek a valószínűségek különböző trükkökkel és bizonyos megkötések-

kel már származtathatóak. Vegyük észre azt is, hogy ha a fenti összefüggésekbe az egyetlen szót jelölő w_i szimbólum helyett a $W = w_1, w_2, ..., w_n$ szósorozatot jelölő szimbólumot helyettesítjük, minden eddigi megállapításunk érvényben marad, de szósorozatok esetén keressük az akusztikai valószínűségeket, illetve szósorozatok valószínűségeit keressük a felismeréshez. Így lépünk tovább az izolált szavas felismerésről a folyamatosra. Ha pedig tudjuk, hogy a szavakat beszédhangok építik fel, akkor semmi akadálya annak, hogy akusztikai modellként ne az egyes szavakra, hanem az egyes beszédhangokra vonatkozó modelleket hozzunk létre, amelyeket összefűzve könnyedén konstruálhatók a szavak modelljei is, ráadásul e módszer jóval rugalmasabb felépítést tesz lehetővé és jelentősen csökkenti a modellek tárolásához szükséges erőforrásigényt. Ez utóbbi megoldás a beszédhangalapú, folyamatos felismerés, amelyet napjainkban a leginkább alkalmaznak. Arra is felhívjuk a figyelmet, hogy ha szósorozatokat tekintünk, akkor a nyelvi modell által meghatározandó $P(W)$ nem más, mint a nyelv egy sajátos – statisztikai – szintaxismodellje: azt adja meg, milyen szósorozatok milyen valószínűség szerint fordulnak elő a felismerő által ismert nyelvben. A nyelvi modell tehát felfogható úgy, mint egy igen rugalmas szabályrendszer, amely a szavak egymás után való fűzésére vonatkozóan tárol információkat.

9.5.2. Beszédfelismerés rejtett Markov-modellel

A továbbiakban W feleljen meg a $W = w_1, w_2, ..., w_n$ szósorozatnak. Láttuk, hogy $P(\mathbf{O}|W)$, illetve $P(W)$ valószínűségeket kellene ismernünk vagy megbecsülnünk a felismerés végrehajtásához. A Markov-modelles beszédfelismerés során azt feltételezzük, hogy a felismerendő beszédjelet, illetve annak $\mathbf{O} = o_1, ..., o_t, ..., o_T$ tulajdonságvektorait egy Markov-folyamat generálta, amely periodikus időközönként (keretidőnként) állapotot változtathat vagy azonos állapotban marad, de minden egyes állapotból, ahol éppen tartózkodik, egy tulajdonságvektort bocsát ki (keretidőnként az új, vagy helybenmaradás esetén a régi-új állapotnak megfelelően egyet). Ez a kibocsátott vektor egy valószínűségi vektorváltozó, amely az adott j állapotra jellemző $b_j(o_t)$ eloszlás szerint generálódik. Az egyes állapotok közötti átmenet is valószínűségi változó ($a_{i,j}$ jelöli az i állapotból a j állapotba történő átmenet konstans valószínűségét), méghozzá valamilyen diszkrét eloszlással. A mintaillesztés során azt az állapotsorozatot keressük, amelyen a folyamat végighaladhatott, miközben a felismerendő beszédjelet generálta. Mivel ez az állapotsorozat előzetesen ismeretlen, ezért a generáló modell *rejtett modell*. A mintaillesztés során azt feltételezzük, hogy az az állapotsorozat felel meg a felismert szószekvenciának (vagy e szószekvencia egyes beszédhangjainak), amelyre teljesül, hogy arra az összes lehetséges alternatíva közül a legnagyobb az éppen megfigyelt tulajdonságvektorokkal azonos tulajdon-

ságvektorok kibocsátásának valószínűsége. Tehát ha feltételezzük, hogy a megfigyelt vektorsorozatot (**O**-t) az m modell $S = S_1, S_2, ..., S_N$ állapotsorozata generálta, akkor az állapotátmeneti valószínűségeket és a $b_j(o_t)$ kibocsátási valószínűségeket összeszorozva kapnánk:

$$P(\mathbf{O}, S|m) = a_{S_1,S_2} b_{S_2}(o_1) \cdot a_{S_2,S_3} b_{S_3}(o_2) \cdots \tag{9.6}$$

Ha a modelljeink (m_i) az egyes szavakra (w_i) készülnek, akkor $P(\mathbf{O}|m_i) = P(\mathbf{O}|w_i)$, a szavakból képzett szó-, illetve modellsorozatokra pedig: $P(\mathbf{O}|M) = P(\mathbf{O}|W)$. Mivel $S = S_1, S_2, ..., S_N$ ismeretlen (sőt éppen ezt keressük, a mintaillesztés eredménye ez lenne), ezért a fenti (9.6) valószínűséget csak úgy tudnánk kiszámítani, ha valamennyi lehetséges állapotsorozatra (azaz az összes elméletileg lehetséges mintaillesztési kombinációra) összegzünk egy adott M modellsorozatra:

$$P(\mathbf{O}|M) = \sum_S \prod_{t=1}^{T} a_{S_t,S_{t+1}} \cdot b_{S_{t+1}}(o_t). \tag{9.7}$$

A fenti (9.7) összefüggésben azonban az egyes lehetséges állapotsorozatokra elvégzett összegzés jól közelíthető maximumkereséssel (minden modellre csak a legvalószínűbb lehetséges állapotsorozat érdekel bennünket), így a legvalószínűbb állapotsorozatra a valószínűség közelítése:

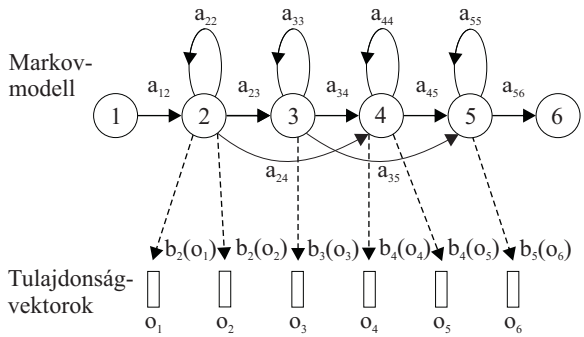
$$\hat{P}(\mathbf{O}|M) = \max_S \prod_{t=1}^{T} a_{S_t,S_{t+1}} \cdot b_{S_{t+1}}(o_t). \tag{9.8}$$

Ezeket az értékeket valamennyi lehetséges M modellsorozatra kiszámítva, közülük a maximális valószínűséget adó modellsorozat a felismerés eredménye:

$$\hat{W} = M \arg\max_M \hat{P}(\mathbf{O}|\mathbf{M}). \tag{9.9}$$

A felismert szósorozathoz tartozó legvalószínűbb állapotsorozat a felismerés eredményének ismeretében visszakövethető. Ennek megkönnyítésére általában egy rögzített 0. állapotból indítják a felismerést, és egy képzeletbeli $N+1$-edik állapotba fogják össze valamennyi lehetséges felismerési útvonal (állapotsorozat) végpontját. A mintaillesztés során a 0. és az $N+1$-edik állapothoz természetesen nem rendelünk megfigyelt tulajdonságvektorokat, így ezeknek az állapotoknak nem lesz a megfigyelések képzelt „kibocsátását" leíró $b_j(o_t)$ sűrűségfüggvénye sem (vö. 9.5. ábra). A fenti összefüggések alapján a rejtett Markov-modelles felismerés akusztikai modelljével készen vagyunk, ha ismerjük valamennyi modell minden i állapotához tartozó $a_{i,j}$ és $b_i(o_t)$ értékeket. Fontos még látnunk, hogy mivel a mintaillesztésben minden lehetséges esetet vizsgálunk, komplex felismerési feladatban (bonyolultabb nyelvi modell esetén) igen nagy a számításigény, ez esetenként akár azt is jelentheti,

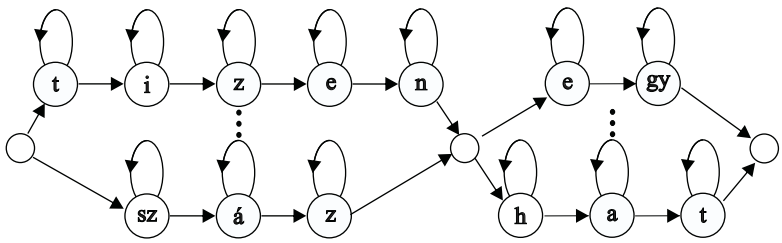
hogy a beszédfelismerő nem lesz képes valós időben, azaz az elhangzó beszéddel párhuzamosan megadni a felismert szósorozatot. Komoly késleltetés származik abból is, hogy szósorozatokra történik az illesztés, emiatt a felismerő a folyamatos beszédhez képest mindig néhány szavas lemaradásban van még valós idejű üzem esetén is, hiszen a mintaillesztés a minta végének a referenciastruktúrához való illesztését is magába foglalja, csak ezt követően derül ki, melyik illesztés adta a végül legjobban illeszkedő állapotsorozatot. Folyamatos beszédfelismerés esetén a megfigyelések szakaszolásáról is gondoskodni kell, egyrész a szósorozat elemszámának kordában tartása, másrészt a túl nagy késleltetés elkerülése érdekében.



9.5. ábra. A Markov-modell állapotai és a tulajdonságvektorok közötti megfeleltetés (mintaillesztés)

Az eddigiekben áttekintettük a rejtett Markov-modelles mintaillesztés matematikai hátterét és az akusztikai modell funkcióját a mintaillesztés során. A következőkben a nyelvi modellt vesszük sorra, jóllehet bújtatva ugyan, de a korábbiak során is feltételeztük, hogy a mintaillesztés során figyelembe veendő lehetséges állapotsorozatok a nyelvi modell alapján állnak össze. A nyelvi modell funkciója éppen ez: azt adja meg, mely szósorozatok milyen valószínűséggel fordulhatnak elő a nyelvben. Beszédhangalapú beszédfelismerés esetén a szótárból meg tudjuk mondani, hogy az egyes szavak milyen beszédhangsorozatból épülnek fel. A beszédhangok elemi akusztikai modellkre, azaz rejtett Markov-modell állapotokra történő közvetlen leképezése alapján ily módon egy gráfot, az úgynevetett *felismerési hálózat*ot konstruálhatjuk meg, amely a mintaillesztés lehetséges útvonalait adja meg – akár beszédhangszintre lebontva (9.6. ábra). Az, hogy egy-egy adott útvonalon melyik a legvalószínűbb állapotsorozat (az adott megfigyelés mellett), már tisztán az akusztikai modellek alapján derül ki. Az egyes útvonalak maguk is súlyozottak, mégpedig értelemszerűen éppen az útvonal által fedett szósorozat valószínűségével, illetve ezzel arányos súllyal. A korábbiakban matematikailag bemutatott mintaillesztési eljárás tehát magában foglalja a lehetséges utak számbavételét (nyelvi modell alapján), az egyes utakhoz tartozó állapotsorozatok közül a legjobb illeszkedést adó meghatározását (akusztikai modellek alapján), majd az egyes utakhoz rendelt optimális állapotsoro-

zatok valószínűségeit az út súlyával kombinálva (ismét a nyelvi modell alapján) az összességében legvalószínűbb út a felismerés eredménye. A rejtett Markov-modelles

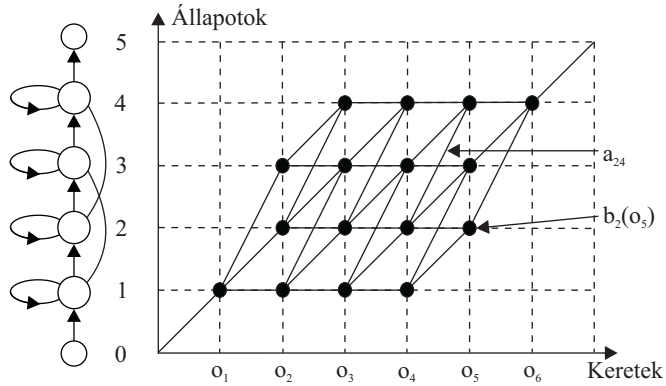


9.6. ábra. Legfeljebb kétjegyű számok felismerésére alkalmas HMM felismerési hálózat

felismerőkben (is) jól alkalmazható, az akusztikai és nyelvi modelleket egységesen kezelő mintaillesztési algoritmust, a *Viterbi-algoritmus*t, a következőkben mutatjuk be. Mivel a Viterbi-algoritmus eredendően a dekódolásban használatos, illetve mivel úgy tekintjük, hogy a megfigyelt tulajdonságvektor-sorozat a rejtett állapotsorozat valamilyen kódolásaként állt elő, ezért a beszédfelismerés mintaillesztési szakaszára gyakran *dekódolás*ként is hivatkoznak (ennek megfelelően a mintaillesztő modult is hívják dekódernek). Definiáljuk először a $\delta_t(i)$ mennyiséget, amely a kezdőtől az i-edik állapotig, az első t megfigyelés során számított legjobb út *jóság*át jelenti. A jóság egy hasonlósági mérték, tulajdonképpen a valószínűség többszörös közelítése, de mivel – mint láttuk – nem a pontos valószínűséget számítjuk, hanem csak közelítjük, ezért ezt a továbbiakban az elnevezésében is hangsúlyozzuk. Legyen a modell állapotainak száma továbbra is N, a tulajdonságvektoroké pedig T. Az egyszerűség kedvéért tekintsük az izolált szavas felismerés esetét! A Viterbi-algoritmus ekkor a következőképpen írható le:

- Inicializálás: $\delta_0(i) = 1$, minden $i = 0 \ldots N+1$-re;
- Indukciós lépés: $\delta_{t+1}(j) = \left(\max_i \left\{\delta_t(i) \cdot a_{i,j}\right\}\right) b_j(o_{t+1})$ minden $i = 0 \ldots N$-re és minden $j = 1 \ldots N+1$-re.
- Leállás: ha $t = T$. Ekkor $\delta_T(N+1)$ adja meg az adott modellre lépésenként maximalizált, kumulált *jóság*ot.

A Viterbi-algoritmus nem más, mint a (9.8) összefüggés rekurzív kiszámítása $P(S = S_0 | t = 0) = 1$ kezdeti feltétel hozzáadásával. Az algoritmust egy TxN-es rácsban való haladással szokásos szemléltetni. A vízszintes tengelyen sorakoznak a tulajdonságvektorok ($\mathbf{O} = o_1, o_2, \ldots o_T$), a függőleges tengely mentén pedig az állapotok ($i = 0, 1, \ldots N+1$) vannak, beleértve a képzeletbeli kezdő- és végállapotokat ($O.$ és $N+1$-edik állapotok) is. (9.7. ábra). A dinamikus idővetemítéshez hasonlóan itt is az a feladat, hogy a bal alsó sarokból eljussunk a jobb felsőbe. Minden egyes ütemben egy oszloppal jobbra lépünk, és az új oszlop minden rácspontjában megvizsgáljuk, hogy hogyan lehet ide jutni, hogy az idáig vezető út a legnagyobb jóságú legyen. Az



9.7. ábra. A Viterbi-algoritmus szemléltetése a lehetséges útvonalak ábrázolásával

algoritmus megvalósításához egy N-elemű táblázatot (oszlopot) kell eltárolni, mely az előző ütem legjobb útjainak a jóságát tárolja állapotonként. Első lépésként a táblázatot egyesekkel töltjük fel, majd az indukciós lépés szerint frissítjük az értékeket. Az indukciós lépésben van az algoritmus lényege: a kiválasztott i állapotban megvizsgáljuk, hogy melyik az az előző ütembeli állapot, amelyre a legjobb út jósága az átmeneti valószínűséggel súlyozva a legnagyobb, vagyis honnan lehet a legnagyobb jósággal ide jutni. Azután ezt a számot ($\max_i \left\{ \delta_t(i) \cdot a_{i,j} \right\}$) beszorozzuk az éppen beérkezett tulajdonságvektor adott állapot szerinti megfigyelési valószínűségével ($b_j(o_{t+1})$), és máris megkaptuk az egy állapottal meghosszabbított út jóságát. Világos, hogy ha olyan állapotban vagyunk, amelyik nem az aktuális tulajdonságvektort modellezi, akkor a megfigyelési valószínűség kicsi lesz, és lerontja az út jóságát. A számítást minden ütemben mindegyik állapotra (az oszlop minden rácspontjára) elvégezzük. Végül, az utolsó lépésben a táblázat legnagyobb értéke ($\delta_T(N)$) adja a teljes tulajdonságvektor-sorozatra számított legjobb út jóságát az adott modellre. Minden modellre elvégezve az algoritmust, a további teendőnk mindössze annyi, hogy kiválasszuk azt a modellt, amelyiknél a legnagyobb ez a számérték – ez azonosítja a felismert szót.

Megjegyzések:

- A gyakorlatban a valószínűségek (és jóságok) helyett azok logaritmusával számolnak, mivel így szorzás helyett összeadásokat kell végrehajtani. A transzformált valószínűségeket költségként is értelmezhetjük, ekkor teljesen hasonló módon a minimális összköltségű útvonalat kell keresni.
- Természetesen egy adott lépésben nem kell minden állapotot végignézni, csak azokat, melyekbe lehetséges az előző lépésből eljutni. (A 0 átmeneti valószínűség miatt amúgy is „kiesnek” az érvénytelen utak.)

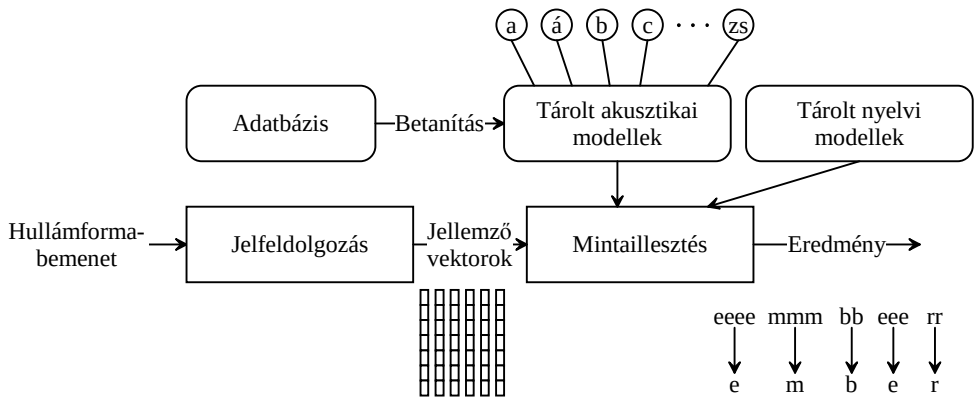
A fenti algoritmus szerint a legjobb út valószínűségének közelítése (jósága) határozható meg adott HMM-nél. Izolált szavas felismerésnél ez elegendő, de ha magára az útvonalra explicite is kíváncsiak vagyunk, akkor ki kell egészíteni az eljárást. Szükségünk van egy TxN-es tömbre (úgynevezett *trace back* tömb), ebbe a kumulált jóságokat tároljuk keretenként a megfelelő oszlopba. Erre azért van szükség, mert így az utolsó, nyerő állapotból a maximumokon visszafelé haladva le tudjuk követni, hogy hogyan is kerültünk oda. Így némi memória és számítási idő ráfordításával a legjobb út valószínűsége mellett az útvonalat is meg tudjuk határozni. Vegyük észre, hogy a legvalószínűbb állapotsorozat meghatározása – az alkalmazott rejtett Markov-modellek topológiájából fakadóan – itt a megfigyelt vektorsorozat időillesztését jelenti. Ilyen értelemben a Markov-modell dinamikus idővetemítést hajt végre, a tárolt referencia azonban nem konkrét sablon, hanem az elemi modellekből a szótárelemeknek (például szavaknak) megfelelően összetett Markov-modelleket képezünk, ezek pedig felismerési hálózattá állnak össze.

Megjegyzések:

- A korábban megismert (9.5) összefüggést szokás a beszédfelismerés MAP (Maximum A Posteriori) alapegyenletének is hívni. $P(W|O)$ értékének keresésekor valójában arról van szó, hogy az $\mathbf{O} = \mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_T}$ megfigyelések által reprezentált paraméterek alapján kellene megadnunk a felismert szót mint egy valószínűségi változó értékét adott paraméterek mellett. Mivel ezt nem tudjuk megtenni, megfordítjuk a gondolatmenetünket: az egyes szavak esetére próbáljuk megbecsülni, mik lehetnek azok a paraméterek, amelyek esetén az adott szó értékét veszi fel a felismerő kimenetét reprezentáló valószínűségi változó (a paraméterek legvalószínűbb értékét keressük rögzített szavak mellett). Ez nem más, mint paraméterbecslés, a maximum likelihood (ML) elv alapján pedig a legjobb becslést úgy kapjuk, ha azokra a paraméterértékekre becslünk, amelyekre az adott szó értékét a lehető legnagyobb valószínűséggel veszi fel a kimenetet reprezentáló valószínűségi változó. A MAP becslés az ML becsléstől annyiban különbözik, hogy az a priori eloszlást ($P(W)$-k) is figyelembe veszi, míg az ML esetén az egyes szavak (vagy szószekvenciák) egyenletes eloszlását feltételeznénk.
- A jóság valójában nem is a valószínűség, hanem a *valószerűség* (likelihood) közelítése (a valószerűséget hasonlóságnak is nevezik). Ezt az elnevezést Fisher (1922) vezette be épp annak érzékeltetésére, amit az előző megjegyzésben megfogalmaztunk: adott paraméterértékekre vagy paraméterérték-tartományokra meg tudjuk adni egy tőlük függő valószínűségi változó valószínűségét, a realizálódott „valószínűségi változó”-hoz azonban a legvalószerűbb paraméterértékeket tudjuk megbecsülni. A magyar szakirodalomban ritkábban tesznek különbséget a valószínűség és a valószerűség (likelihood) között, az angol nyelvű irodalomban azonban ilyen esetekben elterjedten használják a *likelihood* kifejezést.

9.5.3. Beszédhangalapú folyamatos beszédfelismerés

Korábban említettük már, hogy az akusztikai modellek esetében a modellezett beszédegység szinte tetszőleges lehet: szavakra, szóelemekre, szótagokra vagy beszédhangokra is készülhet az akusztikai modell. A beszédfelismerés kutatásának kezdeti időszakában izolált szavas beszédfelismerőket valósítottak meg, így kézenfekvő volt az egyes szavak akusztikai modellezése. Ez gyakorlatilag azt jelentette, hogy minden egyes szónak külön-külön rejtett Markov-modellje volt. Erre az esetre vezettük le a beszédfelismerés 9.5 alapegyenletét is. Nézzük meg, hogy hogyan módosul ez a megközelítés akkor, ha a modellezési alapegységünk a beszédhang. Továbbra is



9.8. ábra. A folyamatos, beszédhangalapú beszédfelismerő felépítése

szósorozatokban gondolkodunk, és lehetséges szósorozatok valószínűségeit kívánjuk összehasonlítani adott akusztikai megfigyelés és adott nyelvi modell mellett. Ehhez szómodellekre van szükség, nem szükséges viszont valamennyi szóról modellt tárolni, elegendő, ha az elemi építőelemek, a beszédhangok modelljei rendelkezésünkre állnak. Az egyes szavak kiejtésének ismeretében ebből könnyedén konstruálható a szó modellje a beszédhangmodellek egymás után fűzésével. Tehát:

$$P(O|W) = \sum_{\Phi} P(O|\Phi) \cdot P(\Phi|W), \tag{9.10}$$

ahol $\Phi = \phi_1, ..., \phi_P$ a W szósorozathoz tartozó lehetséges beszédhangsorozatokat adja meg. Így a (9.5) összefüggés:

$$\hat{W} = \arg\max_{W} P(W) \cdot \sum_{\Phi} P(O|\Phi) \cdot P(\Phi|W). \tag{9.11}$$

Látható, hogy a nyelvi és az elemi akusztikai modell szint közé beékelődött a kiejtési modell ($P(\Phi|W)$), amely egy-egy adott szósorozat esetén egy adott Φ beszédhang-

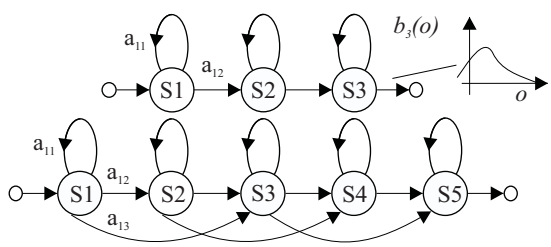
sorozat realizációjának valószínűségét szolgáltatja. A *kiejtési modell* az egyes szósorozatokhoz valószínűségekkel ellátott beszédhangsorozatokat rendel. Ez tipikusan egy kiejtési szótár segítségével történik, melyben az egyes ortografikus szóalakokhoz több, valószínűséggel ellátott beszédhangsorozat is tartozhat.

$$P(\Phi|W) = \max_i \prod_{k=1}^{K} P(\Phi_k^i|w_k), \tag{9.12}$$

ahol Φ_k^i a k-adik szóhoz tartozó i-edik kiejtési variánsnak megfelelő részbeszédhangsorozat. A mintaillesztés az ML (Maximum Likelihood) értelemben optimális szó-beszédhang részsorozat összerendelést, azaz az optimális kiejtési variáns megtalálását is magában foglalja. A kiejtési szótár előállítása történhet kézi, részben vagy teljesen automatikus módszerekkel. A beszédhangszintre transzformált $P(O|\Phi)$ akusztikai modell feladata adott (megengedett) beszédhangsorozathoz valószínűséget rendelni az akusztikai megfigyeléssorozat (**O**) alapján. Formálisan:

$$P(O|\Phi) = \prod_{f=1}^{F} P(O_f|\phi_f), \tag{9.13}$$

ahol O_f az f-edik fonémához tartozó f-edik jellemzővektor-részsorozat, a 9.11 összefüggés szerinti maximalizálás tehát itt az optimális beszédhang–jellemzővektor-részsorozat hozzárendelés megtalálását jelenti. Ha pedig adott egy (hipotetikus) beszédhangmodell– jellemzővektor-részsorozat hozzárendelés, az akusztikus valószínűségek modellenként számolhatók. A gyakorlatban a beszédhangok akusztikai modellezésére is a rejtett Markov-modellek (HMM: Hidden Markov-model) használata terjedt el. Általában úgynevetett balról-jobbra struktúrájú, 3-állapotú HMM-eket használnak, ezekben az átmenet mindig legfeljebb a következő állapotba lehetséges. Egyes rendszerekben az úgynevetett Bakis-struktúra alkalmazása is előfordul (Schramm 2006).



9.9. ábra. A legegyszerűbb balról-jobbra (fent), illetve a Bakis HMM struktúra (lent) beszédhangok akusztikai modellezésére. (A $b_3(o)$ lokális valószínűségsűrűség-függvényt egy dimenziós jellemzővektorral szemléltettük a képen)

Megjegyzések:

- A szavak helyett a beszédhangalapú építőelemek modellezésének számtalan előnye van: a szótár rugalmasan bővíthető, a kiejtésvariációk rugalmasan kezelhetők, áttekinthetőbb a rendszerstruktúra, kisebb a tárhelyigény a modellek tárolásához, kisebb az adatmennyiség-igény a modellek betanításához.
- Elvileg lehetséges az is, hogy a beszédfelismerő kimenetén csupán a felismert beszédhangsorozat jelenjen meg, ebből azonban nagyon nehezen lehetne a közlést rekonstruálni, ugyanis a beszédhangok felismerésének felső határa 70% találati arány környékén van, ráadásul a lehetséges szóhatárokról sincsen információnk. Ha az emberi beszédértést beszédhangszinten vizsgáljuk (például értelmetlen szótagokat, szavakat kell lejegyezni), az ember sem képes a beszédhangok 70–75%-ánál többet helyesen azonosítani (vö. Gósy–Olaszy 1983). Értelmes, felolvasott szövegben a beszédhangok véletlenszerű keverésének egyre növekvő fokával drasztikusan romlik az emberi beszédértés (vö. Tóth et al. 2007). Az, hogy az ember a beszédet ennek ellenére kifogástalanul érti, a nyelv magasabb szintjeinek (értelmes szavak, szintaxis, kontextus stb.) szerepére utal a percepcióban. Mindez azt is jelenti, hogy a gépi beszédfelismerők akusztikai szinten bizonyos körülmények között képesek lehetnek közel ugyanolyan teljesítményt nyújtani, mint az ember. A gépi felismerés mégis nagyságrendekkel gyengébb az emberinél, ez pedig arra utal, hogy a gépi megoldásban a nyelvi tudás reprezentálása, a nyelv modellezése még tökéletlen.

9.5.3.1. Környezetfüggő beszédhangmodellek

Az egyes szavak beszédhangsorozattal történő modellezése azt feltételezi, hogy a szavak diszkrét beszédhangok sorozatából épülnek fel. Ez azonban nem igaz, az egyes beszédhangok között ugyanis koartikulációs hatások lépnek/léphetnek fel (5.2. fejezet). Éppen erre vezethető vissza, hogy miért legalább 3 állapotból álló Markov-modelleket alkalmaznak a beszédhangok akusztikai modellezésére (Mihajlik et al. 2006). A két szélső állapot ugyanis – a pusztán az összefűzést segítő legszélső állapotokat nem számítva (vö. 9.9. ábra) – éppen a koartikuláció kezelésére alkalmas, a beszédhang kvázistacionárius szakaszát a középső állapot modellezi (a betanítás ismertetésénél látni fogjuk, hogy ez hogyan biztosítható). A koartikuláció akusztikai megjelenése a két szomszédos beszédhang tulajdonságaitól függ. Ezek szerint tehát az egyes beszédhangmodellek szélső állapotai együttesen modellezik a különböző beszédhangok kontextusába került központi, modellezendő beszédhang koartikulációs viselkedését. Ha a modelleket kontextusérzékennyé tesszük, azaz a különféle hangkapcsolatokra külön-külön modelleket készítünk, a modellezés pontosabbá tehető. Ennek megfelelően egy beszédhangnak elvileg annyi modellje lesz, ahányféle kontextusba (beszédhang-környezetbe) kerülhet az adott beszédhang. Ezek közül mindig

azt a modellt fogjuk használni, amelyre a kontextus is stimmel. Az ilyen modelleket nevezik *kontextusérzékeny* vagy *környezetfüggő* modelleknek, illetve gyakran *trifón* modelleknek is. Ez utóbbi elnevezés a legelterjedtebb az angol nyelvű szakirodalomban, jóllehet kissé megtévesztő, ugyanis nem három beszédhang együttes modellezése történik, hanem egyetlen beszédhangé, amelynek „széleit" a koartikulációs hatás a szomszédos beszédhangok függvényében megváltoztat(hat)ja. Mint azt a modellek betanítása (készítése) kapcsán rövidesen látni fogjuk, a statisztikai megközelítés miatt nem mindegy, mennyi minta alapján becsüljük meg a modellek paramétereit. Ha a nyelvben 30 beszédhang van, akkor egyetlen beszédhangra $30^2 = 900$ modell készülhetne kontextusérzékeny modellezéskor, ez durván ugyanilyen mértékben növelné a betanító adatbázis megkövetelt méretét is a nem kontextusérzékeny esethez képest. Mivel az adatbázisok drágák, illetve mivel a becslés annál pontosabb, minél nagyobb a rendelkezésre álló mintaszám, ezért a gyakorlatban köztes megoldást alkalmaznak. A koartikuláció ugyanis a szomszédos hang akusztikai sajátságaitól függ, ezeket pedig döntően a képzési hely és mód befolyásolja (Olaszy 2003). Az egyes beszédhangok tehát képzési jegyeik szerint csoportosíthatók (klaszterezhetők), és az egy csoportba tartozó beszédhangokat a kontextusban nem különböztetjük meg (az ezeknek megfelelő szélső állapotokat a modellek közösen használják, megosztják, innen ered az *osztott állapot* elnevezés). Ily módon jól kihasználható a betanító adatbázis a paraméterbecsléshez (relatíve nagy mintaszámmal tudunk becsülni). A csoportosítást gyakran automatizálják, jellemzően döntési fákkal, amelyek az optimális klaszterezést is képesek megtalálni bizonyos rögzített feltételek mellett.

Megjegyzés:

- A beszédadatbázisok készítésénél éppen a kontextusérzékeny modellezés támogatása érdekében fontos, hogy a beszédadatbázis az adott nyelvben előforduló hármas hangkapcsolatokat reprezentatív számban tartalmazza, kellő számú beszélő bemondásában (az akusztikai változatosság miatt).

9.5.3.2. A kényszerített illesztés

A kényszerített illesztés (forced alignment) a mintaillesztés egy speciális esete, amelyet gyakran alkalmaznak a beszédtechnológiában, elsősorban beszédhangszintű szegmentálásra (a beszédhangok „határainak" hozzávetőleges bejelölésére). Ez hasznos lehet beszédadatbázisok feldolgozásában (8.2.3. fejezet), illetve az akusztikai modellek betanításakor is.

A kényszerített illesztés olyan mintaillesztési feladat, amelyben ismert tartalmú hanganyaghoz szeretnénk annak szöveges (fonetikai) átiratát illeszteni. Rendelkezésre áll a beszédminta és annak fonemikus szöveges változata is. Tehát tulajdonképpen rögzített felismerési cél (tudjuk, hogy mi a közlésben a beszédhangok sorrendje) mellett a legjobb illeszkedést adó útvonalat keressük. Az útvonal visszakövetésé-

vel (trace back) egészen az elemi modellekig lebontva meghatározható, mely időpontokban (az aktuális keretknél) mely modellben (igény esetén mely elemi modell melyik állapotában) tartózkodtunk a legnagyobb valószínűséggel a rejtett Markov-modellekből konstruált felismerési hálózatban. A felismerési hálózatot ebben az esetben nem a nyelvtan, hanem az ismert szöveges átirat alapján generáljuk le. A hálózat generálásakor lehetőség nyílik arra is, hogy abban kiejtésváltozatokat is szerepeltessünk, a mintaillesztés pedig ezek közül is kiválasztja a legvalószínűbbet. A kiejtésváltozatoknak megfelelően a felismerési hálózat elágaztatható, így az egyes változatok elméletileg akár beszédhang szinten is megkülönböztethetők lehetnek, a gyakorlatban leginkább szavak egyes ejtésváltozatai közül választhatjuk ki az aktuálisan megjelenőt (például *egyszer* vagy *eccer*).

Megjegyezzük, hogy az akusztikai modellek tanításakor is gyakran alkalmazott lépés a kényszerített illesztés még akkor is, ha előzetesen felszegmentált adatbázison történik a modellparaméterek becslése. Például, a beszédhangszintű szegmentálás alapján tanított modelleket felhasználva a tanítóanyagot újraszegmentálhatjuk kényszerített illesztéssel, majd az így nyert, a gép „ízlésének", tudásának jobban megfelelő szegmentálással pontosabb modellek készíthetőek.

A kényszerített illesztés hatékonysága jó, a hangok azonosítása és a hanghatárok helyes megjelölése 10 ms-os küszöbérték mellett elérheti a 80–85%-ot is.

9.5.3.3. Mintaillesztési példa a Viterbi-algoritmus használatára

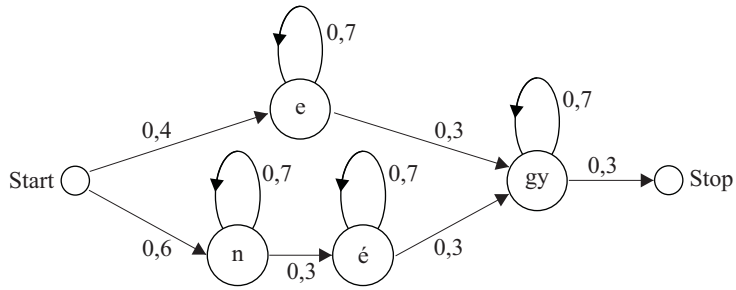
Az alábbi 9.10. ábrán feltüntetett beszédfelismerési hálózaton beszédfelismerési kísérletet végzünk. A beszédhangok modelljei rejtett Markov-modellek, a mintaillesztést Viterbi-algoritmussal végezzük (az algoritmust lásd a 9.5.2. fejezetben). Mi a felismerés eredménye, ha az előfeldolgozó egységből három tulajdonságvektor érkezett (rendre o_1,o_2 és o_3), amelyekre:

$b_{'e'}(o_1) = 0,8 \quad b_{'e'}(o_2) = 0,4 \quad b_{'e'}(o_3) = 0,2$
$b_{'n'}(o_1) = 0,6 \quad b_{'n'}(o_2) = 0,4 \quad b_{'n'}(o_3) = 0,2$
$b_{'é'}(o_1) = 0,3 \quad b_{'é'}(o_2) = 0,6 \quad b_{'é'}(o_3) = 0,4$
$b_{'gy'}(o_1) = 0,1 \; b_{'gy'}(o_2) = 0,5 \; b_{'gy'}(o_3) = 0,8$

A feladat megoldása: a felismerési hálózat topológiájára való tekintettel a 'START' feliratú, nem kibocsátó állapotból indulunk, tehát: $\delta_0(START) = 1$ Az első tulajdonságvektor érkezésekor az 'e' vagy az 'n' állapotokba kerülhetünk. Az első indukciós lépést elvégezve:

$\delta_1('e') = \delta_0(START) \cdot a_{START,e} \cdot b_{'e'}(o_1) = 1 \cdot 0,4 \cdot 0,8 = 0,32$ és
$\delta_1('n') = \delta_0(START) \cdot a_{START,n} \cdot b_{'n'}(o_1) = 1 \cdot 0,6 \cdot 0,6 = 0,36.$

A második tulajdonságvektor érkezésekor az egyes állapotokból továbbléphetünk, vagy akár helyben is maradhatunk. A második indukciós lépés előtt gondolkodjunk kicsit! Tudjuk, hogy három tulajdonságvektor érkezik összesen, sikeres mintaillesz-



9.10. ábra. Felismerési hálózat a mintaillesztési feladathoz.

tés pedig csak úgy történhet, ha az utolsó (a harmadik) tulajdonságvektor érkezésével a végállapotba (END) jutunk. Vegyük észre, hogy ha az 'n' állapot irányába indulunk, akkor egyetlen esetben sem maradhatunk helyben (továbblépési kényszerben vagyunk), hiszen épp annyi állapot van, amennyi tulajdonságvektor érkezik. Emiatt eleve kihagyhatjuk azokat a kombinációkat (állapotsorozatokat), amelyek esetén a mintaillesztés végül eredménytelen lenne. Így a második indukciós lépésben nem számítjuk $\delta_2('n')$-t.

$\delta_2('e') = \delta_1('e') \cdot a_{e,e} \cdot b_{'e'}(o_2) = 0,32 \cdot 0,7 \cdot 0,4 = 0,0896$

$\delta_2('é') = \delta_1('n') \cdot a_{n,é} \cdot b_{'é'}(o_2) = 0,36 \cdot 0,3 \cdot 0,6 = 0,054$

$\delta_2('gy') = \delta_1('e') \cdot a_{e,gy} \cdot b_{'gy'}(o_2) = 0,32 \cdot 0,3 \cdot 0,5 = 0,048$

Az első két indukciós lépésben elhagyhattuk a maximalizálást, miután a maximumfüggvény argumentumában csak egyetlen lehetőség szerepelt volna (egyetlen olyan állapot volt, amelyből az éppen vizsgált állapotba juthattunk). A harmadik indukciós lépésben ez már nem teljesül, ugyanis a harmadik tulajdonságvektor érkezésekor a topológiából adódóan mindenképpen a 'gy' állapotba kell lépnünk a sikeres mintaillesztéshez, tehát csak $\delta_3('gy')$-t számítjuk, viszont a maximalizálást már nem „ússzuk meg", hiszen a 'gy' állapotba 3 helyről is léphetünk az 'e' és 'é' állapotokból, illetve magából a 'gy' állapotból (helyben maradással):

$\delta_3('gy') = \max_{'e','é','gy'} \{\delta_2('e') \cdot a_{e,gy} \cdot b_{'gy'}(o_3),$

$\delta_2('é') \cdot a_{é,gy} \cdot b_{'gy'}(o_3), \delta_2('gy') \cdot a_{gy,gy} \cdot b_{'gy'}(o_3)\} =$

$= \max\{0,0896 \cdot 0,3 \cdot 0,8; 0,054 \cdot 0,3 \cdot 0,8; 0,048 \cdot 0,7 \cdot 0,8\} =$

$= \max\{0,021504; 0,01296; 0,02688\} = 0,02688.$

Esetünkben az utolsó lépés már csak formalitás:

$\delta_{3+}('END') = \delta_3('gy') \cdot a_{'gy',END} = 0,02688 \cdot 1 = 0,02688.$

A felismert szót visszakövetéssel kaphatjuk meg. A 'gy' állapotba legvalószínűbben helybenmaradással a 'gy állapotból jutottunk, ide pedig kényszerűen az 'e' állapotból, a felismert szó tehát az *egy*, a mintaillesztés során kapott legvalószínűbb állapotsorozat pedig: 'e', 'gy', 'gy'. Vegyük észre, hogy az egyes indukciós lépések végén kapott legvalószínűbb állapot részsorozat nem feltétlenül fedi a mintaillesztés legvégén kapott legvalószínűbb állapotsorozatot, hiszen például a második induk-

ciós lépésben úgy nézett ki, hogy az 'e' állapotban történő helyben maradás volt a valószínűbb, ez azonban a harmadik lépésben kapott legvalószínűbb útba már nem illeszkedik.

9.6. A beszéd-szöveg átalakítás alapvető tudásforrásai

Statisztikainak azért nevezik a korszerű beszédfelismerési megközelítéseket, mert az egyes (nyelvi, akusztikai stb.) modellek paramétereinek becslése tanító- vagy mintaadatok (szöveg, hullámforma) alapján statisztikai úton történik. Minél több és változatosabb a tanítóadat, annál pontosabb és jobb általánosító képességű modelleket készíthetünk. Például akusztikai modelleknél néhányszor száz beszélős tanítóadatbázisoknál már beszélőfüggetlennek tekintjük a betanított modelleket. Miután az akusztikai és nyelvi modellek lényegében tudást tárolnak, ezért ezeket *tudásforrás*oknak vagy tudásbázisoknak is nevezik. A következőkben az akusztikai és a nyelvi modellek készítésekor követett lépéseket vesszük sorra, amelyek lényegében azt a célt szolgálják, hogy a beszéd akusztikai jellegzetességeiről és a nyelvről (nyelvtanról, pontosabban főleg a szintaxisról) megszerezhető tudást kinyerjük, absztraktan megfogalmazzuk és eltároljuk.

9.6.1. Az akusztikai modellek betanítása

A fonéma akusztikai modellek paraméterbecslése alatt az elemi HMM-ek a és b paramétereinek elemi modellenkénti vagy együttes meghatározását, beállítását értjük. A fő feladat a $b_j(o_t)$ eloszlások becslése, hiszen a mintaillesztés alapvető része, az osztályozás ezeken keresztül valósul meg. Mivel a becslés iteratívan történik, ezt szokás *betanítás*nak is nevezni, azok a minták, amelyek alapján a becslést végezzük a *tanítómintá*k. Mint később látni fogjuk, minden becslési ciklus után ellenőrizzük, hogy a modellek „megtanulták-e" már a mintákból kinyerhető tudást, azaz megfelelő illeszkedést adnak-e a tanítóminták halmazára. Ha igen, a tanulást befejeztük. A HMM állapotokhoz tartozó folytonos valószínűség-eloszlásokat – egészen pontosan azok sűrűségfüggvényeit – a legelterjedtebben Gauss-függvények szuperpozíciójaként modellezik (Titterington et al. 1985). Ezt a megközelítést a GMM (Gaussian Mixture Model) jelzővel szokták illetni. A tulajdonságvektorok egy adott beszédelemre (illetve a rá referáló állapotra) tehát

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} \cdot G(o_t; \mu_{jm}, \Sigma_{jm}) \tag{9.14}$$

alakúak, ahol c_{jm} az m-edik Gauss-komponens súlya ($\sum_{m=1}^{M} c_{jm} = 1$), $G(.)$ pedig a normális eloszlás (Gaussian) sűrűségfüggvénye, amelynek két paramétere egyértelműen azonosítja a sűrűségfüggvényt: μ_{jm} adja meg a várhatóérték-vektort (mely annyi dimenziós, ahány dimenziósak a tulajdonságvektorok), Σ_{jm} pedig a kovarianciamátrix, amelyeket az o_t tulajdonságvektorok alapján szeretnénk becsülni. Ha a tulajdonságvektor N dimenziós, akkor a normális eloszlás alakja:

$$G(o_t;\mu_j,\Sigma_j) \;=\; \frac{1}{\sqrt{(2\Pi)^N|\Sigma_j|}} \cdot e^{-\frac{1}{2}(o_t-\mu_j)^T\Sigma_j^{-1}(o_t-\mu_j)}. \tag{9.15}$$

A feladat tehát $G(.)$ paramétereinek becslése. A matematikai statisztika alapján a normális eloszlás várható értékének ML becslése a mintaátlag:

$$\mu_j \;=\; \frac{1}{T}\sum_{t=1}^{T} o_t \tag{9.16}$$

a kovarianciamátrixot pedig a korrigált tapasztalati szórásnégyzet alapján tudjuk meghatározni:

$$\Sigma_j \;=\; \frac{1}{T-1}\sum_{t=1}^{T}(\mu_j - o_t)(\mu_j - o_t)^T. \tag{9.17}$$

Mivel az o_t tulajdonságvektorok többé-kevésbé dekorreláltnak tekinthetők, tipikusan diagonális kovarianciamátrixszal szoktak számolni. Az akusztikai modellparaméterek ML becslése történhet a 'k-means' algoritmussal (MacQueen 1967) is, amit az egyes beszédhangmodellek a és b paramétereinek együttes becslésére a Viterbi-újraszegmentálással egészítenek ki. A másik gyakran alkalmazott tanítási módszer a Baum–Welch (Baum–Eagon 1967) újrabecslő algoritmus futtatása először uniform modellparaméterek esetén, majd a Mixture-splitting eljárással növelt komplexitású modelleken való iteratív alkalmazása (Young 2006). Míg az előbbi esetben szükséges a tanító-adatbázisban a modellezett beszédegységek (hangok, esetleg szavak) határainak kijelölése, a második esetben nem, csupán a modellszekvencia ismerete fontos. Ezek a módszerek alapvetően az EM (Expectation Maximization) algoritmus (Dempster et al. 1977) egyes változatai. A betanítás során is problémaként jelentkezik a modellek „rejtettsége", azaz nem tudjuk biztosan megmondani, hogy mely o_t tulajdonságvektorokat kellene felhasználnunk egy-egy állapot paramétereinek becslésére. Például ha egy szó rejtett Markov-modelljének paramétereit szeretnénk meghatározni, több mintára is szükségünk van ebből a szóból, és minden egyes minta tulajdonságvektorait a megfelelő állapothoz kellene hozzárendelnünk. Ha már készen lenne a felismerőnk, azaz a szó Markov-modellje, minden további nélkül elvégezhetnénk a mintaillesztést, majd visszakövetéssel meg tudnánk határozni az optimális állapot-tulajdonságvektor hozzárendelést. E hozzárendelés megteremtésében segít a beszédhangszintű szegmentálás, ekkor a szegmentálás alapján be tudjuk azonosítani

az **O** tulajdonságvektor-sorozatnak azokat a diszjunkt részsorozatait, amelyeket egy-egy állapot a és b paramétereinek becslésére felhasználhatunk. Az így kapott modell azonban még tovább finomítható: a modellel megpróbálhatjuk felismerni azokat a mintákat, amelyeket a modell paramétereinek becsléséhez felhasználtunk. A felismerés eredménye nem kérdés ugyan, de visszakövetéssel megtalálhatjuk a mintaillesztéssel kapott legjobb állapotsorozatot az egyes mintákra. Ez az állapotsorozat nem feltétlenül illeszkedik pontosan a szegmentálás által megadott beszédhanghatárokkal, sőt, attól jellemzően el is tér. Mi is történt tulajdonképpen: készítettünk egy kezdeti modellt, majd mintaillesztést hajtottunk végre, de a mintaillesztés eredménye nem egyezik a modell készítésénél feltételezett állapot-tulajdonságvektor hozzárendeléssel. Egyes tulajdonságvektorok másik (leginkább szomszédos) állapothoz sorolódnak. Ez azt jelenti, hogy a modell még pontosítható. Használjuk a modell újabb elkészítéséhez az illesztett tulajdonságvektor-sorozatot, de a modell paramétereit számítsuk újra. Majd végezzünk ismételt mintaillesztést, illetve iteratívan ismételjük a paraméterbecslés-mintaillesztés ciklusokat mindaddig, amíg a modell paraméterei számottevően változnak. Ez maga a gépi tanulás, az ily módon iteratívan optimalizálódó paraméterbecslést nevezik betanításnak. Szemléletesen az algoritmus valóban tanulás: az első, majd a soron következő paraméterbecsléssel szerzett tudást próbára tesszük a már ismert mintákon, s ha az egyes iterációkban már nem tudunk meg többet, azaz a paramétereink értékei nem változnak, akkor az adott szinten kinyerhető tudást megszereztük. Mindez azt is jelenti, hogy akár pontos szegmentálás nélkül is hozzáfoghatunk a tanuláshoz, hiszen annyit is elég lehet biztosítani, hogy a megfelelő állapotokhoz hozzávetőlegesen az odaillő tulajdonságvektorokat párosítsuk. Ilyen az úgynevezett *flat start* eljárás, amikor is gyakorlatilag hasraütésszerűen, például a minta tulajdonságvektorait a modell állapotainak száma szerint, szekvenciálisan egyenlő részekre bontva kezdjük meg az egyes állapotok paramétereinek becslését. A fix állapot-tulajdonságvektor hozzárendelés helyett azonban még rugalmasabb struktúrában is gondolkozhatunk: ésszerű korlátozásokat betartva megengedjük, hogy az egyes tulajdonságvektorok gyakorlatilag tetszőleges állapothoz legyenek párosíthatóak, bizonyos valószínűséggel (súllyal). A továbbiakban röviden áttekintjük a b paraméterek egy elterjedten alkalmazott, a *Baum–Welch-algoritmus*sal történő becslését. Az eljárások további részletezését viszont mellőzzük, azok kimerítő leírása megtalálható például a Rabiner–Juang (1993) és a Young (2006) irodalmakban. A Baum–Welch-algoritmus lényege tehát, hogy egy adott állapot $b_j(o_t)$ sűrűségfüggvényének becslését nem fix állapot-tulajdonságvektor párosítások mellett végezzük, hanem minden o_t megfigyelt tulajdonságvektort minden olyan j állapothoz tartozó $b_j(o_t)$ sűrűségfüggvény becsléséhez felhasználunk, amelyre pozitív az adott t időben az adott j állapotban tartózkodás valószínűsége. Ezt a valószínűséget $L_j(t)$-vel jelölve a 9.16 és a 9.17 öszefüggések a következőképpen módosulnak:

$$\mu_j = \frac{\sum_{t=1}^{T} L_j(t) \cdot o_t}{\sum_{t=1}^{T} L_j(t)}, \tag{9.18}$$

$$\Sigma_j = \frac{\sum_{t=1}^{T} L_j(t) \cdot (\mu_j - o_t)(\mu_j - o_t)^T}{\sum_{t=1}^{T} L_j(t)}. \tag{9.19}$$

Ha tehát meg tudjuk határozni $L_j(t)$-t, akkor el tudjuk végezni a Baum–Welch-paraméterbecslést. $L_j(t)$ meghatározása egyszerű megfontolásokkal lehetséges az úgynevezett *előre-hátra algoritmus*sal (forward-backward algorithm). Definiáljuk az előre valószínűséget a következőképpen:

$$\alpha_j(t) = P(o_1, ..., o_t, S(t) = j | M), \tag{9.20}$$

azaz $\alpha_j(t)$ jelenti annak valószínűségét, hogy t időpontban a j állapotban vagyunk, és az eddig eltelt idő alatt $o_1, ..., o_t$ tulajdonságvektorokat figyeltük meg. Ezt a valószínűséget egyszer már közelítettük a dekódolásra használt Viterbi-algoritmus ismertetése során. Most annyi a különbség, hogy a valószínűség közelítése (jóság) helyett pontos értékekkel kell számolnunk, azaz a szummázás helyett nem végezhetünk maximumkeresést. Ennek megfelelően $\alpha_j(t)$ rekurzív számítása:

$$\alpha_0(1) = 1 \tag{9.21}$$

$$\alpha_j(1) = a_{0,j} b_j(o_1) \tag{9.22}$$

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \left\{ \alpha_t(i) \cdot a_{i,j} \right\} \right) b_j(o_{t+1}) \tag{9.23}$$

a képzeletbeli $N+1$-edik végállapotra összefogva pedig:

$$\alpha_{N+1}(T) = \sum_{i=1}^{N} a_{i,N+1} \cdot \alpha_i(T). \tag{9.24}$$

A hátra valószínűség definícóját az alábbi formában adhatjuk meg:

$$\beta_j = P(o_{t+1}, ..., o_T | S(t) = j, M), \tag{9.25}$$

azaz $\beta_j(t)$ jelenti annak valószínűségét, hogy t időpontban a j állapotban vagyunk, és a következőkben pedig az $o_{t+1}, ..., o_T$ tulajdonságvektorokat fogjuk megfigyelni. $\beta_j(t)$ is rekurzívan számítható, de most visszafelé elindulva:

$$\beta_i(T) = a_{i,N+1} \tag{9.26}$$

$$\beta_{t-1}(j) = \sum_{j=1}^{N} \beta_t(i) \cdot a_{i,j} b_j(o_t) \tag{9.27}$$

Az előre és a hátra valószínűségeket azért így határoztuk meg, mert ekkor:

$$\alpha_j(t)\cdot\beta_j(t) = P(o_1,...,o_t,S(t)=j|M)\cdot P(o_{t+1},...,o_T|S(t)=j,M) \quad (9.28)$$
$$= P(o_1,...,o_t,...,o_T,S(t)=j|M) \quad (9.29)$$
$$= P(\mathbf{O},S(t)=j|M). \quad (9.30)$$

Ennek segítségével meg tudjuk adni az állapotban tartózkodás valószínűségét:

$$L_j(t) = P(S(t)=j|o_1,...,o_T,M) \quad (9.31)$$
$$= P(S(t)=j|\mathbf{O},M) \quad (9.32)$$
$$= \frac{P(\mathbf{O},S(t)=j|M)}{P(\mathbf{O}|M)} \quad (9.33)$$
$$= \frac{1}{\alpha_{N+1}(T)}\cdot\alpha_j(t)\beta_j(t), \quad (9.34)$$

hiszen $P(\mathbf{O}|M)=\alpha_{N+1}(T)$. Miután az állapotban történő tartózkodás valószínűségét az idő függvényében meghatároztuk, a Baum–Welch-paraméterbecslést is el tudjuk végezni. Kísérletek támasztják alá, hogy a Baum–Welch-algoritmussal végzett paraméterbecslés sokkal pontosabb beszédfelismerést tesz lehetővé. Végezetül összegezzük az akusztikai modellek betanítása kapcsán elmondottakat: az akusztikai modellek a és b paraméterei tárolják az akusztikai tudást, az ebből származtatott referenciához történik a felismerés során a mintaillesztés. Az akusztikai modellek paramétereit iteratív ciklusokban lehet kiszámítani: a tulajdonságvektorokat fixen, vagy állapotban tartózkodás valószínűségével súlyozva rugalmasan a megfelelő állapotokhoz társítjuk, majd becsüljük az a és b paramétereket (ezek közül a b paraméterek becslését mutattuk be részletesen). Az így nyert modellel mintaillesztést végzünk a tanítóanyagon, majd újrabecsüljük a modell paramétereit az új mintaillesztés alapján. Mindezt addig folytatjuk, amíg a modellparaméterek szignifikánsan változnak. A gyakorlatban a betanítás egyébként jellemzően három fázisra oszlik: inicializálás, ciklikus újrabecslés fix hozzárendeléssel, majd ciklikus újrabecslés Baum–Welch-algoritmussal. A $b_j(o_t)$ paraméterek tekintetében a becsült sűrűségfüggvény az első két fázisban egyszerű normális eloszlás, a harmadikban GMM, fokozatosan növelt komponensszámmal (általában kétszerezés történik, majd 2–3 Baum–Welch-ciklus a következő kétszerezés előtt). Nyolc-tizenhat komponensű GMM általában már kellően pontos modellezést tesz lehetővé.

Fontos megjegyezni, hogy az ML akusztikai modellbecslést a korszerű rendszerekben gyakorta diszkriminatív tanítás követi. Ennek lényege, hogy nem a modellek tanító-adatbázishoz illeszkedésének mértékét igyekszik maximalizálni, hanem közvetlenebb módon, például MMI (Maximum Mutual Information), MCE (Minimum Classification Error) vagy MPE (Minimum Phone Error) kritériumoknak megfelelően minimalizálja a felismerési hibát (Bahl et al. 1986, McDermott 1997, Povey–

Woodland 2002). Szintén idekívánkozó megjegyzés, hogy mivel az akusztikai modell az egyes ciklusokban egyre jobban illeszkedik a betanítóanyagra, ezért biztosítani kell, hogy a betanítóanyag kellően mintagazdag legyen (mind a mintaszám, mind az akusztikai változatosság megfelelő legyen). Ezért is különösen fontos a betanításhoz használatos adatbázisok szakszerű elkészítése (lásd a 8.1. fejezetben). Ha ez a követelmény nem teljesül, torz modellek születnek, amelyek a tanítóanyagra jól illeszkednek ugyan, de felismeréskor – a tanítóanyagban nem szereplő akusztikai változatok megjelenésével – gyakorlatilag használhatatlanok.

9.6.2. A nyelvi modell készítése

A folyamatos gépi beszédfelismerők nélkülözhetetlen komponense a nyelvi modell, amelynek feladata a $P(W)$, vagyis adott W szósorozat önmagában vett előfordulási valószínűségének becslése (megadása). Ez történhet statisztikai közelítő módszerek alkalmazásával, vagy nyelvi szabályok definiálásával. Utóbbi esetben formalizáltan adjuk meg a nyelv szintaxisát, míg az előbbiben a szintaxis szabályait közvetlenül nem adjuk meg, hanem statisztikai eszközök segítségével, közvetetten írjuk le. A formalizált megoldás előnye, hogy nem szükséges hozzá szöveges tanító-adatbázis, de komplexebb feladatokban a kordában tartható komplexitású szabályrendszer általában rugalmatlan, nem kellő általánosító képességű. A statisztikai megközelítés fontos tulajdonsága, hogy nyelvi szakértelem nélkül állítható vele elő az adott beszédfelismerési feladathoz tartozó rugalmas nyelvi modell. Nagyszótáras folyamatos beszédfelismerés csak statisztikai alapokon nyugvó nyelvi modellel valósítható meg. A statisztikai nyelvi modell viszont csak akkor hatékony, ha témaspecifikus és nagyméretű szöveges adatbázissal tanítják. Ha megfelelő adatbázis nem áll rendelkezésre, akkor általában jelentős idő- és költségráfordítás szükséges az előállításához.

9.6.2.1. Statisztikai N-gram modellek

Az N-gram modellek statisztikailag modellezik a nyelvet, lényegében az egyes szavak egymás után következésének becsült valószínűségét adják meg. A keresett valószínűség tehát $P(W) = P(w_1, w_2, ..., w_M)$, azaz annak valószínűsége, hogy éppen a $w_1, w_2, ..., w_M$ szavak követik egymást. A láncszabály szerint ezt a valószínűséget az alábbiak szerint írhatjuk:

$$P(w_1, ..., w_M) = \prod_{i=1}^{M} P(w_i | w_1, ..., w_{i-1}), \tag{9.35}$$

azaz lényegében szavak, szókettesek, szóhármasok stb. szó M-esek valószínűségeit kellene ismernünk minden lehetséges esetre. Mivel ez gyakorlatilag lehetetlen, ezért közelítést alkalmazunk, és azt feltételezzük, hogy kellően hosszú szósorozatban a következő szó lényegében független a sokkal korábban elhangzott szavaktól, azaz a nyelv *ergodikus*. A legelterjedtebbek az úgynevezett *N-gram* modellek. Az N-gram modellek közelítő becslést adnak egy tetszőlegesen hosszú szósorozat valószínűségére. A közelítés lényege, hogy az együttes valószínűség kiszámításához használt láncszabályt módosítva alkalmazzuk, a feltételes valószínűségeknél a feltételben a memóriahossz (history) maximum N-1 lehet ($N \geq 1$), az ennél hosszabb tagokat már nem vesszük figyelembe:

$$\hat{P}(w_1,...,w_M) = \prod_{i=1}^{M} P(w_i|w_{i-N+1},...,w_{i-1}), \tag{9.36}$$

A gyakorlatban főként az N = 2 (bigram) és N = 3 (trigram) nyelvi modelleket használják, melyek – főként az előbbi esetén – könnyen integrálhatók a beszédfelismerő rendszerekbe. A memóriahossz ilyen rövidre való választása kényszerű döntés (ilyen rövid szósorozatokra az ergodicitás is viszonylag durvább közelítés), amely két okra vezethető vissza: a túl nagy memóriahossz megnövelheti a mintaillesztés számításigényét (hosszabb felismerési hálózat) és növeli a késleltetést, valamint – és ez a kritikus – jelentősen növeli a tárhelyigényt is (az N-gramokat valahol tárolni is kell, ráadásul felismeréskor praktikusan a memóriában, hogy ne vesszen el értékes felismerési idő a háttértárból való beolvasással).

Megjegyzés: a memóriaigény felső korlátja N szóból álló bi-gram esetén N^2-tel, trigram esetén N^3-nal arányos, ha bináris kóddal sikerül is átlagosan egy byte-on tárolni egy szóhármas (trigram) valószínűségét, egy mindössze 1000 szavas felismerő elméletileg 1 GB nagyságrendjében is igényelhet memóriát csak a nyelvi modell tárolásához.

Az N-gram nyelvi modellhez a valószínűségeket szintén statisztikai úton, becsléssel nyerhetjük. A valószínűségek becslésének alapja a relatív gyakoriság, amely a megfelelő feltételek teljesülése esetén konvergál a valószínűséghez:

$$\hat{P}(w_1,...,w_M) = \frac{C(w_{i-N+1},...,w_{i-1},w_i)}{C(w_{i-N+1},...,w_{i-1})} \tag{9.37}$$

ahol $C(.)$ a számláló operátor és $M = i - N + 1$. Mindez azt is feltételezi, hogy egy olyan korpusz birtokában vagyunk, amely a nyelv egy, a felismerő által ismert részhalmazát lefedi, és alkalmas arra, hogy rajta a fenti (9.37) összefüggéssel a kívánt valószínűségeket a kellő pontossággal megbecsülhessük. Az ennek eleget tevő korpuszok jellemzően igen nagy méretű szöveges adatbázisok (a kellő szintű konvergencia ugyanis magas mintaszámot feltételez). A nyelvi modell készítéséhez az ideális olyan beszéd átiratát tartalmazó korpusz lenne, amely pontosan tükrözi a beszédfel-

ismerő majdani felhasználási területét, s abban minden akusztikai – tartalmilag nyelvi jelentéssel akár nem is bíró (például hezitáció, lélegzetvételi zörejek) – esemény jelölve van. Mivel ilyen korpusz ritkán áll rendelkezésre (az ilyen szintű átirat és annotáció igen lassú és költséges), ezért általában szöveges adatbázisokat alkalmaznak (8.1.1.2. fejezet). Bármilyen nagy is azonban a szöveges korpusz, általában sohasem elég nagy, nem ad kellő fedést, ritka, de a nyelvben elméletileg létező szókapcsolatok hiányozhatnak belőle. A korpusz alapján ezek becsült valószínűsége nulla, így felismerhetetlenek. Ennek orvoslására alkalmazzák a különböző úgynevezett *simítás*i technikákat. Alapvetően kétféle simítási megközelítés használatos: a *visszametszés* (backoff), ilyenkor a rosszul becsülhető N-gram valószínűségeket visszavezetjük a jobban becsülhető (N-1)-gram valószínűségekre (Katz 1987):

$$\hat{P}(w_k|w_{k-N+1},...,w_{k-1}) \;=\; P_{boff,N}\cdot P(w_k|w_{k-N+2},...,w_{k-1}), \tag{9.38}$$

ahol $P_{boff,N}$ valószínűséget úgy kell megválasztani, hogy a nyelvi modellek által szolgáltatott összvalószínűség 1 maradjon. Az eljárás az 1- vagy akár 0-gram szintig folytatható. A másik lehetőség az *interpoláció*: a megközelítés lényege, hogy amikor N-gram szinten egy valószínűség rosszul becsülhető – szemben az előző megoldással – nem zárja ki az adott N-gram szintet a valószínűségi becslésből, hanem interpolálja (mintegy korrigálja) az alacsonyabb rendű értékkel (Jelinek–Mercer 1980):

$$\hat{P}(w_k|w_{k-N+1},.,w_{k-1}) \;=\; (1-\alpha)\cdot P(w_k|w_{k-N+1},.,w_{k-1})+\alpha\cdot P(w_k|w_{k-N+2},.,w_{k-1}). \tag{9.39}$$

A simítási eljárások széles választéka áll rendelkezésre, melyből az aktuális feladathoz illőt kell kiválasztani. Fontos kérdés, hogy mikor tekinthetünk jónak egy nyelvi modellt, mikor mondhatjuk, hogy jól illeszkedik a felismerési feladathoz. Ennek mérésére szolgál a *perplexitás* vagy *bonyolultság*, amelyet az információelméletbeli entrópia alapján adhatunk meg. Tehát a nyelvet információforrásként fogjuk fel, amely adott szótárból szósorozatokat bocsát ki, ekkor az entrópia:

$$H = -\lim_{m\to\infty}\frac{1}{m}\sum_{w_1,...,w_m} P(w_1,...,w_m)\cdot\log_2 P(w_1,...,w_m). \tag{9.40}$$

Említettük már, hogy a nyelv ergodicitását feltételezzük, ekkor az entrópia az alábbi egyszerűbb alakú:

$$H = -\lim_{m\to\infty}\frac{1}{m}\log_2 P(w_1,...,w_m) \tag{9.41}$$

Ha m elég nagy, akkor a határértékképzés elhagyásával jó közelítést kapunk:

$$\hat{H} = -\frac{1}{m}\log_2 P(w_1,...,w_m). \tag{9.42}$$

A perplexitás definíció szerint a következő:

$$PP = 2^H \tag{9.43}$$

Felhasználva az előbbi (9.42) összefüggést:

$$PP = P(w_1, ..., w_m)^{-\frac{1}{m}} \tag{9.44}$$

A perplexitás segítségével jól mérhető, hogy egy adott szót hány különböző, nagy valószínűségű szósorozat követhet. Minél több, annál bonyolultabb (illetve annál kevésbé illeszkedő) a nyelvi modell által megadott nyelvtan. Fontos látnunk azonban, hogy a perplexitás a beszédfelismerés szóhibaarányával csak kismértékben korrelál, hiszen az akusztikai illeszkedést nem képes figyelembe venni.

9.6.2.2. Környezetfüggetlen nyelvtanok

Jól körülhatárolt esetekben úgynevezett környezetfüggetlen nyelvtanokkal is jól megadható a nyelvi modell, ami hatékony feldolgozást tesz lehetővé (Mihajlik 2010). A nyelvtan ebben az esetben szabályok összességeként értelmezendő, amelyeket valamilyen rögzített formalizmus alapján fogalmazunk meg. Általános esetben, illetve komplex feladatoknál, mint például egy diktáló rendszer természetes nyelvi feldolgozó egysége, a korábban tárgyalt N-gram modelleket alkalmazzák, azonban a kutatások előrehaladtával ezeket a feladatokat is igyekeznek formalizálni a nagyobb felismerési biztonság érdekében. A számítógépek világában a nyelvtanok ma már alapkőnek tekinthetőek. A UNIX operációs rendszer alapvető részét képzi egy általános nyelvtani elemző algoritmus. A nyelvtanok elterjedését a fordítóprogramok indikálták. A processzornak magának is van egy saját egyszerű nyelve, azaz a processzor csak a szintaxisának megfelelő memóriatartalmat hajlandó feldolgozni. A programozók munkájuk hatékonyabbá tétele érdekében hamar elkészítették a magas szintű programozási nyelveket. A programozási nyelveknek vannak kulcsszavai (terminálisok) és ezek sorrendjét leíró szabályok. A szabályok megfogalmazását segítik a kulcsszavak sorrendiségének egy halmazát összefoglaló nemterminálisok. A magas szintű programnyelven megírt programot a fordítóprogram első lépésben szintaktikailag ellenőrzi. A fordító beolvassa a kulcsszavak, azaz a terminálisok sorozatát, és leellenőrzi, hogy a nemterminálisok segítségével megfogalmazott szabályok által megadott sorrendben érkeznek-e a terminálisok. A szintaktikai ellenőrzés akkor sikeres, ha a terminálisok elfogytak, és eközben nem volt hiba. Ezt a jellegű feladatot általánosan ellenőrzésnek hívhatjuk, azaz ellenőrizzük, hogy a leírt mondat eleme-e a nyelvnek (szintaktikailag megfelel-e a nyelvi szabályoknak). A fordítás második lépése a transzformáció, amikor a magas szintű nyelv elemeinek sorozatát át akarjuk alakítani a processzor kulcsszavainak sorozatára. Ekkor jutnak a nemterminálisok fontos szerephez. A nemterminálisok tekinthetők részszabályoknak. Például az if-else utasítást elfogadó nemterminális. Amennyiben az elemzett kód egy részletét

elfogadja egy nemterminális, a nemterminálishoz megadott processzor kód generálódik. A kód természetesen nem teljes, hiszen az if-else utasítás szabálya további nemterminálisok segítségével van megfogalmazva (például a logikai feltétel szintaktikáját megadó nemterminális). Ezen nemterminálisok saját fordítása segítségével rekurzívan fordul végül az utasítás a processzor nyelvére. A nyelvtanok ilyen felhasználása, azaz a nemterminálisokhoz vagy szabályokhoz rendelt extra információ alkalmas az elemzett mondat transzformálására. A transzformálás pedig természetesen lehet a mondat által kódolt információ dekódolása, a mondat szemantikus elemzése, azaz a mondat megértése is. A nyelvtanok megértését segítheti a következő példa. A példa egy olyan szabályrendszert valósít meg, amely elfogadja a számokat 0 és 99 között. A végső szabály, a MONDAT, nemterminális, amit két szabály ad meg: érvényes terminális sorozat a számok 0 és 9 között, és a számok 10 és 99 között. A terminálisokat a SZÁMJEGYEK-1–9 és SZÁMJEGYEK-0–9 nemterminálisok halmazai igyekeznek absztraktan kezelni. A nemterminálisok megadása tehát:
SZÁMJEGYEK-1–9 ::= ’1’;
SZÁMJEGYEK-1–9 ::= ’2’;
. . .
SZÁMJEGYEK-1–9 ::= ’9’;
SZÁMJEGYEK-0–9 ::= ’0’;
SZÁMJEGYEK-0–9 ::= SZÁMJEGYEK-1–9;
A MONDAT ekkor az alábbi szintaxissal adható meg:
MONDAT ::= SZÁMJEGYEK-1–9 SZÁMJEGYEK-0–9;
MONDAT ::= SZÁMJEGYEK-0–9; Természetesen igaz, hogy a 0 és 99 közé eső számokat fel is sorolhattuk volna. Felmerülhet, hogy miért van akkor szükség a látszólag bonyolult nyelvtanokra: már ez az egyszerű példa is érzékeltetheti a nyelvtanok absztrakciós képességét, bonyolultabb nyelvtan esetén pedig a felsorolás kezelhetetlen méretű lenne. Amennyiben rekurzív szabályt is alkalmazunk, az összes mondat felsorolásának lehetősége meg is szűnne (azaz ebben az esetben felsorolással nem tudnánk megadni az elvileg megszámlálhatatlanul sok lehetőséget, mert a nyelv mondatai végtelen elemszámú halmazt alkotnak). A nyelvtanok a szabályok struktúrája alapján osztályozhatók. Az egyik osztály a környezetfüggetlen nyelvtanok osztálya. (A környezetfüggetlen nyelvtanok részhalmazát képzik a reguláris nyelvtanok. A reguláris nyelvtanok ezért érdekesek, mert a szabályok egyértelműen meghatároznak egy determinisztikus véges automatát.) A környezetfüggetlenség arra utal, hogy a szabályokban a terminális és nemterminális szimbólumok tetszőlegesen követhetik egymást. A nyelvtanok alkalmazásához természetesen szükség van egy olyan szoftvereszközre, amely fogadja az input terminális sorozatot, és ellenőrzi, hogy az megfelel-e a szabályoknak. Ez az eszköz tehát képes valamely speciális szintaktikával elkészített szabályrendszert értelmezni, és ezek alapján az input mondatokat kiértékelni. A beszédfelismerésben a parancsmódú beszédfelismerő rendszerek alapvető eszköze a nyelvtan. A felismerő által elfogadható parancsokat ekkor

úgynevezett nyelvtanok formájában fogalmazzuk meg. A mintaillesztő rendszer által kiadott alternatívákat az utófeldolgozó egység az aktív nyelvtanok segítségével leellenőrzi, és a legvalószínűbb érvényes mondatot tekinti felismertnek. A következőkben példaként a fájlműveleteket fogalmazzuk meg:

FÁJLMŰVELETEK ::= 'fájl' MŰVELET;
fájlműveletek := CTRL
MŰVELET ::= 'megnyitása';
művelet := O
MŰVELET ::= 'mentése';
művelet := S
MŰVELET ::= 'nyomtatása';

művelet := P azaz a fájl kulcsszót kimondva és a kívánt műveletet megjelölve a megadott vezérlőbillentyű-kombinációkkal azonos hatást érünk el a billentyűzet használata nélkül, tehát például a 'fájl megnyitása' szóbeli parancs ekvivalens a CTRL+O billenytűkombináció használatával. Úgy is fogalmazhatunk, hogy a szóbeli parancs alá beágyazódik a szemantikus információ, hiszen a gép tudja, hogyan értelmezze a szóban adott parancsot. Amennyiben valamelyik felismert hipotézis a 'fájl' szóval kezdődik, az első szabály teljesül, így a CTRL információ letárolásra kerül. A CTRL a parancs gyorsbillentyű-kombinációjának első billentyűje. Ha a kimondott parancs eleget tesz a MŰVELET nemterminális által definiált szabálynak is (azaz létező parancsot mond a felhasználó), akkor a hipotézist elfogadjuk, és a tartalmi információ a gyorsbillentyű-kombináció második tagjával kiegészül. Az alkalmazás természetesen a felismert parancsnak megfelelő szemantikus információ alapján egyértelműen végre tudja hajtani a parancsot. Ahogyan a parancsmódú rendszerek átalakulnak dialógusrendszerekké, a nyelvtanok definíciója és felhasználása is átalakul. Amennyiben megkísérelnénk leírni egy természetes nyelvet környezetfüggetlen nyelvtan segítségével, valószínűleg bevezetnénk a jól ismert nyelvtani fogalmainkat: alany, állítmány, jelző, határozó stb. A nyelvtani fogalmakat formálisan a nemterminálisok fogalmazzák meg. A dialógusrendszerekben a nyelvtanok szerkezete éppen ezt az utat követi. Az ilyen rendszerekben a cél a felhasználó szándékának megértése, és nem feltétlenül a teljes hűségű beszéd-szöveg átalakítás. (Ugyanazt a dolgot számtalan módon és formában lehet kifejezni. Ha a beszélő valamilyen információt kér, akkor az a lényeg, hogy tudjuk, mit szeretne megtudni, és nem az, hogy a kérdését pontosan milyen szavakkal fogalmazta meg.) Ez úgy lehetséges, ha a tipikus kéréseket a rendszert programozók előzetesen kiismerik. Például, egy repülőtéri információs rendszer esetében előre tudjuk, hogy a felhasználó egyszer ki fogja mondani, hogy honnan, hova és mikor akar utazni. Ezek a határozóval analóg fogalmak (concept). Nagyon jól tudjuk, hogy milyen nyelvtani szerkezet azonosítja a honnan, a hova információt ('Budapestről', 'Londonba' stb.). Ekkor tehát a nyelvtan a keresett információt kifejező nyelvtani szerkezeteket fogja tartalmazni a parancsok helyett. A nyelvtan tartalmán kívül módosításra van szükség az elemző eszközben

is. Eddig az elemzés akkor volt sikeres, ha elfogyott a terminális szimbólumsorozat, és eközben nem tapasztaltunk hibát. Az információkiemelés alapvető ismérve, hogy nem igyekszik megadni az érvényes mondatok teljes halmazát, hanem a keresett információ nyelvtani fordulataira szorítkozik csupán. (Ennek is köszönheti sikerességét, mert az élőbeszéd gyakran sérti meg a szabályokat, ami a teljes körű nyelvtani elemzés alkalmazása esetén sikertelen felismerésre vezetne.) Az elemzést tehát nem a felismert mondat elején kell kezdeni, hanem a terminális sorozat minden egyes elemét potenciális mondatkezdetnek kell tekinteni. Más szavakkal a nyelvtani elemző elfogadja, ha egy érvényes mondat a terminális sorozat közepén helyezkedik el. Ezt a keresési eljárást *kulcsszókeresés*nek vagy *word spotting*nak nevezik, hiszen az eljárás szavakat, szókapcsolatokat keres a felismert szósorozatban.

9.7. Zajtűrő beszédfelismerés

A gépi beszédfelismerők rendkívül érzékenyek a környezeti zajra, illetőleg az átvitelt biztosító csatorna (például telefon) műszaki tulajdonságaira (zajos és keskeny az átviteli sáv). Mindkettő jelentősen befolyásolja a felismerés pontosságát. A zajos környezetben is jó teljesítményt nyújtó felismerőket *zajtűrő* (noise robust) felismerőknek hívják. Általánosan igaz, hogy a felismerés abban az esetben a leghatékonyabb, ha a felismerőt olyan akusztikai környezetben használják, amilyen környezetben a betanító mintákat is rögzítették. A környezet szerepe azért jelentős, mert egyrészt befolyásolja a beszéd akusztikai jellemzőit, másrészt a környezetből származó esetleges zajok a beszédjelhez hozzáadódnak. Az akusztikai környezetbe bele szokás érteni a beszédjel rögzítését és feldolgozását, esetleges átvitelét biztosító jelfeldolgozó eszközök és átviteli csatornák hatásait, amelyet e beszédfeldolgozási lánc átviteli karakterisztikája jellemez. A beszédfeldolgozási lánc első eleme tehát a beszédet elektromos jellé alakító mikrofon, melynek karakterisztikája és jellemzői (iránykarakterisztika, torzítások a karakterisztika frekvenciafüggése miatt, érzékenység, közel- illetve távolbeszélőség) jelentősen befolyásolják az átalakított jel akusztikai sajátságait. A láncban a további elemek, a mikrofon jelét továbbító csatorna átviteli karakterisztikája, majd a tulajdonságvektorok előállítását végző előfeldolgozó egység karakterisztikája. Az előfeldolgozó egységben történik az analóg-digitális átalakítás (kivéve telefonos átvitel esetén) – a mintavétel és ezzel sávkorlátozás, kvantálás, majd a szűkebb értelemben vett előfeldolgozás, a tulajdonságvektorok előállítása például szűrősoros elemzéssel. Tehát ezen lépések mindegyike befolyásolja magát a mintaillesztést és így a felismerést is. Speciális eset a telefonos beszédfelismerés, ekkor a telefonhálózaton továbbított (de akár GSM vagy VoIP kódolóval kódolt majd dekódolt) jelen dolgozunk, az átviteli hálózat és a kódolók tulajdonságai nyilvánvalóan jelentősen befolyásolják a beszédjel spektrális sajátságait. A telefonon

átvitt jel jelentős sávkorlátozáson esik át (a vezetékes telefóniában például a 300 Hz és 3400 Hz közötti sávra), ez információveszteséget is okoz. Érdekesség, hogy az emberi kommunikációt ez a veszteség általában nem zavarja, telefonon is jól megértjük beszélgetőpartnerünket, a gépi felismerésben azonban szignifikáns különbséget tapasztalunk a beszédfelismerés pontosságában, döntően a sávkorlátozás miatt. A telefonhálózaton keresztül történő használatra szánt beszédfelismerők akusztikai modelljeit szintén a telefonhálózaton keresztül rögzített tanítómintákkal célszerű betanítani, ugyanis a telefonos átvitel hatása más módon nehezen szimulálható. (Természetesen szűrhető a telefonhálózat karakterisztikáját közelítő eljárásokkal a nem telefonos környezetben rögzített beszédminta is, de általános tapasztalat, hogy az így készített felvételeken betanított akusztikai modellek valós telefonos környezetben gyengébben teljesítenek a mintaillesztés során.) Az akusztikai környezet másik fontos összetevőjét a környezeti zajok adják. A környezeti zajokat általában két csoportra bontják: a hirtelen, lökéshullámszerű, rövid ideig tartó zajokra és az időben hosszabb távon változatlan, tartósan fennálló stacionárius zajokra. A környezeti zajok kezelése a beszédfelismerésben kétféle lehet: taníthatunk olyan akusztikai modelleket, amelyeket zajos minták alapján becslünk, vagy megpróbálkozhatunk a zaj leválasztásával, szűrésével. Általában igaz, hogy a hirtelen zajokkal nem nagyon lehet mit kezdeni, míg a stacionárius zajok esetén mindkét módszer működhet. Fontos tudnunk azonban, hogy napjainkban a legkifinomultabb zajszűrő eljárás sem képes olyan szinten szétválasztani egymástól a beszédjelet és a zajt, mint az emberi percepció biológiai beszédfeldolgozó apparátusa.

9.7.1. Az átviteli csatorna hatását kompenzáló normalizációs eljárások

Ha a beszédfeldolgozási lánc átviteli karakterisztikája nem állandó, valamint ha a betanításra használt minták rögzítésekor, illetve a felismerő felhasználásakor a csatorna és/vagy az előfeldolgozó átviteli karakterisztika eltér, hasznos lehet valamiféle normalizáció alkalmazása a különbségek mérséklése érdekében. Az egyik legrégebbi erre szolgáló technika a kepsztrális átlaggal történő normalizálás (Cepstral Mean Normalization, CMN). Az eljárás alapja a következő: az átvitel után a jel spektrumát úgy kapjuk, hogy a jel eredeti (átvitel előtti) spektrumát szorozzuk a csatorna átviteli karakterisztikájával. Ha a kepsztrális tartományban gondolkodunk, akkor ez a szorzás egyszerű összeadásnak felel meg, nincs más dolgunk tehát, mint az MFC jellemzőket tartalmazó tulajdonságvektorokból levonni a becsült átlagos tulajdonságvektort. Az eljárás hátulütője, hogy az átlag számításához a teljes beszédjelre szükségünk van, így valós idejű működésre az eljárás nem alkalmas (meg kell várnuk az utolsó keretet is, így számíthatunk átlagot, majd kezdhetjük a normalizálást

az első kerettel – ez a beszédminta hosszával egyenlő késleltetést ad).
Megjegyzés:

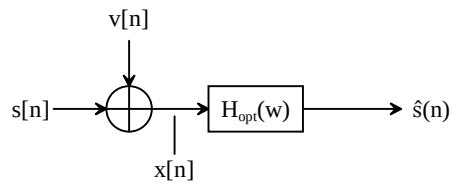
- Általában biztosítható, hogy az előfeldolgozó egység bemenetére érkező jel minden esetben azonos algoritmussal kerüljön feldolgozásra. Ha viszont a jel valamilyen átviteli csatornán érkezik, a csatorna átviteli karakterisztikája időben változhat. Hasonló a helyzet akkor, ha a felhasználó mikrofont cserél, a mikrofont máshogy állítja be vagy más szögből beszél a mikrofonba. Az átviteli csatorna és a mikrofon esetében tehát nem biztosítható az állandóság, a normalizálás célja ezen hatások mérséklése.

Nem kifejezetten csak az átviteli csatorna hatását kompenzálja az energia normalizálása. Ez akkor lehet hasznos, ha a beszélők eltérő hangerővel beszélnek, illetve a mikrofontól más távolságban vannak, a mikrofonba más szögből beszélnek, más érzékenységű mikrofont használnak stb. A normalizálás problémája ugyanaz, mint a CMN esetén, meg kell várni a beszédminta utolsó keretét. A normalizálás rendszerint úgy történik, hogy az adott beszédminta legnagyobb eregiájú keretét fix szintre állítják, a többi keretet az adódó különbséggel szintén módosítják. A számos audio és egyéb eszközben alkalmazott automatikus adaptív erősítés (AGC, Automatic Gain Control) is felhasználható a beszédfelismerésben (az ilyen eszközök a beszédminta egészét tekintve nemlineáris erősítést végeznek), felhasználása azonban körültekintést igényel és nem is feltétlenül előnyös.
Megjegyzés:

- Az átviteli csatorna esetében számos egyéb csatornakiegyenlítő eljárás is használható, mivel ez már átviteltechnikai kérdés, e helyütt részletesen nem foglalkozunk vele.

9.7.2. Zajszűrő eljárások

Napjainkban fontos kutatási terület a zajos közegben történő beszédfelismerés. Különböző zajszűrési algoritmusokkal a jel többé-kevésbé megtisztítható a zajtól, általánosan igaz azonban, hogy a zajos közegben történő beszédfelismerés a zajtalan esethez képest sokkal rosszabb eredményt ad akkor is, ha zajszűrést végzünk. A zaj a gépi beszédfelismerést sokkal jobban zavarja, mint az emberi hallgatót. A zajt additív zajként modellezik, amely a beszédjelhez adódik hozzá. Amint a 9.11. ábrán látható, az $s[n]$ beszédjelhez a $v[n]$ zaj adódik. Feladatunk, hogy az $x[n]$ megfigyelhető jelből előállítsuk a zajtól lehető legjobban megszűrt $\hat{s}[n]$ jelet valamilyen $H_{opt}(\omega)$ optimális szűrővel. A zajszűrés klasszikus megvalósítása a Wiener-szűrő. A Wiener-szűrők központi szerepet játszanak számos alkalmazásban, például lineáris predikció, jelkódolás, visszhangkioltás, jelvisszaállítás és csatornakiegyenlítés



9.11. ábra. A zaj modellezése)

megoldásaiban. A Wiener-szűrőt úgy számoljuk, hogy a szűrő kimenete és a kívánt jel átlagos négyzetes távolsága minimális legyen. Ez tehát azt jelenti, hogy az $e[n] = s[n] - \hat{s}[n]$ hibajel négyzetét minimalizáljuk, az ezt teljesítő $H_{opt}(\omega)$ szűrő a Wiener-szűrő. A szűrő átviteli függvényének meghatározásához ismernünk kell a jelek spektrálsűrűség-függvényeit, amelyeket az autokorrelációs függvényekből Fourier-transzformációval elő tudunk állítani, tehát az autokorrelációs függvények ismerete is elégséges. Az átviteli függvény:

$$H_{opt}(\omega) = \frac{S_{sx}(\omega)}{S_{xx}(\omega)}, \tag{9.45}$$

ahol $S_{xx}(\omega)$ az $x[n]$ jel autokorrelációs függvénye, $S_{sx}(\omega)$ pedig az $s[n]$ és $x[n]$ jelek keresztkorrelációs függvénye alapján számítható. A Wiener-szűrő együtthatói meghatározhatók pusztán a jel és a zaj autokorrelációs függvényeiből is, feltételezve, hogy azok nem korreláltak:

$$H_{opt}(\omega) = \frac{S_{ss}(\omega)}{S_{ss}(\omega)S_{vv}(\omega)}. \tag{9.46}$$

A Wiener-szűrő átviteli függvényének utóbbi alakja jól mutatja, hogy ha a zajteljesítmény egy adott f frekvencián (vagy körfrekvencián: ω, $\omega = 2\pi f$) nagy, akkor a nevezőben $S_{vv}(\omega)$ is nagy, így a szűrő erőteljesebb elnyomást alkalmaz. Ha a zaj teljesítménye elhanyagolható, akkor az átviteli függvény 1-hez közelít. A Wiener-szűrőt diszkrét idejű rendszerekben leginkább véges impulzus válaszú (FIR) szűrőként szokták megvalósítani:

$$\hat{s}[n] = \sum_{k=0}^{P-1} w_k x[n-k], \tag{9.47}$$

ahol P a szűrő fokszáma, w_k-k pedig a szűrő együtthatói. Ha a négyzetes hibát szeretnénk minimalizálni, akkor az együtthatókat

$$w_k = \arg\min_k E(e^2[n]) \tag{9.48}$$

alapján keressük, ahol $E(.)$ a várható érték képzésének operátora (a várható értéket azért kell képeznünk, mert sztochasztikus jelekkel dolgozunk). A minimalizálás tehát úgy végezhető el, ha $k = 0...P-1$ értékekre w_k szerint deriváljuk és megkeressük a derivált azon zérushelyét, ahol a második derivált nagyobb, mint 0 (ekkor biztosan minimumot találtunk). A hibajelet felírva:

$$E(e^2[n]) = E\left((\hat{s}[n]-s[n])^2\right) = \tag{9.49}$$

$$= E\left(\hat{s}[n]^2\right) + E\left(s[n]^2\right) - 2E\left(\hat{s}[n]s[n]\right) = \tag{9.50}$$

$$= E\left(\left(\sum_{k=0}^{P-1} w_k x[n-k]\right)^2\right) + E\left(s[n]^2\right) - 2E\left(\sum_{k=0}^{P-1} w_k x[n-k]s[n]\right). \tag{9.51}$$

Tehát a w_k szerinti parciális derivált:

$$\frac{\partial}{\partial w_k} E(e^2[n]) = 2E\left(\sum_{l=0}^{P-1} w_l x[n-k]x[n-l]\right) - 2E\left(x[n-k]s[n]\right) = \tag{9.52}$$

$$= 2\sum_{l=0}^{P-1} E\left(w_l x[n-k]x[n-l]\right) - 2E\left(x[n-k]s[n]\right) = \tag{9.53}$$

$$= 2\sum_{l=0}^{P-1} R_{xx}[l-k] - 2R_{sx}[k], \tag{9.54}$$

hiszen $x[n]$ autokorrelációs függvénye, valamint $s[n]$ és $x[n]$ keresztkorrelációs függvénye éppen:

$$R_{xx}[m] = E(x[n]x[n-m]), \tag{9.55}$$

$$R_{sx}[m] = E(x[n]s[n-m]). \tag{9.56}$$

Az egyes parciális deriváltakat nullával egyenlővé téve P egyenletből álló lineáris egyenletrendszert kapunk, melynek megoldása a szűrő együtthatóit szolgáltatja. Természetesen léteznek más, szofisztikáltabb szűrési eljárások is, ezekre részletesen nem térünk ki, a szűrés főbb elveit azonban bemutatjuk. A zajszűrő eljárások általában a következő sémán alapulnak (Weiss et al. 2002): a zajos jelet transzformáljuk, majd a transzformált tartományban szűrést, illetve egyéb jelfeldolgozási lépéseket végzünk, amelyekkel célunk a zaj minél jobb eltávolítása. Ezután inverz transzformációval ismét időtartományba képezzük a jelet. A Wiener-szűrés is ezt a sémát követi, ekkor a spektrálsűrűség-függvénnyel végeztük a transzformációt, de alkalmazható *wavelet-transzformáció*, illetve az úgynevezett *modulációs spektrum* előállítása is. A szűrők tervezésénél általános elv a már megismert négyzetes hibaminimalizálás. Megjegyezzük, hogy a wavelet-transzfromáció a Fourier-transzformáció

alternatívája, a modulációs spektrum pedig a spektrum egyes frekvenciakomponenseinek időbeli változását adja meg, de itt részletesen nem tárgyaljuk őket. Ha a zaj additív és időben viszonylag állandó, a spektrális kivonás (spectral subtraction) is alkalmazható, feltéve, hogy lehetőségünk van a zajból mintát venni (azaz azonosítani tudunk valamely zajjal terhelt, de beszédet nem tartalmazó részletet): ekkor a zaj spektrumát kiszámítva ezt a spektrumot a teljes, zajjal terhelt beszédjel spektrumából levonva tisztíthatjuk meg a beszédjelet. Alapvető zajszűrési technika a feldolgozott frekvenciasávra történő szűrés. A zajszűrő eljárásokat általában adaptívan valósítják meg, azaz a $H_{opt}(\omega)$ karakterisztika folyamatosan, az aktuálisan mért vagy becsült zaj jellemzőinek megfelelően változik.

9.7.3. A beszélő személytől származó zajok kezelése

A beszélő személytől származó zajok a zajok egy különleges csoportját képezik. Ilyenek a levegővétel zaja, a kilégzés zaja, nyelvcsettintés. Idetartozik a köhögés, torokköszürülés is, sőt tágabb értelemben az olyan hangesemények is, amelyeknek a közlésben is funkciója van, de nem írható le klasszikus beszédhangok sorozataként, tipikusan a kitöltött szünetek, a hangos hezitálás, a hümmögés (vö. Markó 2006). E hangjelenségek közös vonása, hogy a beszéddel nem lapolnak át, hiszen a beszélőtől származnak. A háttérbeszélgetés épp ezért nem ilyen zaj! A zajtűrő beszédfelismerés egyik legnagyobb kihívása a beszéd kiemelése háttérbeszédből. Az emberi hallgató képes a számára releváns beszélő beszédét kiemelni a zavaró háttérbeszédből – ezt tréfásan „koktélparti-effektusnak" is nevezik – a gépi rendszerek azonban erre nem képesek). Épp azért, mivel e zajok a beszédtől jól elkülönülnek, könnyedén modellezhetők maguk is a beszédhangokhoz hasonlóan, akár rejtett Markov-modellekkel is. Az egyetlen nehézség velük kapcsolatban, hogy a mintaillesztéshez generált hálózatba, illetve ehhez a nyelvi modellbe is be kell csempészni őket. Szabályalapú nyelvi modellezésben ez talán könnyebb feladat, statisztikai nyelvi modellben azonban nem, mert a nyelvi modell betanítására szolgáló korpusz rendszerint írott, és abban nincsenek jelölve a beszélőtől származó zajesemények. Mivel a szöveg írásbeli tagolása nem feltétlenül egyezik meg a beszédbeli tagolással, még az írásjelek sem adnak biztos támpontot arra vonatkozólag, hol lehet inkább egyes beszélőtől származó zajesemények (például levegővétel) megjelenésére számítani. A legjobb megoldás természetesen az, ha a nyelvi modell olyan korpusz alapján készül, amelyben a beszélőtől származó zajok is jelölve vannak (BEA adatbázis 8.5. fejezet), ennek híján egyéb trükkökkel közelíthető valamennyire a nyelvi modell a valósághoz.

9.7.4. Beszéd-nem beszéd detektálás

A beszéd-nem beszéd detektálás – azaz annak meghatározása, mikor van beszéd a kimeneten és mikor nincs – a beszédfelismerés egyik alapfeladata. Azért e helyütt tárgyaljuk, mert – egyrészt – a feladat része lehet a beszédet, illetve a zajt, valamint a beszédet, zajos beszédet és a csak zajt tartalmazó részek elkülönítése is, másrészt a beszéd-nem beszéd detektálás egyik lehetséges módja a csönd modellezése a beszélőtől származó zajok analógiájára. Ez utóbbi megoldást szinte valamennyi rejtett Markov-modell alapú beszédfelismerőben használják, ekkor a csöndnek külön modellje van. A csöndmodell általában „rövidre zárható" topológiájú, azaz a modell első állapotából közvetlen átmenet biztosított az utolsóba, ily módon a csönd el is maradhat. Ez a modellezési trükk lehetővé teszi, hogy a nyelvi modellben minden egyes szó után feltételezzük a csönd tartását, ami azonban el is maradhat. Mivel a beszédszünetek hossza igen tág határok között változik, a csönd (beszédszünet) modellezése önmagában általában nem elegendő. A hosszabb beszédszünetek esetén (körülbelül a másodperces vagy magasabb nagyságrendű szünetek esetén) tanácsos a felismerőben a mintaillesztést felfüggeszteni, ekkor ugyanis a felismerő mintegy „illesztési kényszerbe" kerülhet. Ennek oka, hogy a modellek alapján akusztikailag jellemzően kevésbé valószínű a hosszabb szünet (ez a ritkább), a bemenetre pedig csendes környezetben is érkeznek különböző környezeti zajok, emiatt a mintaillesztő a zajokat is hajlamosabb értelmes szavaknak megfeleltetni. A beszéd-nem beszéd detektálás legegyszerűbb módja a bemenetre kerülő jel energiájának mérése, bizonyos küszöbszint alatt a jelet csendnek tekintjük. A módszer hátránya, hogy zajosabb környezetben nem használható, viszont irodai diktálásra (vagy egyéb kis zajú környezetre) tervezett felismerőkben igen hatékony. Zajos környezetben a beszéd-nem beszéd elkülönítése nehézkesebb. Ilyen esetekben általában automatikus osztályozókat használnak fel, amelyek gépi tanulással is előállhatnak (szupport vektor gépek, döntési fák, multi-layer perceptron stb.). Az osztályozókban olyan jellemzőket számítanak a beszédjelből, amelyek alapján a leghatékonyabb beszéddetektációt tudják megvalósítani (ez nem más, mint a feladathoz optimalizált lényegkiemelés, akárcsak a beszédfelismerés esetében). Az osztályozást esetenként zajszűrés is megelőzheti. Az osztályozás történhet keretenként vagy nagyobb egységekre is. A beszéddetektálásra az angol szakirodalom gyakran a VAD (Voice Activity Detection) terminussal hivatkozik.

9.8. Beszélőadaptáció

Az emberi beszéd igen változatos, bizonyos mértékig pedig egyénre jellemző, ez igaz mind az akusztikailag mérhető paraméterekre, mind a nyelvhasználatra. Ésszе-

rű feltevés, hogy a gépi beszédfelismerő alkalmazások sokkal hatékonyabban működhetnek, ha a bennük alkalmazott akusztikai és/vagy nyelvi modelleket az egyéni jellegzetességeknek megfelelően módosítjuk, azaz *adaptáljuk*. Elméletben az is járható út, hogy az egyéni jellegzetességek hatását megpróbáljuk semlegesíteni a figyelembe vett paraméterekben, ehhez valamilyen *normalizáció*s eljárást alkalmazhatunk. Az adaptáció célja tehát az egyénspecifikus modellezés, míg a normalizálásé a *beszélőfüggetlen*ség biztosítása. A normalizálás korlátozottabban alkalmazható és kevésbé hatékony mint az adaptáció. A beszéd lényegi tulajdonsága a változatosság, egy ponton túl „kiszűrhetetlen" a jelenleg alkalmazott algoritmusokkal. Ez magyarázza, hogy a napjainkban kereskedelmi forgalomban elérhető, egyéni használatra készült beszédfelismerők (például diktálórendszerek) használata előtt a felismerőt a felhasználónak rendszerint saját hangjára adaptálnia kell. Természetesen nincs lehetőség adaptációra az olyan információszolgáltatást nyújtó alkalmazásokban, ahol a gyors kiszolgálás a fontos, illetve ha egy-egy felhasználó olyan rövid ideig használja a beszédfelismerőt, hogy az adaptálás értelmetlen lenne (az adaptáció időigényes folyamat). Ilyen esetekben a lehető legnagyobb fokú beszélőfüggetlenség biztosítása jön szóba.

Beszélőfüggetlen beszédfelismerőről, illetve beszélőfüggetlen modellekről (akusztikai vagy nyelvi) akkor beszélhetünk, ha reprezentatív (a becsülendő paramétereket kellően fedő) tanítóanyagon készítettük a modelleket, amelyek így univerzálisan használhatók a beszélők széles körére. Nyilvánvaló, hogy nagy számú beszélő esetén valószínűleg nagyobb a becsült paraméterek szórása, illetve a becsült átlagok sem olyan pontosak ahhoz az esethez képest, ha egy-egy konkrét beszélőtől származó tanítómintákat használnánk. Ez utóbbi esetben a felismerés a kiválasztott beszélőre jobb, a többi beszélőre viszont előreláthatólag rosszabb pontossággal működne. Az optimális megoldás az lenne, ha minden beszélőre a saját mintái alapján úgynevezett *beszélőfüggő* modelleket készítenének. Ez azonban hosszadalmas és költséges, gyakran kivitelezhetetlen. Ehhez képest hatékonyabb a beszélőadaptáció, mert a modellparamétereket nem teljes egészében tanítjuk újra, hanem a beszélőfüggetlen modellt módosítjuk vagy legfeljebb részlegesen tanítjuk újra. Az eljárás során bizonyos transzformációs kényszereket alkalmazunk, ezért kevesebb beszédminta is elegendő, így az eljárás gazdaságosabb és gyorsabb.

Felügyelt adaptáció. A felügyelt adaptáció feltétele, hogy ismert – vagy legalábbis részben ismert – tartalmú (értsd szövegesen is átírt) felvételek állnak rendelkezésünkre attól a beszélőtől, akinek a hangjára szeretnénk adaptálni a beszédfelismerőt. Ezt a típusú adaptációt azért nevezzük felügyeltnek, mert a beszédminta tartalmát a szöveges átirat pontosan megadja. Szövegesen is rendelkezésre álló hanganyagot például előre megadott, fonetikailag gazdag szöveg felolvastatásával nyerhetünk – jellemzően a beszédfelismerő első használata előtt. Lehetőség van arra is, hogy az általános, beszélőfüggetlen modellekkel működő felismerő használat közben elmentse a beszélő által bemondott mintákat. Utóbbi esetben akkor tudunk felügyelt

adaptációt elvégezni, ha szöveges átiratunk is van, ez lehet a felismerési eredmény utólagosan ember által ellenőrzött és javított átirata is. Ha biztosított, hogy a beszélő személye nem változott a felvétel során, az eredetileg beszélőfüggetlen modellünket „ráhúzhatjuk a beszélő akusztikai profiljára". Felügyelt adaptáció esetén kontrollált módon újrabecsüljük a meglévő modellek paramétereit, eközben felhasználjuk a hanganyaghoz tartozó szöveges átiratot. Így az esetleges további, szöveges átirattal már nem rendelkező részeken futtatott felismerés nemcsak pontosabb, de gyorsabb is lehet.

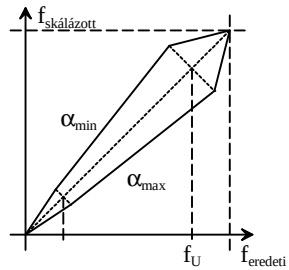
Felügyelet nélküli adaptáció. Amikor csak a hanganyag áll rendelkezésre a beszélőtől, akkor csupán felügyelet nélküli adaptációra van lehetőségünk. Ebben az esetben először beszélőfüggetlen modellekkel becslést adunk a beszédminta szöveges átiratára hagyományos beszédfelismeréssel, majd az így kapott átiratot használjuk fel egy – formálisan – felügyelt adaptációra, azaz a beszélőfüggetlen akusztikai modellparamétereinek transzformációjára. A felügyelet nélküli adaptáció hatékonysága megközelítheti a felügyelt adaptációét, de tipikusan sokkal több adatra van szüksége.

Statikus és dinamikus adaptáció. Az adaptáció *statikus*, ha egy lépésben történik, előzetesen összegyűjtött minták alapján, és *inkrementális* (dinamikus), ha a felismerés során a felhasználótól nyert adatokat folyamatosan, több lépésben újra és újra felhasználjuk a modellek adaptációjára. Az adaptáció *offline*, ha nem futási időben, és *online*, ha használat közben, automatikusan történik.

9.8.1. Az artikulációs csatorna normalizálása

A beszéd akusztikai változatossága a hangképző szervek méreteiben, alakjában és egyéb jellemzőiben meglévő interindividuális különbségekre is visszavezethető (Paczolay et al. 2003). Ezeket az egyéni különbségeket próbálja kiszűrni a tulajdonságvektorokból az artikulációs csatorna hossza szerinti normalizálás (Vocal Tract Length Normalization, VTLN). Az eljárás azon alapul, hogy a frekvenciatengelyt vetemíti, általában lineárisan, vagy szakaszonként lineárisan. A lineáris vetemítés egyébként megfelelő abból a szempontból, hogy a hangszalagok méretére vagy az artikulációs csatorna (vokális traktus) hosszára vonatkozóan valóban lineáris transzformáció kívánatos (magasabb alapfrekvencia esetén a felhangszerkezet arányosan nyúlik), viszont az artikulációs csatorna egészét tekintve ez már nem igaz (a csatorna alakja, üregei szempontjából). Ha a vetemítés lineáris (eltolást meg nem engedő) transzformációval történik, akkor egy konstans szorzófaktort (α) kell csak meghatározni. Ez a tanítás során könnyedén megtehető, amennyiben azt az értéket választjuk, amelyre a modell a tanítóadatra a legjobb illeszkedést adja (általában megkötés, hogy $0,75 < \alpha < 1,25$). Mivel a sávszélesség általában nem változhat, ezért a vetemítést

úgy kell elvégezni, hogy az elemzett frekvenciaintervallum szélső értékeit ne változtassa meg (lásd a 9.12. ábrán). A VTLN eljárás hatékonyságát illetően eltérőek a tapasztalatok.



9.12. ábra. Háromszakaszos lineáris VTLN frekvenciatengely-transzformáció megengedett tartománya

9.8.2. Akusztikai adaptáció

A korábbi fejezetekben bemutattuk, hogyan lehet az akusztikai modelleket úgy elkészíteni, hogy azok a beszéd e változatosságát jól lefedjék. Bármilyen pontosan is tudjuk azonban az akusztikai modellek paramétereit megbecsülni, attól még igaz marad, hogy ha az akusztikai változatosság tág határok között mozog, akkor a modelleknek is igazodniuk kell ehhez. Emiatt a modellek „elmosódnak", relatíve nagyobb szórást tolerálnak, kevésbé pontosan illeszkednek, így egymástól kevésbé élesen határolódnak el. Ez nehezíti a pontos mintaillesztést. Tudjuk viszont, hogy a beszéd változatossága interindividuális szinten sokkal nagyobb fokú, mint intraindividuálisan (beszélőn belül). Tehát, ha a beszédfelismerő leendő használójától származó adatbázissal tanítanánk, akkor pontosabban illeszkedő modelleket kapnánk. Mivel ez a gyakorlatban nem kivitelezhető, ezért az akusztikai modellek tanítása nem történhet egyénre szabottan. Különböző technikákkal azonban lehetőség van arra, hogy a több beszélőtől származó adatbázison betanított modelleket utólag módosítsák oly módon, hogy az az adott beszélőre jellemző akusztikai sajátosságokhoz jobban igazodjon. Ez a beszélőadaptáció, ilyenkor – szemléletesen fogalmazva – a modelleket a beszélő akusztikai profiljára húzzuk.

9.8.2.1. Akusztikai adaptáció lineáris regresszióval

Az akusztikai adaptációk alapja a lineáris regresszió maximum likelihood kritériummal, ezek az úgynevezett MLLR (Maximum Likelihood Linear Regression) eljárá-

sok. A transzformáció alapja, hogy a b paraméterek egyes normális (Gauss) komponenseiben a modellek által képviselt μ értékeket közelítsük a beszélő mintáiból becsült értékekhez. Tehát, a várhatóérték-vektorokat lineárisan transzformáljuk:

$$\hat{\mu} = \mathbf{A}\mu^T + k, \tag{9.57}$$

ahol k az eltolást adja meg, $\mathbf{A}$ pedig a transzformációs mátrix. A teljes transzformáció egyetlen mátrixszal való szorzással is megadható:

$$\hat{\mu} = \mathbf{W}\xi^T, \tag{9.58}$$

ahol $\mathbf{W} = [k,\ \mathbf{A}]$ és $\xi = [1,\ \mu]$. $\mathbf{W}$ becslése a beszélőtől származó minták alapján történik ML értelemben, tehát úgy, hogy

$$\hat{W} = \arg\max_W P(\mathbf{O}|\hat{\mathbf{M}}_\mathbf{W}) \tag{9.59}$$

teljesüljön, ahol $\hat{\mathbf{M}}_\mathbf{W}$ az adaptált modell, $\mathbf{O}$ pedig a tulajdonságvektorokból álló mátrix. A kovarianciamátrixok transzformálásához is transzformáló mátrix ($\mathbf{H}$) becslésére van szükség:

$$\hat{\Sigma} = \mathbf{H}\Sigma\mathbf{H}^T. \tag{9.60}$$

Mivel a számítások műveletigénye igen nagy, ezért gyakori megkötés, hogy a kovarianciamátrix diagonális legyen. Ugyanezen megfontolásból (számítási igény mérséklése) az MLLR adaptációnak létezik kényszerített változata is (Digalakis et al. 1995), ilyenkor a $\mathbf{H}$ mátrixot nem becslik külön, hanem az $\mathbf{A}_c$ transzformációs mátrixot és a k_c eltolásvektort úgy határozzák meg, hogy a transzformáció az alábbi alakban elvégezhető legyen:

$$\hat{\mu} = \mathbf{A_c}\mu^T + k_c \tag{9.61}$$

$$\hat{\Sigma} = \mathbf{A_c}\Sigma\mathbf{A_c}^T. \tag{9.62}$$

A transzformáció ily módon történő megfogalmazásának további előnye, hogy a modellek transzformálása helyett a tulajdonságvektorok is transzformálhatók:

$$\hat{o}_t = \mathbf{A}_c^{-1}o_t + \mathbf{A}_c^{-1}k_c. \tag{9.63}$$

9.8.2.2. Maximum a posteriori adaptáció

A maximum a posteriori (MAP) adaptáció nagyon közel áll a betanítás során becsülendő a és b paraméterek becsléséhez, ebben az esetben arról van szó, hogy az adaptálásra vállalkozó beszélőtől származó minták alapján is becslünk paramétereket, majd ezeket kombináljuk az eredetileg rendelkezésre állókkal:

$$\hat{\mu}_j = \frac{N_j}{N_j + \tau} \cdot \mu_{j,bs} + \frac{\tau}{N_j + \tau} \cdot \mu_j, \tag{9.64}$$

ahol τ súlytényező, μ_j az eredeti várhatóérték-vektor, $\mu_{j,bs}$ pedig a beszélő mintáinak átlaga (a várhatóérték-vektor becslése pusztán a beszélő mintái alapján), N_j pedig:

$$N_j = \sum_{t=1}^{T} L_j(t), \tag{9.65}$$

ahol $L_j(t)$ a j állapotban történő tartózkodás valószínűsége a t időpontban. Mivel a MAP adaptáció tulajdonképp magában foglalja egy beszélőfüggő modell tanítását is, ezért sokkal nagyobb az adatigénye (több minta szükséges a beszélőtől), mint az MLLR eljárásoknak, cserébe viszont jobb adaptációt biztosít. A két eljárás kombinálható is.

Megjegyzések:

- μ természetesen mindig egy adott állapotot jellemez (j-t), ráadásul az arra jellemző tulajdonságvektor-eloszlás Gauss-keverékének is csak egyetlen (m-edik) komponensét, így a pontos jelölés minden esetben μ_{jm} lenne. Az indexeket a korábbi levezetésekben azonban csak akkor tettük ki, ha az az összefüggés áttekinthetőségét nem rontotta.
- Az elméleti hátteret tekintve a MAP adaptáció esetén az az a priori feltevésünk, hogy a becsülendő paraméterek közel lesznek az eredetiekhez, ezért lehet kevesebb tanítóadat is elegendő ahhoz az adatmennyiséghez képest, ami a teljes beszélőfüggő paraméterbecsléshez lenne szükséges. A 9.64 összefüggésben τ tulajdonképpen az a priori tudás súlyát adja meg.

9.8.3. Nyelvi adaptáció

Az akusztikai modellekhez hasonlóan a nyelvi modell is adaptálható. Elviekben a MAP alapú akusztikai adaptáció során megismertek itt is érvényesek: célszerű, ha a felhasználótól származó írott szövegek alapján konstruált modelleket kombináljuk a már meglévőekkel. Mindez elsősorban a statisztikai nyelvi modellekre igaz, szabálybázisú modellekben valószínűleg emberi beavatkozásra van szükség a nyelvi modell igazításához. A nyelvi modell esetén az adaptáció értelme a beszélő jellemző szókincsének reprezentálása, a jellemző szóhasználati, mondatfűzési szokásainak tükrözése a modellben. A nyelvi adaptáció kötetlen szórendű nyelvekben fontosabb, mert a tág szintaktikai keretek nagyobb szabadságot adnak az egyes beszélőre jellemző kifejezésformáknak. A toldalékoló nyelvekben azért lehet fontos a nyelvi adaptáció, mert a szóalakok rendkívül nagy száma miatt nagyon nehéz akkora korpuszt összeállítani,

amelynek alapján általános N-gram nyelvi modell lenne készíthető. A nyelvi adaptáció során azt is biztosítani kell, hogy az esetlegesen megjelenő új szavak kiejtését rögzítsük a kiejtési modellben. Egyes esetekben ehhez az is szükséges, hogy a felhasználó adja meg az egyes szóalakok kiejtését valamilyen ortografikus formában.

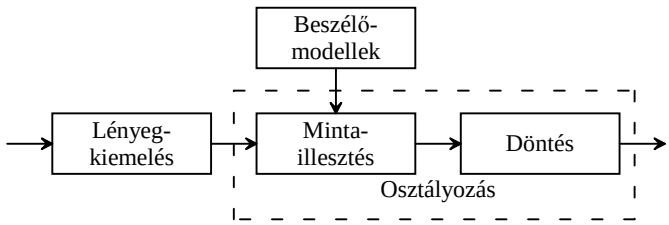
9.9. Beszélőfelismerés

A mindennapi kommunikációban az ember igen gyorsan (néhány másodperces beszédminta alapján) és meglehetősen biztosan képes az általa ismert hangú beszélőt azonosítani. Azt feltételezhetjük tehát, hogy bizonyos egyénre jellemző akusztikai sajátosságok alapján gépi úton is megvalósítható a beszélő felismerése. A mai ilyen eljárásokat három csoportba oszthatjuk: beszélőazonosítás (speaker identification), beszélőhitelesítés (speaker verification) és beszélőváltás-detektálás (speaker change detection). A *beszélőazonosítás* célja, hogy a beszélők egy ismert köréből kiválasszuk az aktuálisan beszélőt. Lehetséges, hogy a beszélők halmaza nem zárt, hanem nyílt, azaz a rendszer számára ismeretlen hangú beszélőtől származó minta is felbukkanhat. A beszélőazonosítás tehát osztályozási feladatra vezethető vissza. A *beszélőhitelesítés* esetén a beszélő személyazonosságát vizsgáljuk (ez a biometrikus személyazonosítás egyik módja, hasonlóan az ujjlenyomathoz vagy a szem íriszéhez). A beszélőhitelesítés döntési feladatra vezethető vissza (igen/nem). A *beszélőváltás detektálásá*nak több-beszélős beszédmintában van érteme (dialógusok, több beszélő váltakozása). Az elsődleges cél ekkor a beszélőváltás minél pontosabb behatárolása, akár úgy is, hogy egyetlen beszélő hangját sem ismerjük előzetesen (tehát észre kell vennünk, ha más beszél, annak ellenére, hogy azonosítani nem tudjuk az ismeretlen beszélőt). Ez osztályozási és illesztési feladatot feltételez. A beszélőazonosítás és a beszélőhitelesítés lehet *szövegalapú*, ekkor a beszélőnek előre rögzített szöveget kell bemondania. Ha a beszélőfelismerés *szövegfüggetlen*, akkor tetszőleges tartalmú beszédminta alapján végezzük a beszélő felismerését. A szövegfüggetlen megoldás rugalmasabb ugyan, de kevésbé megbízható. Érdekességként megjegyezzük, hogy a beszélőfelismerést igencsak megnehezíti, hogy a beszéd akusztikai jellegzetességei – ellentétben a szem íriszével vagy akár az ujjlenyomattal – kevésbé állandóak, illetve nehézségekbe ütközik a környezetfüggetlen beszélőfelismerés, mivel a környezeti zajok maradéktalan kiküszöbölése nem biztosítható. A beszéd sokszor hűen tükrözi a beszélő érzelmi-fizikai állapotát, így például előfordulhat, hogy az alaposan megfázott és rekedt hangú felhasználót a beszélőhitelesítő rendszer mindaddig visszautasítja, amíg meggyógyul és a hangja a régi lesz. Korlátozó tényező lehet a beszélő felismerésében a beszédminta hossza is. A gyakorlati életben, főleg a bűnügyi területen sokszor csak rövid idejű beszédhangminta áll rendelkezésre a vizsgálathoz (Nikléczy 2001), ami nehezíti az analízist. Mindezek miatt az automatikus

beszélőhitelesítés ma még nem tekinthető igazán hatékonynak, inkább még az emberi vizsgálatokra támaszkodnak.

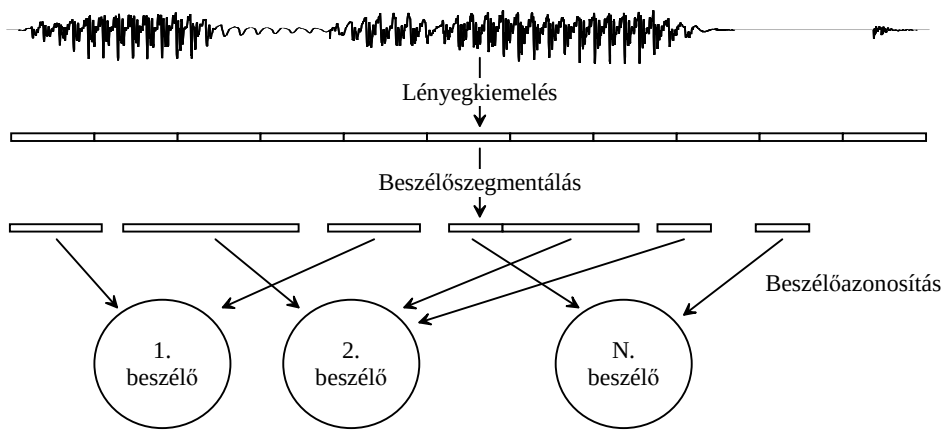
A beszélőfelismerés algoritmikus szempontból rokon a beszédfelismeréssel. Hasonló eljárásokat alkalmaznak, mint a beszédfelismerésben, hangsúlyeltolódás azonban lehet a két feladat között. A beszélőfelismerés is feltételezi (az ismeretlen beszélők váltásainak detekcióját kivéve) valamiféle referenciamodell tárolását, amely alapján a beszélő felismerése a hasonlóság mérése alapján történhet. Nagyobb szerepet kap azonban a *konfidencia* (az osztályozás, illetve a döntés megbízhatósága), azaz az, hogy az osztályozást/döntést mennyire biztosan (mekkora szignifikanciaszinten) tudtuk elvégezni. Beszédfelismerésnél ez általában másodlagos volt, hiszen a legvalószínűbb mintaillesztést kerestük. Most elsősorban az érdekel bennünket, hogy mekkora ez a valószínűség, és mekkora a különbség az egyes lehetőségek valószínűségei között. A másik fontos különbség a beszéd- és a beszélőfelismerés között az akusztikai előfeldolgozás, azaz azok a jellemzők (features), amelyek alapján a döntést/osztályozást/felismerést végezzük. Beszédfelismerésben olyan jellemzőknek örültünk, amelyek minél beszélőfüggetlenebbek, sőt, normalizálással (például VTLN, CMN) is próbálkoztunk a minél egységesebb reprezentációt elősegíteni, az egyéni specifikumokat háttérbe szorítani. Most épp az ellenkező a cél: megkeresni azokat a beszédjellemzőket, amelyek a beszélők között a legmarkánsabb eltéréseket képesek kimutatni, alátámasztani. E jellemzők egyéb kívánatos tulajdonságai: az egyszerű mérhetőség, az univerzalitás (minden beszélőnél mérhetőség), a beszélő tekintetében állandóság és megbízhatóság, nehezen utánozhatóság, a beszélő érzelmi/egészségi állapotától való lehető legnagyobb mérvű függetlenség. Ezeknek a követelményeknek számos szemantikai, illetve fonetikai-akusztikai szinten kinyerhető jellemző eleget tesz. A szemantikai jellemzők a szóhasználatot és mondatfűzési szokásokat tükrözik jól. Ezek alapján a beszélő szociális-gazdasági státusa, képzettségi háttere, származása is behatárolható (Nolan 1983). A fonetikai-akusztikai jellemzők a beszéd időtartamarányait, a prozódiát (főleg az intonáció) mutathatják meg, valamint az egyéni hangképzési sajátosságokat (Rabiner–Juang 1993).

A beszélőfelismerők felépítésükben nagyon hasonlítanak a beszédfelismerőkre: egy előfeldolgozó és egy osztályozó modulra bonthatók (9.13. ábra). Nyílt rendszerekben (ismeretlen beszélő is előfordulhat) az ismeretlen beszélőkre illeszthető modellt is el kell készíteni. Az előfeldolgozó modul végzi azon jellemzők kinyerését, amelyek alapján a beszélőfelismerés történik, az osztályozó pedig elvégzi a kiértékelést: ha a beszélő hitelesítése a cél, akkor egy hasonlósági mértéket ad a beszélő eltárolt mintája és az aktuális bemeneti minta között. Ha a beszélő azonosítása a cél, akkor valamennyi beszélő referenciamintáival el kell végezni az összehasonlítást, és így kiválasztani az aktuális beszélőt. Az osztályozás vagy döntés módszere gyakorlatilag tetszőleges lehet (rejtett Markov-modell, döntési fa, neurális háló, szupport vektor gép stb., illetve ezek hibridjei), sőt szövegalapú beszélőfelismerésre a DTW is jól használható (Jin 2007). A beszélőváltás-detektálást általában két további alfel-



9.13. ábra. Beszélőfelismerő rendszer blokkszintű felépítése

adatra bontják (Jin 2007): először elkülönítik az azonos beszélőtől származó részeket (beszélő szerinti szegmentálás), majd meghatározzák, melyik szegmens melyik beszélőhöz tartozik (beszélőazonosítás). Egy ilyen rendszer vázlatos működése látható a 9.14. ábrán. A beszélőfelismerő rendszerek teljesítményét a feladattól függően



9.14. ábra. Beszélőváltás detektálása és a beszélők azonosításának algoritmusa sematikusan

mérhetjük: beszélőazonosításban a helyesen azonosított beszélők aránya jó mérőszám, míg a beszélőhitelesítésben a tényleges beszélő hibás visszautasítása és más beszélő hibás elfogadása mérőszámkettős a használatos (ezek gyakorlatilag megfelelnek a hipotézisvizsgálat első- és másodfajú hibáinak). Ez utóbbi esetben a két hiba nem feltétlenül egyforma súlyú, a beszélő visszautasítása a személyazonosság esetén többnyire inkább bosszantó, míg a jogosulatlan személy számára hibásan megadott hitelesítésnek komolyabb következményei lehetnek. A beszélőváltás-detektálás eredményességének szintén két mérőszáma lehet: az elmulasztott váltások aránya és a tévesen feltételezett váltások száma vagy aránya. A beszélőváltás detektálásban az időbeli pontosságot (a tényleges váltás és a detektált váltási időpont közötti különbséget) is érdemes lehet vizsgálni.

9.10. A prozódia szerepe a beszédfelismerésben

Szaszák György

Korábban már megismerkedtünk a beszéd szupraszegmentális szerkezetével – más néven a prozódiával –, és azt is előrebocsátjuk, hogy a gépi beszéd-előállításban jó minőségű beszédet csak akkor kapunk, ha a prozódiát is megfelelően modellezzük. A prozódia felhasználása a gépi beszédfelismerésben korántsem ennyire elterjedt, a szűk értelemben vett beszéd-karaktersorozat átalakításhoz pedig akár nélkülözhető is lenne, hiszen ehhez elegendő lehet pusztán a beszéd spektrális paramétereinek modellezése. Miután a prozódia szerepe a beszédfolyamatban elsődlegesen a beszédfolyam értelmi tagolása, illetve a nyelvi hierarchiában magasabban, jellemzően a mondatszinten elhelyezhető információ közvetítése, a gépi beszéd*értés* megvalósításában már nem kerülhető meg. A beszédértés ugyanis nem csupán a karaktersorozattá való alakítást foglalja magába, hanem a közlemény értelmezését, jelentésének megértését is. Ehhez igen sok segítséget adhat a prozódia, hiszen a beszédfolyamot értelmileg tagolja, jelzi a mondat típusát (például állítást vagy kérdést fogalmaz-e meg a beszélő). A prozódia felhasználása a beszédfelismerésben napjainkban olvasott szövegre értendő, a spontán beszéd ugyanis egészen más prozódiai sajátságokkal rendelkezik, amelyek gépileg egyelőre nem kezelhetők hatékonyan.

A prozódia beszédfelismerésbeli felhasználásával tehát az alábbiakat érhetjük el:

- A beszédfolyamot prozódiai és/vagy hangsúlyfrázisokra (HF) tagolhatjuk, mivel a frázishatárokat és a hangsúlyt az alapfrekvencia és az energia többnyire jól jelzi. Hangsúlyfrázison két hangsúlyos szótag közötti szakaszt értünk. Ez a tagolás felhasználható többek között az írásjelek elhelyezésére (vessző és mondatzáró írásjelek), témaváltás vagy több beszélő esetén a beszélőváltások detektálására;
- A szupraszegmentális szerkezet felgöngyölítésével támogathatjuk a szövegalapú gépi nyelvi elemzést (ebben különösen problémás az esetenként számos helyes, de különböző elemzés közül az aktuálisan alkalmazandó kiválasztása, amelyben a prozódia nagyban segítheti az egyértelműsítést), javíthatjuk a beszédfelismerés eredményességét (például a prozódiailag inadekvát felismerési hipotézisek kiszűrésével);
- A mondattípusnak megfelelő írásjeleket helyzehetjük el az egyes mondattípusokra jellemző beszéddallam modellezésével és felismerésével.

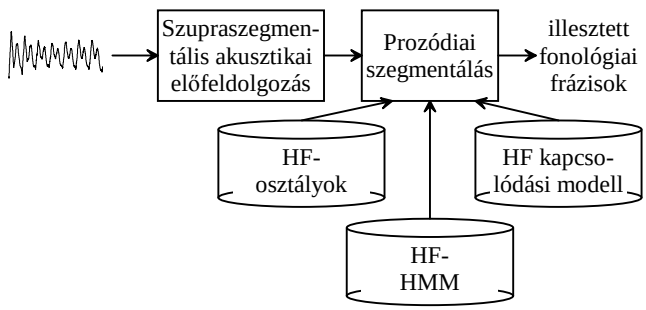
Nézzünk meg ezek közül néhány alkalmazást részletesebben! A beszédfolyam tagolása például annak alapján történhet, hogy a szupraszegmentális jellemzők, illetve ezek fizikailag is mérhető paraméterei – az alapfrekvencia, energia és az időtartamok alakulása – jellegzetes sajátosságokat mutatnak az egyes frázishatárokon. Prozódiai frázisok határán például a nyelvek széles körére jellemző az alapfrekvencia megemelése, ha a mondat még nem fejeződik be, de a beszélő – például levegővétel végett – rövid szünetet tart. A kijelentő mondatok végén ezzel szemben – legalább-

is a magyar nyelvben – mind az alapfrekvencia, mind az energia értékei csökkenő, ereszkedő jellegű tendenciát mutatnak. A detektálást általában gépi tanulás alapján valósítják meg akár döntési fákkal (Veilleux–Ostendorf 1993), akár neurálishálóalapú (Gallwitz et al. 2002), akár rejtett Markov-modell alapú osztályozóval (Szaszák 2008). Hasonlóan, a hangsúly detektálása és a hangsúlyfrázisokra történő felbontás az alapfrekvencia- és az energiaértékek alapján történhet, ezek ugyanis a hangsúlyos szótagon csúcsot – bizonyos esetekben völgyet – adnak. A detektálás történhet csúcskereséssel, vagy az egyes hangsúlyfrázisszakaszok alapfrekvencia- és energiamenetét modellező sémák beszédfolyamhoz történő illesztésével. Ez utóbbi eljárás elvi alapjaiban megegyezik a beszédfelismerésben alkalmazott mintaillesztési eljárással, az illesztett elemek azonban nem szó- vagy beszédhangmodellek, hanem most hangsúlyfrázisok prozódiai modelljei. A feldolgozott paraméterek tekintetében a beszédfelismerés MFC-együtthatói helyébe prozódiai jellemzők vagy azokból származtatott paraméterek kerülnek, és ezek alapján készülnek a frázismodellek. Kötött hangsúlyú nyelvekben a hangsúly detektálásánál messzebb is mehetünk: ha a hangsúlyt detektáljuk (vagy a hangsúlyfrázis elejét, a kettő ekvivalens), akkor – ha a hangsúly például az első szótagon kötött – egyben a szó elejét is detektálhatjuk (Szaszák 2008). A beszéd frázisokra történő tagolása a nyelvi elemzés és a beszédfelismerés javításának eszköze is lehet. A frázishatárok ismeretében például könnyedén ellenőrizhető, hogy az megfelel-e egy adott nyelvi elemzésnek (vagy a felismerési hipotézisnek): egy adott nyelvi elemzéshez ugyanis generálható a helyes ejtéshez megkívánt prozódiai struktúra, ha ehhez mint referenciához jól illeszkedik a beszéd elemzéssel nyert tényleges prozódiai szerkezete, akkor a prozódiai elemzés megerősíti az adott nyelvi elemzést vagy felismerési hipotézist (például egybeeső frázishatárok, hangsúlyok), ha pedig nem, akkor a nyelvi elemzés vagy felismerési hipotézis kevésbé valószínűen lesz helyes. A prozódia gépi beszédfelismerésbeli felhasználásában magyar nyelvre olvasott beszédben hangsúlyfrázisokra történő felbontást és mondattípus-felismerést valósítottak meg. A hangsúlyfrázisokra (hangsúlytól a következő hangsúlyig tartó szakaszokra) történő felbontás alapján pedig gépi beszédfelismerők keresési terét módosították a dekódoláskor a beszédfelismerés pontosabbá tétele érdekében (Szaszák 2008). A hangsúlyfrázisokra történő felbontás az alábbiak szerint történhet: a hangsúlyfrázisokat rejtett Markov-modellekkel modellezzük az alapfrekvencia és az energia mint akusztikai jellemzők, valamint ezek első és második deriváltjai alapján. A modellezésre elkülönített hangsúlyfrázistípusok a 9.1. táblázatban láthatók (vö. 6.2. fejezet). Ezt követően a bemenetre kerülő beszédjelből kinyerjük az alapfrekvenciát és az energiát, majd a modellekkel mintaillesztést végzünk. A mintaillesztés eredménye az illeszkedő hangsúlyfrázisokat adja meg, kezdő és végidőpontjukkal együtt. Mivel a magyar nyelv első szótagon kötött hangsúlyozású (azaz a hangsúlyt minden esetben egy szó első szótagján találjuk a durva kivételektől eltekintve), ezért a hangsúlyfrázis eleje szinte pontosan egybeesik valamely szó elejével. Ily módon tehát tulajdonképpen szavak kezdőpontjait kapjuk meg a beszédjelből. Ha a beszéd-

9.1. táblázat. Hangsúlyfrázistípusok magyar nyelvre a gépi osztályozáshoz.

Címke	Leírás
me	Mondat elejére eső HF
fe	Erős hangsúlyú HF
fs	Hangsúlyfrázis
fv	Prozódiai frázis végére eső HF
s	Hangsúlytalan szakasz vagy rövid szünet
mv	Mondat végére eső HF
sil	Szünet (csend)

felismerés során előálló hipotézisgráfot tekintjük, abban a lehetséges felismerési útvonalak találhatóak meg. A gráf csomópontjai pedig gyakran éppen szavak közötti átmeneteknek felelnek meg. Kézenfekvőnek tűnik ezért, hogy a hipotézisgráfot összevessük a hangsúlyfrázisok elhelyezkedésével, és ennek alapján módosítsuk azt: a hangsúlyfrázisok által megadott prozódiai szerkezetbe jól illeszkedő útvonalak súlyát növelve, a nem illeszkedőkét pedig csökkentve a hipotézisgráf újrasúlyozható, ezzel a módszerrel pedig a beszédfelismerés eredményessége jelentősen növelhető. A hangsúlyfrázisok illesztését végző rendszer blokkvázlata a 9.15. ábrán látható.

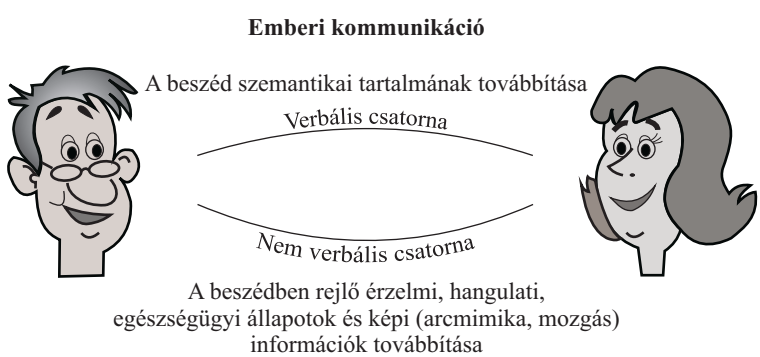


9.15. ábra. Hangsúlyfrázisok (HF) illesztését végző modul blokkvázlata

9.11. Érzelemfelismerés

Vicsi Klára

Az emberi beszédkommunikációban a beszédinformáció feldolgozása két egymástól elkülönült módon történik (9.16. ábra). Az egyik feldolgozási mód esetében az üzenet nyelvi tartalmát dolgozzuk fel (verbális csatorna); a másik információfeldolgozási mód (a nem verbális csatorna) ahol a beszélő általános érzelmi, egészségi állapotát, hangulatát érzékelik (Burkhardt et al. 2005). Az utóbbi évtizedben óriási erőfeszítések történtek a verbális csatorna működésének megértésére. A nem verbális csatorna jelentősége ez idáig kisebb volt, és működését kevésbé értjük. Az emberi beszéddel a



9.16. ábra. Az emberi kommunikáció két egymástól elkülönült csatornája

beszédtartalmon túl sok mást is ki lehet fejezni. Ezeket a különböző beszédformákkal (változatok) tudja a beszélő érzékeltetni. A hangszínezet, az intonáció, a ritmusváltozások mind széles körben használatosak arra, hogy a beszélő kognitív, érzelmi, vagy egészségi állapotát vagy a kommunikációban szerepet játszó nem verbális jelzéseket, mint az egyetértést, illetve egyet nem értést is egyidejűleg kifejezzék. Csak néhány éve növekedett meg a jelentősége a beszéd különböző, nem verbális nézőpont szerinti vizsgálatának. Ebben a fejezetben csak az érzelmekkel foglalkozunk. A beszédben előforduló kognitív állapotok, különösen az érzelmek vizsgálatának számos nehézsége van, melyeket az alábbiakban sorolunk fel.

Statisztikai feldolgozásra elegendő érzelmet kifejező beszédanyag gyűjtése nehéz. Az irodalomban található ugyan néhány kutatási leírás, amely a beszéd emóciótartalmának vizsgálatával és az emóció automatikus, gépi felismerésével foglalkozik, de ezek az eredmények mind laboratóriumi körülmények között elhangzó tiszta beszédre vonatkoznak (Campbell 2004, 2007a, Douglas-Cowie et al. 2003, Hozian–Kacic 2006). A publikációk legtöbbjében szimulált emóciótartalmú beszédet használnak, leggyakrabban művészek bemondásmintáit. Például magyar nyelvre professzionális és amatőr bemondókkal szimulált emóciótartalmú beszédadatbázist építettek (Fék et al. 2007)

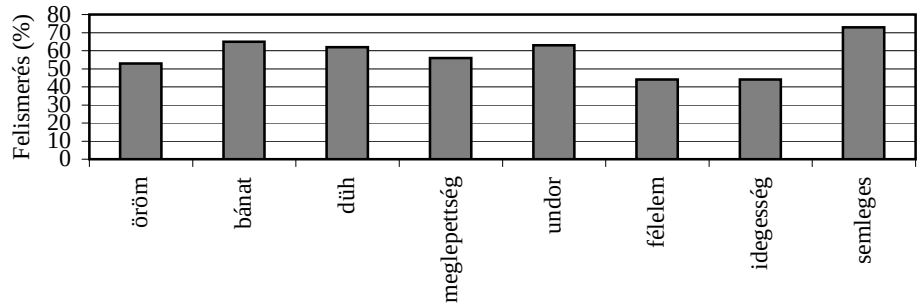
A valós szituációkban elhangzó, spontán beszédre jellemző adatok jelentősen különböznek a színészek által produkált beszédtől (Kostoulas et al. 2007). A természetes ember-gép kommunikációban pedig a valóságos spontán beszédet kell fogadni és feldolgozni. Az utóbbi években már megjelent néhány olyan publikáció, amely a spontán hétköznapi beszéd vizsgálatával (Navas et al. 2006) és információtartalmainak felismerésével (Kohavi 1995) foglalkozik. Problémát jelent továbbá a kognitív és az érzelmi kategóriák változatos megjelenése a beszédben. Az emóció jellemzésére a pszichológiában, nyelvészetben és audiovizuális jelfeldolgozásban, hagyományos emóciókategóriákat használnak, úgymint boldogság, szomorúság, düh, meglepetés, undor. Eredetileg az MPEG-4 szabványban (IEC 1999) e kategóriákat az arcmimi-

ka jellemzésére szolgáló virtuális paraméterek (facial animation parameters, FAPs) megjelenítésére használták. A beszédtechnológiai szakemberek ezeket a kategóriákat vették át a beszédben rejlő érzelem vizsgálatára is. Ha ezt összevetjük a valós helyzettel, az látszik, hogy a spontán beszédben sokkal változatosabb a kognitív kategóriák tárháza, és ezek a téma szerint erősen változhatnak is. Kutatási céllal a spontán beszédben leggyakrabban előforduló érzelmi kategóriákat gyűjtöttek ki a PHYSTA 2001 adatbázisból (Cowie–Douglas-Cowie 2001). Ez az adatbázis spontán társalgást, televíziós beszélgetőműsorok és különböző vallási műsorok gyűjteményét tartalmazza (298 egység, 1 egység 10–60 s hosszú). A kiválasztott leggyakoribb kognitív állapotok és azok gyakorisága a 9.2. táblázatban látható. Ezek egy része megfelel a hagyományos értelemben vett érzelem kategóriáknak (például dühös), mások azonban nem (például magabiztos).

9.2. táblázat. Kognitív állapotok csoportosítása és gyakoriságuk a PHYSTA 2001 spontán audiovizuális adatbázisban

Címke	Használati gyakoriság	Csoport
Semleges	273	Nem erősen érzelemvezérelt
Dühös	114	Erősen negatív
Szomorú	94	Erősen negatív
Örvendező	44	Nem orientáltan pozitív
Boldog	37	Nem orientáltan pozitív
Jó kedélyű	26	Nem orientáltan pozitív
Aggódó	19	Erősen negatív
Csalódott	17	Nem erősen érzelemvezérelt
Izgatott	17	Orientáltan pozitív
Félelem	13	Erősen negatív
Magabiztos	13	Nem erősen érzelemvezérelt
Érdeklődő	12	Nem erősen érzelemvezérelt
Gyengéd	10	Orientáltan pozitív
Elégedett	4	Nem erősen érzelemvezérelt
Szeretetteljes	3	Orientáltan pozitív

További problémát jelent a beszédben kifejezésre kerülő érzelmek vizsgálatánál, hogy a szemantikus tartalom (verbális csatorna) és a beszélő hangulatának, általános érzelmi állapotának a tükröződése (nem verbális csatorna) egyazon beszédfolyamatban valósul meg, és a szemantikus tartalom hozzájárul a beszéd emóciótartalmának a felismeréséhez is. Amennyiben kiküszöböljük a nyelvi tartalom hatását, akkor az emberi emóciófelismerés sem jobb, mint 60–65% (Tóth et al. 2007). Vicsi és munkatársai ugyanazt a mondatot rögzítették különböző érzelmekkel színészekkel és átlagemberekkel (3 mondat, mondatonként 8 érzelem, 15 személlyel). Ezeket a mondatokat meghallgattatták érzelem szerinti megítélésre 20 személlyel. A színészek és átlagemberek produkcióinak lehallgatási eredményei között szignifikáns eltérés nem volt. A nem színészek által bemondott mondatok szubjektív lehallgatásának eredményeit a 9.17. ábra mutatja.



9.17. ábra. Az átlagemberek bemondásainak érzelmek szerinti felismerési eredményei percepciós teszt alapján

A helyzetet tovább bonyolítja, hogy az érzelmeinket a kommunikáció során több érzékszervi csatornán keresztül juttatjuk el a másik félhez, e csatornák közül a legjelentősebb a beszédhang maga és az arcmimika, de még a testbeszéd, bőrpír és egyéb tényezők is szerepet játszhatnak a kognitív állapot és az érzelem kifejezésében. Agyunk az összes érzékszervi csatornán keresztül kapott információ együtteséről dönt (Hozian–Kacic 2003). Például egyes esetekben hallás után az ember maga sem tud különbséget tenni két érzelem között, de látva az arckifejezést, már könnyebben dönt. Az is megfigyelhető, hogy az ember érzelemfelismerési képessége csupán az arckifejezést látva meglepően jó. Az, hogy a hang vagy pedig a kép ad több információt az emberi érzelem felismeréséhez, attól függ, hogy a hangban a nyelvi tartalom is benne van-e, vagy nincs. Amennyiben a hang nyelvi tartalmat is ad, akkor annak alapján lényegesen jobb a felismerés, mint csak az arckifejezés alapján. Ha viszont a hang nyelvi tartalmat nem ad, például idegen nyelv esetén, akkor az arckifejezés alapján lesz jobb felismerés (Esposito 2009). A hang- és képinformációt kombinálva javul a legjobban a felismerés minősége, eddig az automatikus felismerésben a kutatóknak megközelítőleg 80% körüli felismerést sikerült elérniük a kombinált információ felhasználásával (Douglas-Cowie et al. 2003). Továbbiakban célunk csak a hang alapján történő érzelemfelismerés tárgyalása. A fent felsorolt problémák talán magyarázatul szolgálnak arra, hogy az eddig végzett kutatások, kizárólag hang alapján, 60% körüli gépi felismerést értek el legjobb esetben is (Burkhardt et al. 2005, Campbell 2007b, Hozian–Kacic 2003, Tóth et al. 2007).

Beszédérzelmeket tükröző jellemzővektorok. A gépi érzelemfelismerés során a meglévő hanganyagból jellemző vektorokat nyerünk ki, és ezeket használjuk fel az automatikus felismerő tanításához, majd ezekkel hajtjuk végre a felismerést. Ehhez persze tudni kell, hogy mik azok a jellemzők, amelyek jól leírják az emberi beszéd érzelmi tartalmát. Tehát először a beszédérzelem jellemzőit kell definiálni, kategorizálni.

Amint azt már az előző fejezetekből tudjuk, a beszéd semleges érzelem kifejezésekor is rendkívül változatos, két különböző személy ugyanazt a mondatot másképp ejti ki, továbbá ugyanazt a mondatot, ugyanaz a személy sem ejti kétszer ugyanúgy (9.2. fejezet). A kiejtett hangok fizikai paraméterei függhetnek a beszélő egészségi, fizikai állapotától is (megfázás, stressz, fáradtság, légúti betegségek). Mindezekhez hozzájárul még az a tény, hogy a beszélő a szándékától, érzelmi állapotától függően is változtathat egy mondat hangzásán, ezzel is kifejezve érzelmi állapotát. A beszédhang fizikai jellemzői tehát ugyanannál a szemantikai tartalomnál is sokfélék lehetnek. Ez megnehezíti az érzelem gépi felismerését, hiszen meg kell tudnunk mondani, hogy mely változások játszanak fontos szerepet az érzelem kifejezésben, és melyek nem. A mai napig az idevonatkozó szakirodalom egyik fő kérdése, hogy az automatikus érzelemfelismeréshez milyen jellemzőket kell kigyűjteni, amelyek alapján majd a felismerés működni fog. A következőkben az irodalomban fellelhető, az érzelmek jellemző fizikai paramétereivel foglalkozó idézett cikkek összefoglalását adjuk meg. (Hozian–Kacic 2003, Álvarez et al. 2007, Seppänen et al. 2003).

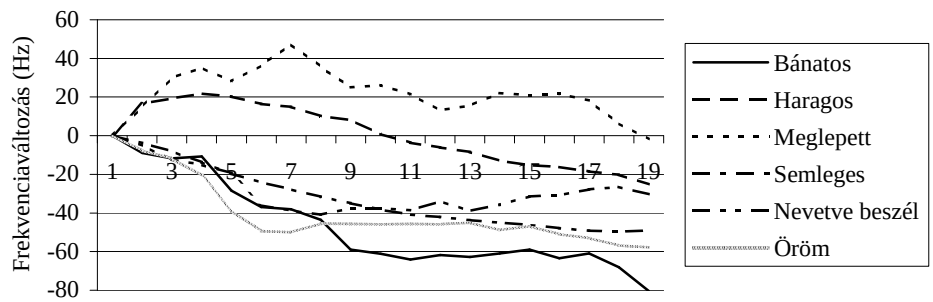
Alapszintű adatok a jellemzővektorokban. Az úgynevezett alapszintű jellemzők közé tartoznak a keretenkénti alapfrekvencia-értékek, a hangintenzitás-értékek, valamint a beszédhangok időtartama. Az alapfrekvencia erősen beszélőfüggő, személyenként és időben változó érték. Mégis az irodalomban érzelmet tükröző alapszintű jellemzőnek tekintik. A beszédhangok intenzitása és annak deriváltja is fontos paraméter, kifejezi a nyomatékokat, azaz hangsúlyokat. A témával foglalkozó cikkek ezt a paramétert mind besorolják a vizsgálandók közé. A harmadik alacsony szintű jellemző a beszédhangok, szótagok időtartama. Ezek meghatározzák a beszéd tempóját, ritmusváltásait.

Származtatott adatok a jellemzővektorokban. A származtatott jellemzőket az alapszintűekből képezzük, azok valamilyen változását, statisztikáját tekintve, melyet jellemzően a mondatnyi hosszúságú beszédre számítanak ki. A cikkek szerint ezek a származtatott jellemzők meghatározzák az egyén beszédének prozódiai jegyeit. Információt hordoznak az intonációról, a tempóról és a hangerőről. Ilyen származtatott jellemzők az alapfrekvencia- és az intenzitás maximuma, minimuma, átlagértéke, deriváltja, értéktartománya egy hosszabb közlésre, például a mondatra. Újabban már a spektrális jellemzőket, például a mel-skála együtthatóit (MFCC együtthatók) is besorolják az érzelmek jellemző paraméterei közé (Tóth et al. 2007). A származtatott jellemzők, amelyeket az irodalomban mondategységekre számítottak ki, folyamatos spontán beszédben nem igazán vezettek eredményre, mivel a hosszabb összetett mondat szerkezete függvényében a mondat más-más részében jelenik meg az érzelem kifejezése. Éppen ezért, a legújabb kutatások szerint (Vicsi–Sztahó 2009) az érzelem kifejezésének alapegységeként a frázist tekintjük. A frázis két levegővétel közötti szakasznak felel meg. Amennyiben frázisonként vizsgáljuk az érzelmek kifejezését, akkor nagyobb részben már ki tudjuk küszöbölni a mondat szerkezetétől

való függést, ugyanakkor a frázis már elég hosszú beszédegység ahhoz, hogy érzelmet tükrözhessen.

A kérdés tehát az, hogy milyen fizikai paraméterek és azok milyen kombinációi tükrözik az egyes érzelmeket a frázisokban? Egy kísérletben 43 beszélő spontán beszédből származó 1000 frázisát vizsgálták (Vicsi–Sztahó 2010). Öt különböző érzelemi kategóriát próbáltak elkülöníteni: bánatos, haragos-ideges, meglepett, nevetve beszélő, örömöt kifejező. Az volt a tapasztalat, hogy az alapfrekvencia és az intenzitás változása egy frázison belül jellemző.

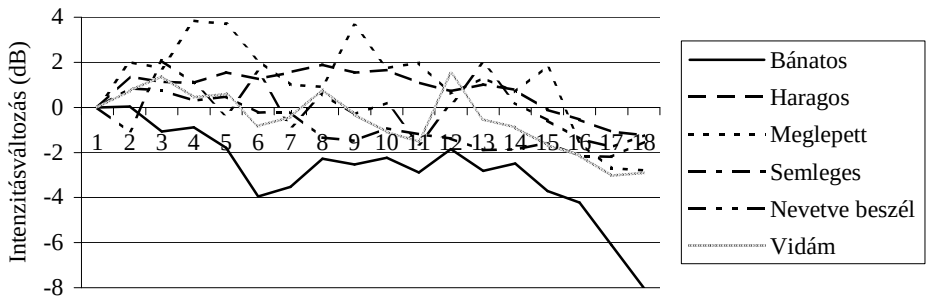
A vizsgálati anyagban a különböző hosszúságú frázisokat *n* egyenlő számú részre osztották. Ezzel lineárisan vetemítést hajtottak végre az időtengelyen. Minden részben végeztek mérést, az adatok normálták a frázisban mért első átlagadat értékére úgy, hogy a mintavételezési pontoknál mért adatokból az első minta értéke levonásra került. Végül átlagolták az érzelem szerinti csoportok frázisonkénti értékeit, vagyis minden érzelemre elkészült az adott érzelemre jellemző, úgynevezett átlagos hangminta-dinamika.



9.18. ábra. A különböző érzelmek átlagos alapfrekvencia-dinamikája. A vízszintes tengelyen a mintavételezési pontok láthatók

Alapfrekvencia dinamikája (n=19): Az alapfrekvencia dinamikája n = 19 esetén a 9.18. ábrán látható, ahol az alapfrekvencia szórásértékei 5–10 Hz között adódtak. Az alapfrekvencia-dinamika érzelem szerint szépen elkülönül az alábbiak szerint.
Bánatos. Alapfrekvencia folyamatosan és erősen csökken. Majd körülbelül a frázis felénél, 60 Hz-es csökkenés után egy stagnálást, majd a végén újabb csökkenést.
Haragos. Az elején nő az alapfrekvencia, majd folyamatosan csökken.
Meglepett. Az elején nagymértékű alapfrekvencia-növekedés látható, majd valamelyes csökkenés. Ennél az érzelemkategóriánál figyelhető meg leginkább az alapfrekvencia növekedése.
Semleges. Az alapfrekvencia folyamatos szabályos csökkenése figyelhető meg, bár annak mértéke nem igazán jelentős.
Nevetve beszél. Az alapfrekvencia csökkenése, majd körülbelül a frázis felétől lassú növekedése jellemzi.

Öröm. Az elején az alapfrekvencia lényeges (körülbelül 50 Hz) csökkenése figyelhető meg, majd utána stagnál, hasonlóan a nevetve beszél kategóriához. Tehát az alapfrekvencia dinamikája jól jellemzi az érzelmeket.
Az intenzitás dinamikája (n=19): alapvetően az egyes érzelemkategóriák intenzitásának dinamikái nem különülnek el olyan szépen, mint az alapfrekvenciaváltozásának esetében, amint ez a 9.19. ábra alapján látható. Itt az értékek nem az első mintavételezési helytől kerültek ábrázolásra, hanem a másodiktól, emiatt az utolsó mintavételezési hely sorszáma a 18-as. A szórásértékek körülbelül 3 dB értékűek



9.19. ábra. A különböző érzelmek átlagos intenzitásdinamikája. A vízszintes tengelyen a mintavételezési pontok láthatók

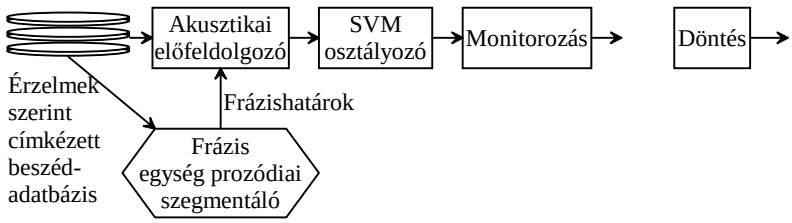
voltak. Ez itt relatíve magas érték. Amit érdemes megfigyelni, az az, hogy a bánatos érzelemnél jól látható és a többi érzelemtől elkülönült az intenzitás csökkenése, stagnálása, majd újabb csökkenése, illetve a haragos érzelemnél az intenzitás növekedése körülbelül a frázis feléig. A semleges érzelemnél az elején kicsi növekedés figyelhető meg, majd az érték folyamatos csökkenése. A nevetve beszél és a vidám érzelmeknél, az intenzitás folyamatos változása figyelhető meg. Az intenzitásértékek kevésbé tükrözik a különböző érzelmeket, bár azért jellemző dinamikajegyek az intenzitásnál is fellelhetők. Érzelmekre jellemző lényeges spektrális változás az idő függvényében a frázison belül nem tapasztalható, ugyanakkor egy frázisra átlagolt spektrális paraméterek már érzelemre jellemző eltéréseket mutatnak.

Összefoglalva, a 43 beszélő 6 különböző spontán beszédben felvett érzelmi kategóriáinak statisztikai vizsgálata alapján elmondható, hogy az alapfrekvencia és az intenzitás frázison belüli időbeli változása, valamint egy frázis egészére átlagolt spektrális paraméterek együttesen jellemzik a különböző érzelmeket. Az, hogy meg tudjuk mondani, melyik paraméter mikor és milyen súllyal járul hozzá a komplex érzelmi jellemzés kialakításához, még további kutatást igényel.

A beszédérzelem gépi felismerésének kutatási eredményei. Ma a beszédérzelem gépi felismerésére kidolgozott megoldások lényegében mind statisztikai osztályozáson alapulnak, ezért az ilyen vizsgálatokhoz nagy méretű tanító-adatbázisra van szükség.

Az adatbázis kiválasztása döntő fontosságú, illeszkednie kell az adott alkalmazáshoz. Lényegében a statisztikus osztályozó eljárások sikere azon múlik, hogy az adott feladathoz illeszkedően milyen jó hanganyagot gyűjtünk össze, és az összegyűjtött hanganyagot hogyan, milyen részletességgel készítettük elő, dolgoztuk fel. A mai rendszerek legtöbbjében az akusztikai előfeldolgozás után, a betanítás nagy mennyiségű érzelem szerint szegmentált és címkézett adatbázissal történik. Erre a szakirodalomban többféle statisztikai osztályozó eljárást alkalmaznak. Széles körben elterjedt a rejtett Markov-modell (HMM) használata, de ezenkívül alkalmazzák a neurális hálók, az MQDF (Modified Quadratic Discriminant Function), az MMD (Modified Mahalanobis Distance), vagy az SVM (Support Vector Machines) eljárásokat is. Ezek az eljárások körülbelül hasonló eredményeket tudnak produkálni, bár bonyolultságuk eltérő lehet. Az eddigi legjobb felismerési eredmények, amelyek az alapfrekvencia, az intenzitás és az időtartam származtatott jellemzőit használták, 60 százalék körül vannak (Burkhardt et al. 2005). Ez alacsonyabb, mint az emberi felismerés foka. Oka az lehet, hogy az alapfrekvencián és intenzitáson kívül más tényezők is hozzájárulnak az érzelem kifejezéséhez, mégpedig a spektrális szerkezet. A csak alaphang és intenzitás alapján történő felismerés elégtelenségét jól mutatja, hogy amikor a semleges szintetizált beszédet csak alaphang és intenzitás szerint módosították, és az így előállított érzelmeket lehallgattatták, az érzelem szerinti felismerés csak 30 és 40 százalék körüli volt (Fék et al. 2005). Az automatikus érzelemfelismerés bemeneti jellemzői közé a spektrális jellemzőket is bekapcsolva a felismerés lényegesen javul (Tóth et al. 2007).

Továbbá általános tapasztalat szerint a beszélőfüggő és beszélőfüggetlen megvalósítások között óriási szakadék tátong teljesítmény szempontjából. Egy valós, telefonos ügyfélszolgálati dialógusban az ügyfél érzelmi állapotát kívánták automatikusan detektálni a telefonbeszélgetés során (Vicsi–Sztahó 2009). Mivel követni akarták a beszélő érzelmi változásait, szegmensekre kellett felosztaniuk a beszédfolyamot, hogy meg tudják vizsgálni, hogyan változik szegmensről szegmensre a beszélő érzelmi állapota a beszélgetés alatt. A rendszerben a frázist választották szegmentálási egységként (lásd még a 9.10. fejezetben). A frázis méretű egységek szegmentálásakor az egységekre való osztást prozódiai szegmentálóval végezték (Vicsi–Szaszák 2008). Az akusztikai előfeldolgozás után a frázis méretű szegmenseket azok érzelmi töltete szerint osztályozták, SVM (support Vector Machine) gépi osztályozó felhasználásával. A rendszer folyamatábráját a 9.20. ábra szemlélteti. Az akusztikai előfeldolgozáskor a keretenkénti alapfrekvenciát ($F0_i$), intenzitásértékeket (E_i), 12 MFCC együtthatót és deriváltjaikat mérték, 150 ms-os időablakot használva 10 ms-os időkeretekben, összesen 28 elemű tulajdonságvektorral 10 ms-onként. Ezután a frázis prozódiai szegmentáló kijelöli a frázishatárokat a beszédben, frázisok sorozatát hozva ezzel létre. A 10 ms-onkénti tulajdonságvektorok alapján minden egyes frázist egy multidimenzionális statisztikai tulajdonságvektor jellemez. Ezeket a statisztikai tulajdonságvektorokat a következők szerint számították ki: az $F0_i$ értekeit az első



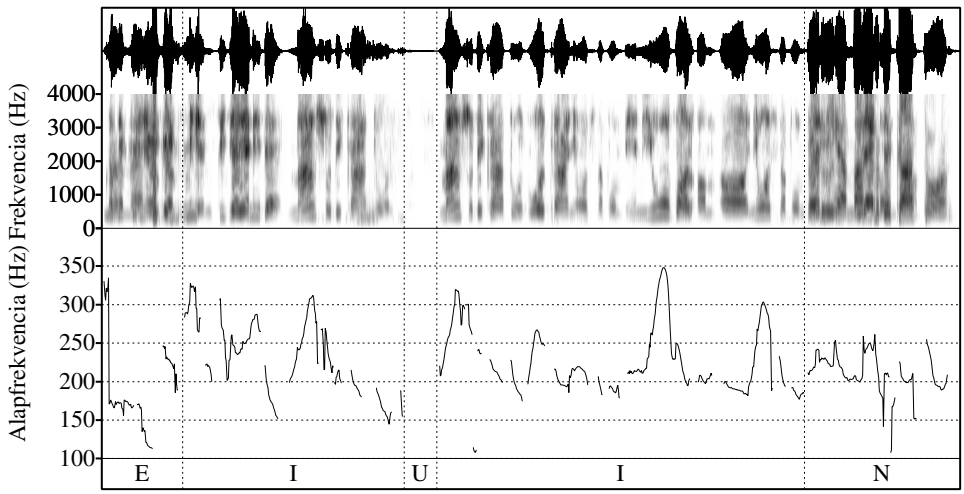
9.20. ábra. Beszédérzelem-osztályozó blokkvázlata

időkeret F0 értékei, az E_i értékeket pedig az E maximumérték szerint normalizálták minden egyes frázis esetében. Majd e normalizált paraméterekből számították ki a következő statisztikai adatokat minden egyes frázisnál: A frázisonként számolt mul-

9.3. táblázat. Statisztikai tulajdonságvektorokba besorolt paraméterek

Fizikai jellemző	A jellemzőből származtatott paraméterek
$F0_i$	Maximum, minimum, közép, medián
$\Delta F0_i$	Maximum, minimum, közép, torzulás (skew)
E_i	Közép, medián
ΔE_i	Maximum, minimum, közép, torzulás (skew)
$MFCC_i$	Maximum, minimum, közép
$\Delta MFCC_i$	Maximum, minimum, közép

tidimenzionális statisztikai tulajdonságvektor kerül az SVM osztályozó bemenetére, amely minden egyes frázist besorol egy-egy érzelmi osztályba. A betanításhoz használt (a diszpécserek és az ügyfél között rögzített párbeszédekből kialakított) adatbázisban négy különböző érzelmi állapotot különböztettek meg: semleges (N), ideges (I), panaszkodó (P) és egyéb (E) kategóriákat a 9.21. ábrán bemutatott példa szerint. Az ezzel az adatbázissal készült betanítás után az osztályozás eredménye érzelemtől függően 44 és 66% között ingadozott. Az I és P érzelmeket nemcsak az osztályozó, de az adatbázist feldolgozó emberek is alig tudták differenciálni. Így az I, P és E osztályok egy osztályba kerültek, mint elégedetlenséget kifejező érzelmek. Tehát végül az „elégedetlen” osztályt és a semleges érzelmek osztályát különböztették meg, és így tanították be az SVM osztályozót. Frázisonként változhat, ugrálhat a megítélt érzelem. Biztos döntés az ügyfél állapotáról akkor hozható, ha több frázison keresztül többségében egy típusú érzelem fordul elő. Ehhez 15 másodperc hosszúságú időablakot választottak, és mérték az ablakon belül az „elégedetlen”-nek osztályozott frázisok számát. Ez a szám %-ban kifejezve adta meg az „elégedetlenség” mértékét. (Az elégedetlenség akkor volt 100%-os, amikor a monitorozó ablakban az összes frázis elégedetlennek lett minősítve.) Azután az ablakot továbbmozgatták, 10 másodperc időlépéssel. Néhány példát mutatunk a 9.22. ábrán arról, hogyan változik meg az ablakban mért szám, vagyis az elégedetlenség mértéke a beszélgetés során. Egy-egy beszélgetésben az automatikusan nyert eredményeket hasonlítottuk össze a kézzel



9.21. ábra. Példa a betanításhoz használt adatbázis szegmentálására és címkézésére. U: szünet, N: semleges, I: ideges és E: egyéb

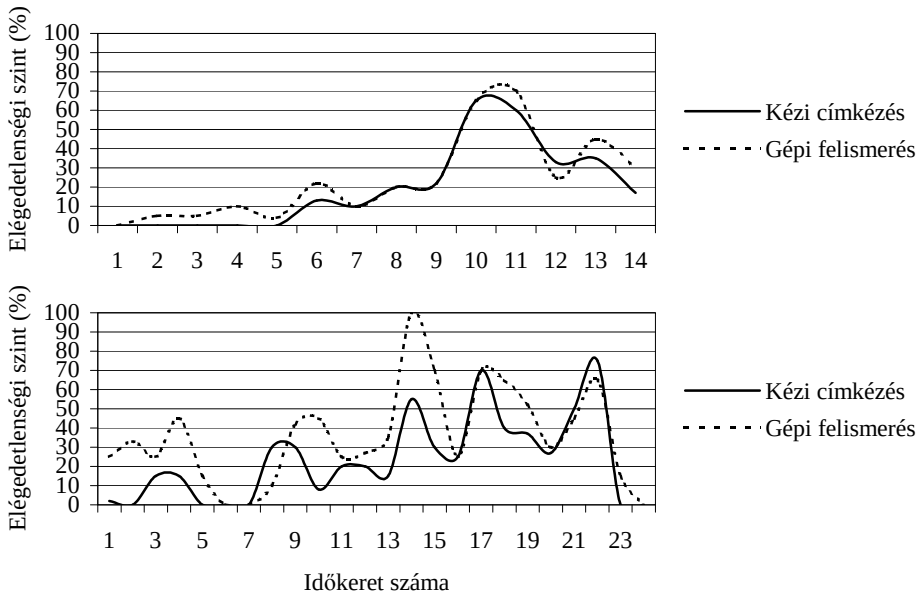
felcímkézett eredményekkel. Az automatikusan nyert eredmények ezzel a módszerrel már csak átlagosan 11,3%-ban tértek el a kézzel felcímkézett eredményektől. Általánosságban elmondható, hogy a beszédtechnológiai feladatokban a spontán beszédérzelem felismerésére van igény, másfelől pedig gyakran nem az alapérzelmek felismerése a feladat, hanem valamilyen kombinált érzelemváltozaté vagy általánosabb kognitív állapoté.

Összegezve, az érzelem automatikus felismeréséről azt kijelenthetjük, hogy spontán beszédnél ma még aligha várható sokkal jobb eredmény, mint a fenti megoldás. Világos, hogy a felismerési eredmény lényegesen jobb lehet, ha a verbális csatorna információja, vagyis a nyelvi tartalom is a rendszerhez integrálódik. A beszéd érzelmi tartalma a nyelvi tartalom szerves része, felismerésbe történő bevonásával a felismerés még biztosabbá tehető. A jövő útja mindenképpen az, hogy a beszédben lévő verbális és nem verbális jegyek együttesen kerülnek feldolgozásra és értékelésre. Így az átadandó üzenet komplex kezelésével biztosabban felismerhető az üzenet teljes nyelvi és nem nyelvi tartalma.

9.12. Beszédfelismerés támogatása multimodális paraméterekkel

Czap László

Zajos környezetben, változó csatornaparaméterek és beszédstílus mellett a beszédfelismerők megbízhatósága jelentősen romlik, meg sem közelítik az emberi beszéd-



9.22. ábra. Az ügyfél elégedetlenségének mértéke egy-egy beszélgetés során. (Az elégedetlenség akkor volt 100%-os, amikor a monitorozó ablakban az összes frázis elégedetlennek lett minősítve)

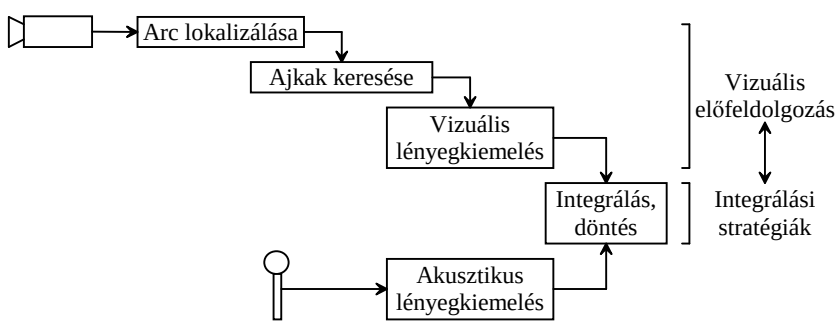
értés alkalmazkodó képességét (Hermansky–Morgan 1994). A beszéd vizuális modalitása az egyik ígéretes kiegészítő információforrás, amely mentes az akusztikai környezet és a zaj zavaró hatásaitól. A multimodális emberi kommunikációban az akusztikai és vizuális jelet zseniálisan kombináljuk a maximális érthetőség érdekében. Még a gépi kétmódusú beszédfelismerés megszületése előtt, az ötvenes években demonstrálták, hogy zajos környezetben a vizuális jel segíti a beszéd jobb megértését (Sumby–Pollack 1954). A vizuális modalitás előnyei az emberi beszédfelismerésben elsősorban három területen mutatkoznak meg: segíti a hangforrás, a beszélő helyének meghatározását, megkönnyíti az akusztikai jel szegmentálását, kiegészítő információval szolgál az artikuláció helyének meghatározásához (Summerfield 1987). Massaro–Stork (1998a) kísérletekkel igazolta, hogy a modalitásokat egymás kiegészítésére használjuk. Ha a hang gyenge minőségű, vagy hallássérült a megfigyelő, jobban hagyatkozik a szájról olvasásra. *Jobban hallom a televíziót, ha felteszem a szemüvegem.* Az emberi beszédértést meg sem közelítő gépi felismerőket hasonlíthatjuk a környezet, vagy képességei által korlátozott emberi felfogóhoz abban a tekintetben, hogy a kiegészítő vizuális jel a gépi beszédfelismerők felismerési hatékonyságát is javíthatja, különösen zajos környezetben. A szájról olvasás régről ismert technikájának a gépi beszédfelismerés szolgálatába állítása mintegy két évtizeddel ezelőtt kezdődött. A számítógépek sebessége és tárkapacitása, a képfelvevő

és képmegjelenítő eszközök fejlődése lehetővé tette a képfeldolgozás eredményeinek szélesebb körű alkalmazását.

9.12.1. A vizuális lényegkiemelés

A kétmódusú beszédfelismerés egyik kulcskérdése a vizuális lényegkiemelés. A kutatások egyik jelentős területe a vizuális információt hordozó jellemzők meghatározása. Az akusztikai jel leírására kipróbált módszerek állnak rendelkezésre, a vizuális jellemzők kiválasztása és ezek leolvasása a képről azonban még kevésbé kidolgozott, tehát a másik lényeges kérdés, hogy a választott jellemzők hogyan nyerhetők ki a képből. Ehhez szükség van az arc lokalizálására, a száj körüli terület kijelölésére és a jellemzők meghatározására (9.23. ábra).

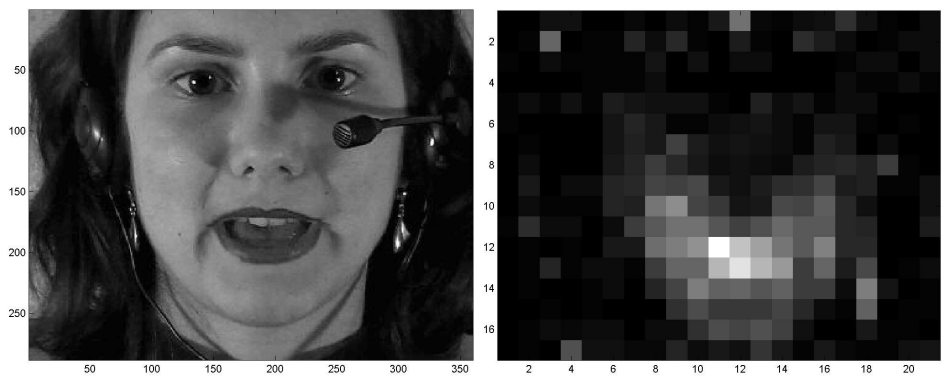
Integrálási stratégiák. A lényegkiemelés célja a több tíz megabit per szekundumos sebességű videojel adatmennyiségének radikális csökkentése és az artikuláció szempontjából releváns, a vizuális beszédet leíró jellemzők kinyerése. A kétmódusú beszédfelismerés kezdetén a videojel feldolgozását a beszélő arcán színes „szépségflastrommal" megjelölt nevezetes pontok (például az állcsúcs és a szájsarkak) azonosítása és néhány geometriai jellemző mérése jelentette. A vizuális beszédfelismerés egyik atyja, Petajan (1984) küszöbdetekcióval határozta meg az ajkak szélességét, nyitását és területét. Úttörő munkájának rövid idő alatt számos követője akadt, és kialakultak az audiovizuális beszédfeldolgozás kutatóműhelyei. A lényegkiemelés és az integrálás különböző módszereit dolgozták ki, amelyek azonban nehezen hasonlíthatók össze, mivel más-más adatbázison dolgoztak. Nemzetközi projektek keretében létrehoztak audiovizuális adatbázisokat, de a vizuális lényegkiemelésre még nem alakult ki követendő módszer. Kedvelt egyszerűsítés a jól azonosítható markerjelek



9.23. ábra. Az audiovizuális beszédfelismerés sémája

(Benoit et al. 1996) arcra festése. Például a kék rúzs (Dalton et al. 1996) színméréssel egyszerűen elkülöníthető, hiszen ez a színösszetevő egyébként jellemzően nem

fordul elő a bőr színében. Ma is elsősorban kutatási célú elemzés folyik, a beszélő által tolerálható előkészítés elfogadható. Régebben nem volt ritka a beszélő fejének fix pozícióba rögzítése sem, ma enyhe természetes mozgást kezelni lehet. A technikai háttér és az eljárások fejlődése ma már lehetővé teszi a kötöttségek enyhítését. A következő feladat az artikuláció szempontjából fontos terület (region of interest) kijelölése. Beszéd közben az arc legintenzívebb mozgást végző területe a száj és az áll környezete. Képsorozat vizsgálata alapján mozgásbecsléssel megállapíthatjuk, hogy egy képrészlet mennyivel tolódott el az előző képhez képest. Az így kapott mozgásvektorok összegzéséből megállapítható a bemondó képein a legintenzívebb mozgást mutató terület. A mozgásbecslés a mozgóképek tömörítésére szolgáló eljárások egyik szokásos lépése. Az előző képkockán szereplő képrészletek eltolásával próbálja összerakni a képet. A mozgásvektorok azt fejezik ki, hogy a például 16x16 pixeles képrészlet az előző képből milyen irányú és nagyságú eltolással állítható elő. A 9.24. ábrán 8 félképből (160 ms) előállított mozgásvektorok abszolút értékének összegét, ennek intenzitásképét látjuk a beszélő képe mellett. A legintenzívebb mozgás a száj területén tapasztalható (Czap 2004). Másik lehetőség a vizsgálandó terület



9.24. ábra. A beszélő képe és a mozgásvektorok abszolút értékének összege

kijelölésére az arc jellegzetes helyeinek azonosítása alapján megtalálni a feldolgozandó képrészletet. Olyan alakzatot kell keresni, amely nagy megbízhatósággal azonosítható a képen. Erre a szemek és az orrhegy-orrlyuk területe a legalkalmasabb (Iyengar et al. 2001). A szem szemüvegeseknél és pislogás idején nehezen található meg. A képkockánkénti feldolgozás, a választott jellemzők leírása, diszkrét idejű változókat eredményez. (A váltottsoros letapogatással kapott képkocka páros és páratlan sorainak szétválasztásával két félkép nyerhető, ezzel a 40 ms-os képidővel 20 ms lesz a félképek periódusideje.) Rendkívül fontos a hang és kép közötti szinkron, valamint többkamerás (például elöl- és oldalnézet) felvételek esetében a vizuális jelek egymás közötti szinkronizálása. McGrath–Summerfield (1985) kutatásai szerint az audiovizuális beszédfelismeréshez a hang és a kép 40 ms-nál kisebb

elcsúszása még nem okoz veszteséget, ez PAL rendszerű felvételnél egy képidőnek felel meg. A lényegkiemelés egyik fő kérdése, hogy milyen jellemzők hordozzák a beszédfelismerés szempontjából lényeges vizuális információt. A másik fontos kérdés, hogy a kiválasztott vizuális jellemzők hogyan nyerhetők ki a képből. A magas szintű, vagy geometriai alapú feldolgozás ismert eljárásainak általános jellemzője, hogy az ajkak külső és belső kontúrjának követésével az artikulációs szervek látható részeinek méretét vagy helyzetét olvassák le. A geometriai alapú felismerésnél az artikuláció különböző paramétereivel próbáljuk leírni a vizuális beszédjel változásait. A leggyakoribb vizuális jellemzők:

- az ajkak szélessége, vagyis az ajkak belső kontúrjának vízszintes mérete
- az ajaknyílás, az ajkak belső függőleges mérete
- az előbbi jellemzők az ajkak külső kontúrjára vonatkozóan
- az ajkak belső kontúrja által bezárt terület
- az ajkak külső vonalával határolt terület
- az áll mozgása, például az orrhegyhez képest
- oldalnézeti képen az ajkak előre mozgása
- a nyelv és a fogak láthatóságának leírására általánosan alkalmazott jellemző még nem alakult ki

A felsorolt paraméterek erősen redundánsak, szerepüket több kutató vizsgálta (Hennecke et al. 1996, Lavagetto–Lavagetto 1996). Az egymással összefüggő jellemzők közül például a szájnyílás belső méretei meghatározzák a területet, így az nem jelent új információt. Az ajkak szélességének és nyitásának aránya eltérő ajakkerekítéses és ajakréses hangoknál. Ajakréses hangoknál a külső kontúr méretei, területe csak kevéssel nagyobb a belső méreteknél, ajakkerekítéses hangoknál a külső méretek számottevően nagyobbak. Az ajkak előre mozgása az ajakkerekítéses hangokra jellemző. Az áll mozgása összefüggésben áll a száj nyitottságával és a fogak láthatóságával, hiszen az alsó fogsor az állal együtt mozog. Redundáns jellemzők alkalmazása robusztusabb felismeréshez, vagy ellentmondó jellemzők esetén az eredmények romlásához vezet. A vizuális jellemzőknek az akusztikai jelet kell erősíteniük, a két paraméter egymásra hatása módosíthatja egyes jellemzők hasznosságát. A vizuális lényegkiemelés másik iskolája a pixelbázisú megközelítés. Az alacsony szintű vagy pixelalapú feldolgozás a száj környezetének képpontjait veti valamilyen kétdimenziós (Fourier-, diszkrét koszinusz) transzformáció alá, és a transzformált jellemzők egy redukált készletét használja a felismeréshez például főkomponens-analízis után. Ebben az esetben a száj és környezete, de akár az egész kép minden pontja részt vehet az elemzésben. A pixelbázisú feldolgozás előnye, hogy az artikulációs terület feldolgozásával a képi információt veszteség nélkül a felismerés szolgálatába állíthatja. A teljes kép feldolgozása esetén a gesztusok figyelembevételére is lehetőség nyílik. Ma a pixelbázisú jellemzők hatékonyabb felismerést tesznek lehetővé (Matthews et al. 2001). A geometriai alapú feldolgozás segíti az artikuláció elemzését, a hangkép-

zés statikus és dinamikus jellemzőinek mérését. A pixelalapú feldolgozás ezeket a lehetőségeket nem kínálja. További hátránya a pixelbázisú feldolgozásnak, hogy érzékenyebb a megvilágítás változásaira és személyfüggő felismeréshez használható. A pixelbázisú feldolgozás hátrányaival kapcsolatban meg kell említeni, hogy a megvilágítás változását a geometriai bázisú rendszerek is csak kismértékben tolerálják. Ha a beszélőt szemből nem éri fény, a szájnyílása mindig sötét. Ha azonban éppen a kamera irányából világítjuk meg, a nyelv a hátul képzett hangoknál is látható lesz, a kétdimenziós képen a mélység nem érzékelhető. A vizuális lényegkiemelés eredményeképpen két paraméterrendszert kapunk, egyet az akusztikai, egyet a vizuális jel leírására. A legnagyobb kihívás a két modalitás integrálása a legjobb felismerés érdekében, az ember ugyanis a legkülönbözőbb minőségű modalitásokat mindig úgy integrálja, hogy a kétmódusú beszédfelismerés felülmúlja mindkét modalitás külön mért eredményeit (Potamianos et al. 2004). A két modalitás integrálása a vizuális lényegkiemelésnél is kevésbé kidolgozott terület.

9.12.2. A vizuális és akusztikai modalitás integrálása

Az audiovizuális beszédfelismerés az akusztikai beszédfelismerésből fejlődött ki, annak eredményeire épít. Kézenfekvő megoldásnak tűnt, hogy az akusztikai és vizuális jellemzőket összekapcsolva a bevált rejtett Markov-modell vagy neurális hálózat alapú felismerőt a megnövelt dimenziójú paraméterekkel tanítjuk és teszteljük. Ezt korai integrálási modellnek nevezték el (Hennecke et al. 1996). A másik szélsőséges lehetőség, hogy külön-külön döntést hozunk az akusztikai és vizuális modalitás alapján, és a két eredményt utólag egyesítjük (kései integrálás). A korai integrálás esetében az akusztikai és vizuális jellemzők egyesítése megelőzi a felismerés alapegységeire való osztályozást. A kései integrálásnál a két modalitás alapján külön-külön elvégzett osztályozás után történik az egyesítés (Benoit et al. 1998). Próbálkoztak közbenső megoldásokkal is, amikor a feldolgozás valamely fázisában figyelembe veszik a másik modalitást (Massaro–Stork 1998a, Benoit et al. 1996). Egyes eljárások lehetővé teszik a két modalitás megbízhatóságának figyelembevételét, és például a jel/zaj viszony becslése alapján súlyozhatjuk az akusztikai és vizuális jeleket (Glotin et al. 2001, Potamianos–Neti 2000). A Bayes-féle döntési eljárás alkalmas két független változó együttes valószínűségének meghatározására. Az akusztikai és vizuális csatorna a posteriori valószínűségeit a két modalitásra külön-külön meghatározva utólagos integrálással megkereshető a legvalószínűbb jelölt. Massaro és Stork Massaro–Stork (1998a) kimutatta, hogy a vizuális és az akusztikai jel között erős korreláció áll fenn, egy kategórián (szó, szótag, hangpár, hang) belül azonban a jellemzők véletlen eloszlásúak, feltételesen függetlenek. Ebben az esetben a két modalitás valószínűségei szorzatának maximuma szolgáltatja az optimális döntést. Az

utóbbi években a webkamerák elterjedésével új lendületet kapott az audiovizuális beszédfelismerés. A monitorra szerelt, vagy beépített kamera az előtte ülő beszélő képét konszolidált körülmények között rögzíti, így a képpel kiegészített beszéd felismerése a pusztán akusztikai jelre támaszkodó beszédfelismerésnél hatékonyabb (Potamianos et al. 2004).

9.13. Beszédfelismerők minősítése

Csapó Tamás Gábor

A gépi beszédfelismerés minősítése komplex feladat, mivel minden alkalmazás egyedi megoldásokat tartalmaz. Általános minősítéshez olyan paramétereket határoznak meg, amelyek függetleníthetők az alkalmazástól. Ilyen a szóhibaarány (WER = Word Error Rate). A WER érték a helytelenül felismert szavak és az eredeti beszédrészlet szavainak arányát adja meg. Ezt a következő módon lehet kiszámítani:

$$WER = \frac{S+D+I}{N}, \tag{9.66}$$

ahol „S” (Substitution) jelöli a felcserélt szavak számát, „D” (Deletion) jelöli a törölt szavak számát, „I” (Insertion) jelöli a beillesztett szavak számát, és „N” jelöli a valódi beszédrészlet összes szavainak számát. A szóhibaarány számításához elengedhetetlen a referenciaként szolgáló átirat szavainak és a felismerő kimenetén kapott szósorozat egymáshoz illesztése, amelyet dinamikus programozással valósíthatunk meg. A dinamikus programozás a két szóláncot illeszti egymáshoz úgy, hogy azok között a legkisebb hibájú illeszkedést kapjuk, tehát lényegében párba állítja a helyesen felismert szavakat, és meghatározza a felcserélt, törlődött, illetve beszúrt szavakat is. A beszédfelismerési eredmények kiértékelésére néha a szófelismerési pontosságot (Word Recognition Accuracy, WAcc) adják meg, ez lényegében egyenrangú a szóhibaarány megadásával, mivel $WAcc = 1 - WER$. A szófelismerési arány (Word Recognition Rate, WRR) a helyesen felismert szavak arányát adja meg, számítása megegyezik a WER értékével azzal a különbséggel, hogy a beszúrási hibát nem vesszük figyelembe (nem adjuk hozzá a számlálóhoz a 9.66 összefüggésben). Emiatt a WRR érték nem is igazán tekinthető egzakt mérőszámnak, viszonyalg ritkán is használják. Speciális rendszerek esetén a fenti mennyiségek helyett az alacsonyabb szintű fonémahiba-arány (Phoneme Error Rate, PER), valamint a betűhibaarány (Letter Error Rate, LER), illetve magasabb szintű mondathiba-arány (Sentence Error Rate, SER) hibatípus vizsgálata is lehetséges. A betűhibaarány (LER) jelentőségét az adja, hogy – például diktálórendszerek esetén – ez az érték jobban korrellál a kézi hibajavítás költségével, mint a szóhibaarány (WER), továbbá az erősen tolda-

lékoló nyelvek esetén – ilyen a magyar is – a betűhibaaránnyal korrektebben mérhető a beszédfelismerési pontosság változása. A fonémahiba-arány (PER) elsősorban fonémafelismerő rendszerekben használatos. A felismerés minősége mellett fontos tényező a beszédfelismerés gyorsasága is, melyet a valós idejű működéshez képest szokás mérni (Real Time Factor, RTF). Az RTF nem más, mint a beszédfelismerési számításokra fordított idő és a felismerendő beszéd időben mért hosszának aránya. Az RTF < 1.0 tulajdonságú beszédfelismerők képesek a beszédet annak elhangzása alatt felismerni, így valós idejű rendszerekben is használhatóak.

A beszédfelismerők pontossága és teljesítménye sok tényezőtől függ:

- az elhangzott beszéd témája,
- a beszédjel akusztikai minősége,
- zavaró környezeti tényezők,
- a tanítóadatbázis mérete és bonyolultsága,
- a szótáron kívüli szavak aránya,
- rendelkezésre álló memória és számítási erőforrás,
- történt-e akusztikai és nyelvimodell-adaptáció,
- történt-e rátanulás az adott személy beszédére.

A beszédfelismerő rendszereket tipikusan egy-egy konkrét alkalmazáshoz illesztve használják, ezért a minősítéskor is célszerű figyelembe venni a működés körülményeit. Elengedhetetlen az is, hogy szisztematikusan és teljesen objektív módon lehessen mérni, melyik beszédfelismerő motor felel meg leginkább egy konkrét feladattípus elvégzésére. Fontos továbbá, hogy a kiértékelés könnyen megismételhető legyen. Három minősítési munkát mutatunk be a legújabbak közül. A MERL (2008) projekt célja olyan szabványos interfészek és tesztek tervezése, kifejlesztése és megvalósítása volt, amelyek szabványos korpuszokon képesek mérni a különböző felismerő motorok teljesítményét, memóriára és számításigényre robusztus módon. A rendszerek teljesítményét több szabványos adatbázison tesztelik (angol nyelvre például a TI digits, a Macrophone, a Broadcast News, valamint a realisztikusabb, telefonos vagy autós környezetben gyűjtött adatbázisok).

A beszédfelismerő motorok egyik tipikus felhasználási területe a diktálórendszerek. A különböző kereskedelmi szoftverek akár 98%-os felismerési pontosságot is ígérnek, azonban ez a gyakorlatban nem feltétlenül elérhető. Egy minősítési kísérletben (Burger et al. 2006) ötféle beszédstílus (többek között felolvasott és spontán beszéd) felismerését vizsgálta három kereskedelmi szoftverben, nyolc nyelven alkalmazva. Az már a teszt elején kiderült, hogy az egyes nyelvekben előforduló dialektusokat meglehetősen rosszul kezelték a rendszerek. A vizsgálatot átlagos irodai körülmények között végezték (PC, irodai zaj, olcsó fejhallgatós mikrofon). Először elvégezték a gyártók által javasolt minimumadaptációt az egyes beszélőkre. Az elemzést a WER metrika mérésével kezdték, majd ezt a mértéket szófelismerési arányra váltották át. Az átlagos eredmény 66–76% körüli felismerési arány volt (amely jóval

alacsonyabb a hirdetett 98%-nál). A beszédtípus (olvasott/spontán) és a felismerés pontossága erősen összefügg: a spontán beszéd átlagosan 13%-kal rosszabb eredményt ért el az olvasottnál. A kutatás fő eredménye, hogy még távol állunk a spontán beszédet jól felismerni képes rendszerektől.

Yao és kollégái a gyakorlati felhasználás szempontjából vizsgáltak néhány szabad forráskódú beszédfelismerő rendszert (Yao et al. 2010). Az értékelést nem beszédtechnológiai szakértőkkel végeztették, hanem olyan felhasználókkal, akik éles környezetben, virtuális ember-gép dialógus formájában használták a rendszert. Összesen hat adathalmazt használtak három szabad forráskódú felismerő motor teszteléséhez. Eredményeik azt mutatják, hogy a társalgás témája nagyban befolyásolja a beszédfelismerés sikerességét. Fő következtetéseik közé tartozik, hogy általános témájú beszélgetés felismerésére a vizsgált beszédfelismerő motorok még meglehetősen rosszul működnek. Összefoglalásként elmondhatjuk, hogy a 21. század elején a gépi beszédfelismerési technika már számos gyakorlati feladatra jól alkalmazható, az általános beszéd-szöveg átalakítás problémája azonban még messze nem tekinthető megoldottnak.

10. fejezet
A beszéd gépi előállítása

Olaszy Gábor

Az ember régi vágya, hogy a gépeket megtanítsa beszélni. Noha a gépi beszéd-előállításra tett kísérletek már több mint 200 évre nyúlnak vissza, az igazi „beszélő” gépek megvalósítására csak a 20. század második felétől volt mód. Ezt a számítástechnika fejlődése tette lehetővé. Négy alapszabályt lehet megfogalmazni a korszerű gépi beszéd-előállítással mint technológiával kapcsolatosan.

a) A géppel előállított beszéd létrehozása interdiszciplináris tudást igényel és csapatmunkában lehet csak sikeres. Mérnök, nyelvész, akusztikus, informatikus, fonetikus alkothatja a csapatot.

b) A beszédszintetizáló rendszert – legyen az bármilyen egyszerű vagy bonyolult – a hangminősége alapján ítéli meg a felhasználó. A hangminőség és az érthetőség alakítja ki a véleményt. Ennek minőségi szintje nemcsak a tervezőtől függ, hanem az alkalmazótól és a technikai lehetőségektől is. A magas minőségi követelményű rendszerekben kölcsönös együttműködésre van szükség.

c) A gépi beszéd-előállítás kompromisszumokat követel mind a fejlesztőtől, mind az alkalmazótól. Az optimális kompromisszumok kialakítása határozza meg a legmegfelelőbb megoldást. A fejlesztőnek számolni kell az alkalmazó technikai követelményeivel (memória, vezérlés, átvitel), az alkalmazónak a nyelvi szabályokkal, korlátokkal (csak magyarul tud a rendszer, nem tud stílusok között váltani, nincs felkészítve érzelmek kifejezésére stb.).

d) A gépi beszéd-előállítás kulcskérdése, hogy a beszéd négy legfontosabb paraméterének fizikai megvalósítása során a folyamatosságot kell megvalósítani. A négy paraméter a következő: beszédhangok spektrális szerkezete; alapfrekvencia; intenzitás és időszerkezet. Mindegyik paraméter állandóan változik, tehát ezeknek a módosulásoknak a folyamatosságát kell megvalósítani (nem lehetnek hirtelen változások egyikben sem). Ennek megvalósítása nehéz, minden esetben kompromisszumokat követel.

A gépi beszéd-előállítás technológiájával kapcsolatosan két alapvető kérdésre kell megoldást találni. Az egyik, hogy mi hangozzék el a gép „szájából”, a másik,

hogy azt hogyan állítsuk elő. Azt, hogy mi hangozzék el a gép „szájából", a számítógépes rendszereknél általában szöveges formában adják meg. A 21. század elejére jellemző gondolkodás e tekintetben elvárja, hogy a számítógépek versenyezzenek az emberrel (a gép úgy olvasson, mint egy ember). Ezt csak bizonyos esetekben lehet megtenni, hiszen az ember beszélni és olvasni tudását sok évtizedes folyamatos információszerzés támogatja. Tudjuk, hogy például bizonyos idegen neveket hogyan kell kimondani, tudjuk, hogy a *220V* betűsort hogy kell felolvasni (és főleg tudjuk, hogy milyen fogalom rejlik mögötte), de nem biztos, hogy az átlagember tudja, hogy a műszaki kifejezéseket, a gyógyszerek hatóanyagait, a gépkocsikkal kapcsolatos szakkifejezéseket stb. hogyan kell helyesen ejteni. A gépi szövegfelolvasás általános és teljes körű megoldása tehát még várat magára. De gondoljunk bele az ember teljesítményébe. Mi, emberek sem tudunk mindent azonnal hibátlanul felolvasni. Mindenki csak a saját életteréhez, szakmájához tartozó dolgokat tudja helyesen értelmezni, felolvasni. A jogász belebonyolódik, ha egy építési technológiai leírást kell felolvasnia, az orvos nemigen birkózik meg számítástechnikai szövegekkel, az autóhirdetések felolvasásához pedig ismernünk kell a különböző márkákat, motortípusokat, egyéb jellemző szakszavakat. Mindezekkel a példákkal azt akarjuk érzékeltetni, hogy a gépi szövegfelolvasás közel hibátlan megvalósításához minden konkrét feladat esetében a gépi felolvasórendszert rá kell illeszteni a feladatra. Ha híreket kell felolvastatni, akkor a felmerülő nevek, tulajdonnevek felolvasására kell megtanítani a gépet, és a hírolvasás prozódiai stílusát kell utánozni, ha például tudakozórendszerben használjuk a gépi felolvasót, akkor cégnevek, személynevek, utcák, települések, postacímek helyes kiejtésére kell rátanítani a rendszert. Mindezt különböző írásformákból kell kinyerni, hiszen az írott információ azokban az adatbázisokban, ahonnét a gépnek azt átadják, nem gépnek készült eredetileg, hanem embernek. Lássunk egy példát arra, hogy az operátor hányféleképpen rögzíthet egy postacímet, ami emberi olvasatban abszolút nem látszik problémásnak: *Kispipa u. 3/1em. 4; Kispipa 3 I.4; Kispipa utca 3-I/4; Kis pipa u. 3. 1/4*. Ha például mobiltelefon-készülékekkel kapcsolatos automata információs szolgáltatást akarunk megvalósítani, akkor a készülékek fantázianeveinek kiejtésére kell a gépet megtanítani, és ezt folyamatosan kell tennünk, mert új fantázianevek jelennek meg időről időre (Németh et al. 2009). A tényleges hang előállítása szempontjából a következő két megoldási stratégia lehetséges.

Az emberi hangképzés és artikuláció utánzása. Ezzel kapcsolatosan a fejlődés két végpontját említhetjük. Egyrészről Kempelen mechanikus gépe ezt próbálta imitálni a 18. század végén, másrészről ez lehet a jövő technológiája, amikor algoritmusokkal fogják utánozni a biológiai beszédképzést, vagyis artikulációs beszédszintetizátorokat hoznak létre (jelenleg ilyen rendszer gyakorlati felhasználásban nem működik sehol a világon). Artikulációs beszédszintézissel kapcsolatos algoritmusok laboratóriumi fejlesztéseiről folyamatosan hírt ad a szakirodalom,

azonban az ilyen kutatások még csak szórványosnak mondhatók. Példaként említjük, hogy a 2009-ben megrendezett Interspeech világkonferencia 756 előadásából mindössze néhány foglalkozott közvetlenül vagy közvetve az artikulációs csatorna modellezésével, illetve az ilyen modelleken alapuló artikulációs beszédszintézissel (Bawab et al. 2009, Cai et al. 2009). Kempelen gondolatmenetének folytatására is van példa. Robotok beszédének előállítására artikulációs elektromechanikus beszédkeltőkkel kísérleteznek (Nakamura–Sawada 2006), amelyek az emberhez hasonló módon adnak hangot.

Az akusztikai produktum utánzása. Ennél a megközelítésnél a beszéd hullámformáját használják kiindulásként. Kétfajta megoldás alakult ki az ilyen beszédépítésre. Az úgynevezett forráskódolású megoldásokban a beszédjelből kivonják a lényegi információt és mint adatsorozatot használják a szintézis során. Ennél a megoldásnál minden esetben gondoskodni kell arról, hogy az adatokból ismét hullámforma álljon elő (szűrőrendszer, kódoló). Ilyen elvet követ a formánsszintézis, az LPC alapú beszéd-előállítás, de a HMM alapú megoldás is.

A másik megközelítésben az emberi hangot közvetlenül használják fel a beszédépítéshez. Ehhez hullámforma-elemtárakat készítenek, amelyekben különböző hosszúságú hullámforma-részleteket tárolnak, és ezek megfelelő összefűzésével hozzák létre a kívánt beszédhullámot. Ilyen megoldásokat alkalmaznak az elemösszefűzésen alapuló eljárások, illetve az elemkiválasztás-alapú technológiák. Ezeknél a megoldásoknál esetenként jelfeldolgozást is alkalmaznak a hullámforma végleges kialakításánál (például dallam, hangsúly előállítása).

A hullámformát közvetlenül felhasználó megoldásokban külön osztályozást adhatunk a beszéd szegmentális és szupraszegmentális szerkezetének kezelése szempontjából. A szintézis során kezelhetjük külön-külön a két szintet (egymásra építve), vagy egyszerre (hasonlóan, ahogy az ember teszi). Mindkét eljáráshoz valamilyen prozódiai modellre van szükség. A gyakorlatban inkább a két szint elválasztása a jellemző (forráskódolású és elemösszefűzéses megoldások). Az elemkiválasztásos technológia olyan modellt használ, amelyik egyszerre veszi figyelemben a hangtestet és a prozódiai tartalmat.

Módszertan. A beszéd-előállítás módszere tekintetében kétfajta megoldást különböztetnek meg: szabályalapú rendszerek, illetve statisztikai elven működő megoldások (ilyen a gépi tanulás is). A szabályalapú megoldásnál megfigyelések alapján szabályokat alakítanak ki a beszéd-előállítás minden szintjére, mozzanatára. Ennek a megoldásnak az előnye, hogy a szabályok működése és hatóköre ismert, ezért a rendszer tesztelése során felmerülő hibák javítás hatékony lehet. A hátránya, hogy szabályokkal nem lehet leírni a beszéd finom részleteit. A statisztikai elvű megoldásoknál a finom részletek is megmaradhatnak (például ráismerünk a beszélő hangjára), viszont az elvből adódik, hogy a hibajavítás nem annyira egyszerű. Ez

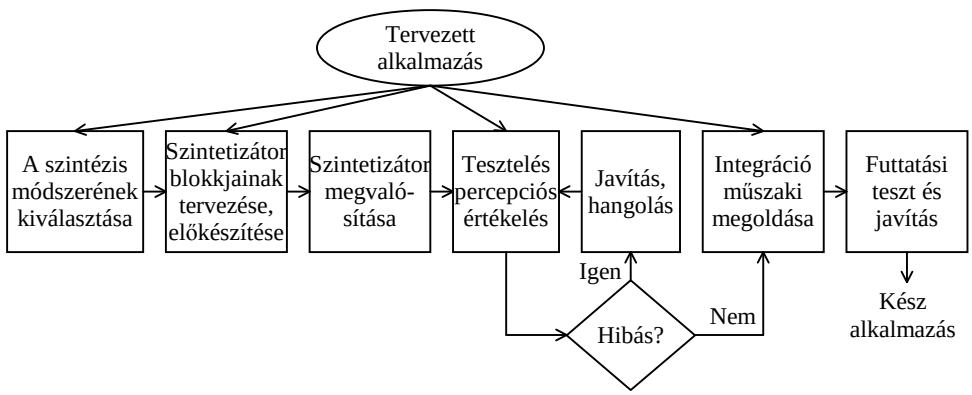
azért van, mert a belső, valószínűségeken alapuló rendszerállapotok működését kívülről nem lehet befolyásolni.

Gyakorlati megvalósítások. A gépi beszédkeltés gyakorlati megvalósításait három csoportba sorolhatjuk: kötött szótáras rendszerek, szövegfelolvasók (általában ezt szokták beszédszintézisnek nevezni) és a kettő kombinálásából adódó úgynevezett hibrid rendszerek. A kötött szótáras rendszerek csak előre meghatározott üzeneteket szólaltatnak meg. A szövegfelolvasókat Text-to-Speech (TTS) további alosztályokba sorolják:

- Általános szövegfelolvasó, amely nincs témakörhöz kötve. Ez általánosságban egy 12 éves gyerek szókincsét és olvasási képességét valósítja meg.
- Szövegfelolvasó az alkalmazáshoz hangolva. Ennél a megoldásnál az alkalmazás szövegtípusaihoz adaptálják a szövegfelolvasó algoritmusait, a kiejtési kivételek szótárait, a prozódiai modult. Ilyen alkalmazások a szépirodalmi felolvasók, a turistatájékoztatók, a hírolvasók, a hangos időjárás-jelentés.
- Erősen a témakörre szűkített szövegfelolvasó (Concept-to-Speech). Ilyen lehet például a fent említett árlistafelolvasó mobiltelefonokra vagy az orvosi, jogi, műszaki témájú szűkített tematikájú szövegeket meghangosító megoldás. Van példa mesemondásra irányuló kísérletre is (Székely 2009).
- Több nyelvű felolvasók (Multilingual TTS). Vannak olyanok, amelyek ugyanazon technológiát alkalmazva ugyanazon hangkarakterrel szólnak. Ilyen rendszer volt a MultiVox magyar rendszer, amely 12 nyelven tudott beszélni (Olaszy et al. 1992). Mindegyik nyelv ugyanazzal a formánsszintetizátorral szólalt meg, tehát a hangzásban lényeges eltérés nem volt. Más technológia esetén viszont a többnyelvű rendszerekben más-más hangon is szólhat a rendszer. Ilyenre példa a ProfiVox technológia, amelynek hullámforma-elemtárát minden nyelvre más ember beszédéből alakították ki.
- Több nyelvű, azonos hangú (poliglott) felolvasók. Ugyanazon személy hangján szólnak több nyelven (nehéz megvalósítani). Az ilyen megoldások azt a célt tűzik ki, hogy kevert nyelvű szövegek gépi felolvasása is lehetővé váljon (Pfister–Romsdorfer 2003).
- Hibrid rendszerekben a kötött szótáras megoldást és a szövegfelolvasást kombináltan alkalmazzák. Ilyenre példa a magyar automatikus szám szerinti tudakozóban használt név- és címfelolvasó beszédszintetizátor megoldása (12.3.4. fejezet).

Alapvető tervezési szempontok. A tervezés során az alkalmazásból kell kiindulni. Az itt elvárt követelményrendszer határozza meg a felolvasó tervezésének, tesztelésének, integrálásának fő körvonalait. Ezek minden beszédszintetizátorra érvényesek (10.1. ábra). Első lépésben meghatározzuk a módszert, az alkalmazandó technológiát. Itt döntjük el a szintézis alapelemét is, vagyis hogy milyen építőelemeket használunk majd a beszéd felépítéséhez. A tervezés második lépése a szintetizátor

moduljainak meghatározása, az alkalmazandó algoritmusok kialakítása. A harmadik lépésben összeállítjuk a beszédszintetizátort. E fejlesztési fázis végén már megszólal a rendszer. A negyedik lépés a tesztelés, ennek során a kitűzött alkalmazási feladatnak megfelelően beszéltetjük a szintetizátort és értékeljük a hallottakat mondatról mondatra. A kiejtési hibákat gyűjtjük, majd kategorizáljuk (hanghibát és prozódiait egyaránt). Ezt csak ember tudja elvégezni. Az ötödik fázis a hibajavítás. A célkitűzés, hogy a tesztelés előrehaladtával minél ritkábban találkozzunk kiejtési hibával, hangcsuklással, rossz szünettel stb. A tesztelés és a hibajavítás munkaigényes feladat, türelmet és szakértelmet kíván. A hatodik munkaszakaszban a végleges rendszert integráljuk az alkalmazás éles környezetébe, és ott futtatási és terhelési tesztnek vetjük alá.

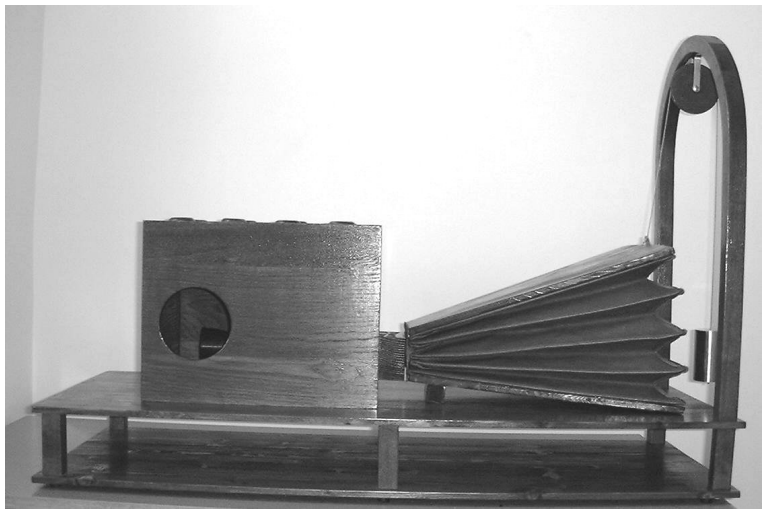


10.1. ábra. Beszédszintetizátorok fejlesztésének általános folyamata

10.1. Kempelentől napjainkig

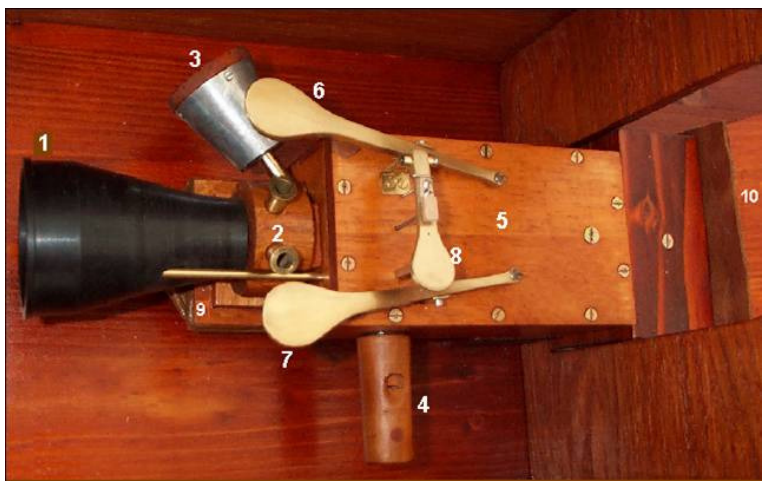
Olaszy Gábor

A világ első beszédkeltő gépét Kempelen Farkas készítette a 18. század második felében (Kempelen 1791). Mintegy 22 éves kísérletezés eredménye volt az 1791-ben bemutatott végleges szerkezet (több gépet készített a kutatásai során). A gép pontosan rekonstruált és működő változata (10.2. ábra) ma is látható az MTA Nyelvtudományi Intézetében (Nikléczy–Olaszy 2002). Kempelen ezzel a kutatásával 200 évvel előzte meg korát, és mai szemmel azt mondhatjuk, hogy ő készítette a világ első artikulációs, többnyelvű beszédszintetizátorát. Kempelen munkássága alapozta meg a fonetika tudományát, és néhány megfigyelése ma is megállja a helyét. Az idézett könyve tulajdonképpen az első komoly, tudományos fonetikai leírás a beszédet



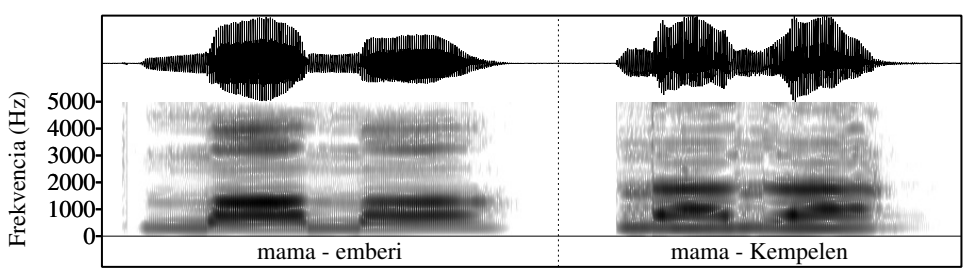
10.2. ábra. Kempelen beszédkeltő gépének rekonstruált, működő változata eredeti méretben

illetően. Ő mondta ki először, hogy a beszédhangokat a két hang között elhelyezkedő hangkapcsolódási rész köti össze, e nélkül nem létezik folyamatos beszéd. Ezzel definiálta a mai hangátmenet fogalmát (Dudley–Tarnóczy 1950). Kempelen gépe az emberi artikulációs csatorna felépítését követte (korlátozottan, az akkori lehetőségekhez mérten), vezérléséhez egy képzett kezelőre volt szükség, aki mindkét kezét (és agyát) használva működtette a gépet (10.3. ábra).



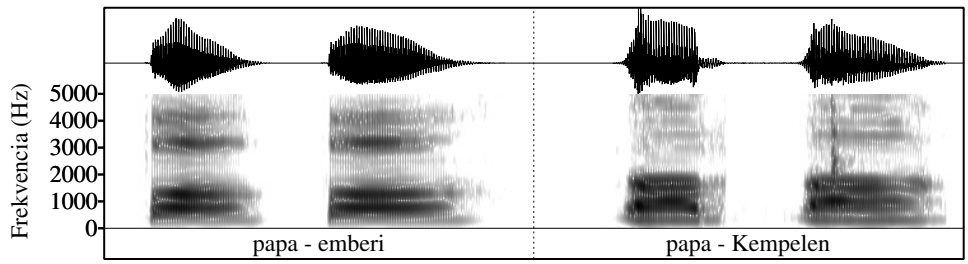
10.3. ábra. Kempelen gépének hangkeltő része, az emberi artikulációs csatorna felépítésének utánzása

A gép részegységei a következők: 1 – szájüreg és ajkak imitálására szolgáló gumitölcsér, 2 – orrcimpákat képviselő két csövecske, 3 – az [f] hang hangkeltő rezonátora és 6 – működtető billentyűje, 4 – az [ʃ] hang hangkeltő rezonátora és 7 – billentyűje, 5 – a gége tere, benne a hangszalagokat reprezentáló rezgő nyelvvel, 8 – az [r] hangot létrehozó billentyű, 9 – pótfújtató a zárhangokat létrehozó nyomás növeléséhez, 10 – fújtató a tüdő imitálására. Kempelen gépével természetesen nem lehetett hosszú mondatokat kimondatni, de a szerkezet elégséges volt ahhoz, hogy a tudós bebizonyítsa az emberi beszédképzés alapvető működését (soros elrendezésű rendszer, amelyben az alapot a zöngehang jelenti, majd ezt módosítja az artikulációs csatorna pillanatnyi fizikai formája, amit a nyelv, állkapocs, ajkak mozgatásával változtatunk). A gép fizikai és elvi korlátait csak a rekonstrukció után lehetett megállapítani 2002-ben (visszamentünk az időben mintegy 200 évet). Azokat az állításokat lehetett ellenőrizni, amiket Kempelen a könyvében leírt. A lefolytatott kísérletek megmutatták, hogy Kempelen is tisztában volt gépe korlátaival, noha nem ismerhette olyan pontosan a hangképzés minden részmozzanatát, ahogy a mai kutatók. A géppel végzett kísérletek azt mutatták, hogy bizonyos magánhangzókat is és mássalhangzókat is nehéz, illetve egyáltalán nem lehet előállítani ilyen mechanikus szerkezettel (Nikléczy–Olaszy 2004). A legeredményesebben a CVCV jellegű rövid hangsorok hozhatók létre, ahol a C bizonyos nazális-, rés-, illetve zárhang, a V pedig az [ɔ], illetve az [aː] hang. Mondatokat nem lehet a géppel megszólaltatni. A 10.4. ábrán bemutatjuk a *mama* szó emberi és gépi kiejtésének rezgésképét és hangspektrogramját, a 10.5. ábrán pedig a *papa* szót mutatjuk be. A rezgéskép és a hangspektrogram alapján szemre is megállapítható, hogy a két hangzás nagyon hasonló lehet. Kempelen

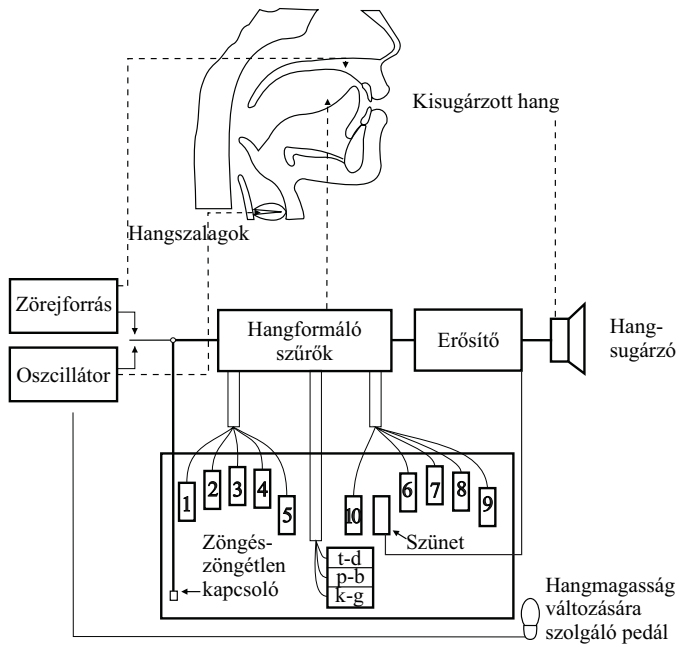


10.4. ábra. A *mama* hangsorok rezgésképe és hangspetrogramja emberi ejtésben és Kempelen gépével előállítva

gépének hangja meghallgatható a `http://fonetikai.nytud.hu` honlapon. Kempelen követői számos mechanikus szerkezetet hoztak létre, majd a forradalmi változást ezen a területen az elektronika fejlődése hozta el a 20. század első felében. A Bell Laboratóriumban mutatták be 1939-ben az első elektromechanikus angolul beszélő gépet VODER (Voice Operation DEmonstratoR) néven (Homer et al. 1939). Elviekben ezzel a géppel már folyamatos beszédet tudtak létrehozni (tetszőleges hangsort



10.5. ábra. A *papa* hangsorok rezgésképe és hangspetrogramja emberi ejtésben és Kempelen gépével előállítva



10.6. ábra. A VODER elektronikus áramköreit vezérlő pult és az emberi beszédképző szervek megfeleltetése (Homer et al. 1939) alapján

lehetett vele előállítani), bár a hangja alig volt érthető. A gép működtetését még ekkor is ember látta el, egy képzett hölgy kezelte a billentyűket tartalmazó egységet (hangokat lehetett megszólaltatni), egy lábpedállal pedig a hangmagasságot lehetett változtatni (10.6. ábra). Ez többek között annyit jelentett, hogy a beszédhangok időtartamát az ember határozta meg, továbbá azt is, hogy minimális dallamot tartalmazott csak a generált beszédhullám. A gyakorlatban a VODER hangja – hasonlóan Kempelen gépének a hangjához – nehezen volt érthető. A VODER 1939-es bemu-

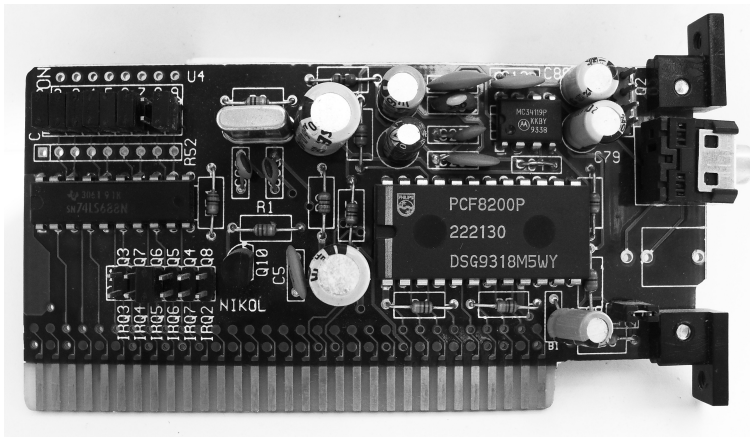
tatóján előre bemondták, hogy milyen mondatot fog megszólaltani a kezelő hölgy a géppel (lásd http://blog.makezine.com/archive/2009/02/speech_synthesis_in_the_year_1939.html).

A következő nagy ugrást a számítógép alkalmazása jelentette. Többek között ez váltotta ki az emberi kezelőt. Számítógéppel már sokkal gyorsabb vezérlést lehetett megvalósítani, ezzel lehetőség nyílt a beszéd finomabb részleteinek a megvalósítására. Ehhez viszont modellezni kellett a beszédképzés összes paraméterét az idő függvényében. Az elméleti megalapozást egy svéd tudós dolgozta ki (Fant 1960). Az első számítógéppel vezérelt beszédszintetizátorok az emberi beszédképző mechanizmust utánozták egyszerű eszközökkel, egy zönge/zaj generátor hajtott meg egy lineáris, idővariáns, vezérelhető többkomponenses szűrősort. A szűrők hozták létre a formánsokat, tehát az eljárást formánsszintézisnek nevezték el (Rabiner 1968). Egy évtizeddel később lépett be a témakör kutatásába Magyarország. Az MTA Nyelvtudományi Intézetének fonetikai laboratóriumában készítették el az első magyar szövegfelolvasó rendszert – Hungarovox –, amely magyarul olvasott fel tetszőleges szöveget. A Hungarovox rendszer egy nagy méretű beszédszintetizátorból (10.7. ábra), egy PDP 11/34-es DEC gyártmányú (könyvszekrény nagyságú) számítógépből, valamint a szintézis Fortran nyelven írt szoftveréből állt. A számítógép központi memóriája 64 kByte volt. A Hungarovox formánsszintézis elven működött, tehát alapvetően fonetikai jellegű modellezést valósítottak meg a szoftverben (Kiss–Olaszy 1984). Nyilvánosan 1983-ban mutatták be. Magyar szabadalom (Olaszy–Kiss 1982), lajstromszáma 185527.

10.7. ábra. A Hungarovox szövegfelolvasó beszédszintetizátor hangkeltő egysége (30x20x10 cm), amely párhuzamos porton csatlakozott a vezérlő számítógéphez

Ugyancsak 1983-ban készült el az első magyar fejlesztésű, kötött üzeneteket meghangosító beszédszintetizátor is LIAVOX néven a BME Híradástechnikai Elektronika Intézetben (Gordos et al. 1983). Ez tárolt beszédből építkezett, és meghatározott témakörben tudott mondatokat generálni. A gyakorlati felhasználására 1985-ben került sor (Gordos–Sándor 1985). Az elkövetkező években a fonetikai laboratórium és

a BME akkori távközléssel foglalkozó tanszéke (Híradástechikai Elektronika Intézet, majd később Távközlési és Telematikai Tanszék) összefogott és több generációt is kifejlesztett a szövegfelolvasó rendszerekből, amelyek mindig igazodtak a technika fejlődéséhez. Akkoriban az MTA Kísérleti Fizikai Kutatóintézetben is fejlesztettek magyar szövegfelolvasót, főleg vakok segítésére (Kiss et al. 1987). A PC-k megjelenésével igény volt önálló beszédszintetizátor kártya megvalósítására is. Ez a 80-as évek végén született meg PCROBOT néven (10.8. ábra). A beszédszintetizátor a Philips cég második generációs formánsszintetizátorát (PCF 8200 chip) alkalmazta a beszédkeltésre (Olaszy 1994a). Az első gépi beszédelő rendszert távközlési al-



10.8. ábra. A PCROBOT beszédszintetizátor kártya a 80-as évekből

kalmazásban állították üzembe Magyarországon, ahol számváltozással kapcsolatos információt mondott el a gép (Takács 1989). Az 1980-as évek végén jelentek meg az első hangkártyák, amelyek lehetővé tették a hangok elektronikus tárolását. Magyarországon ekkor született meg az első, emberi beszédhullámból építkező beszédszintetizátor (Király 1989), amely ugyan erősen robotos hangon beszélt, azonban helyfoglalása összesen 140 kByte volt. A hangkeltés módszere viszont forradalmian újnak számított, egyrészről emberi hangból építkezett, másrészről csak szoftverből állt. Ebből az elvből fejlődtek ki később az emberi hangon beszélő beszédszintetizátorok (lásd a későbbi fejezetekben). Az ezutáni két évtizedben a BME fent említett tanszékén (ma Távközlési és Médiainformatikai Tanszék) folyamatos kutatások folytak, és a világ új kutatási irányaihoz igazodva, valamint lépést tartva a technikai fejlődéssel a szövegfelolvasó technológiák három új generációja született meg. Ezek elviekben is élesen különböztek a korábbi, formánsszintézises technikától. Az első generációváltás során egy bemondó hangjából (logatomokat olvasott fel) célzottan kivágott hullámforma-részleteket használták a hanghullám felépítésére. Ehhez a technológiához köthetők a 8.2.2, és a 10.3.6. fejezetek. A prozódiát jelfeldolgozással

alakították ki, tehát az eljárás kétlépcsős. A második generációváltásnál új gondolat lépett előtérbe, főleg a memória rohamos növekedésének köszönhetően. Kezdtek elhárulni a tárolási korlátok, megnyílt az út hatalmas beszédadatbázisok létrehozására. Egy kiválasztott bemondó hangját rögzítették (felolvasással), és szavakat, mondatrészeket válogattak ki a szintézishez (minden esetben a szövegnek megfelelőt). Ennél a technológiánál a prozódiai változásokat nem különválasztva, utófeldolgozással hozzák létre, hanem – megfelelő modellezéssel és keresési módszerrel – a beszédhangsor létrehozásával egy időben, ahogy az ember teszi. Ez a technológia adja a teljesen emberi hangzást, de jelenleg csak szűk témakörre működik hatásosan. Ehhez a technológiához köthetők a 8.2.3, és a 10.3.7. fejezetek. A harmadik generációváltást ismét új gondolat jellemzi. Statisztikai elvű gépi tanuláson alapul és rejtett Markov-modelleket használ fel a szintetizálandó jel létrehozására. Ehhez a technológiához köthető a 10.3.8. fejezet. A prozódiai megformálást itt a tanulás alapján végzi a rendszer, utólagos jelfeldolgozásra nincs szükség. A szintetizált hullámformát egy kódoló (hangvisszaállító) kimenete szolgáltatja. Mindezek mellett fontos látnunk, hogy noha újabb és újabb megoldások születnek, a technológiai generációk életképessége ezzel nem szűnik meg, mindegyiknek van valamilyen előnye, ami egyes alkalmazásoknál kritikus lehet (Németh et al. 2009). Például a formáns (vagy más parametrikus) szintézis esetében könnyedén lehet suttogó hangot előállítani és viszonylag egyszerűen gyorsítható, illetve lassítható az előállított beszéd, ami a közvetlen elemkiválasztás-alapú vagy a HMM megoldásoknál csak nehezen valósítható meg. A Magyarországon fejlesztett szövegfelolvasó technológiákat időrendi sorrendbe szedve a 10.1. táblázatban adjuk meg.

Magyarul beszélő, ingyenes szövegfelolvasó szoftver tölthető le a

`http://speechlab.tmit.bme.hu` címről.

A szoftver létrehozását a Miniszterelnöki Hivatal Informatikai Kormánybiztossága támogatta 2000–2002 között.

10.2. Kötött szótáras beszédszintetizátorok

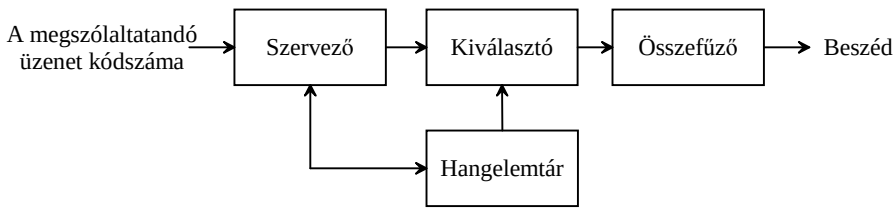
Olaszy Gábor

A kötött szótáras beszédszintetizátor-technológia megszületését a műszaki fejlődés két tényezője segítette a 20. század közepén. Egyrészről felmerült az igény a gépi beszéd alkalmazására (egyszerű távközlési szolgáltatásokban, mint pontosidőbemondó), másrészről a számítógépes hangkártyák is megjelentek. Ez lehetővé tette, hogy előre felvett hangüzeneteket automatikusan távközlési csatornára továbbítsanak. Ezt a technológiát ma is széles körben használják egy adott szűk témához kapcsolódó, véges számú hangüzenet előállítására (bankszámla, óraállás, indulási idő, megálló neve, pontszámjóváírás, navigációs célravezetés [GPS] stb.). Ilyen rend-

10.1. táblázat. Szövegfelolvasó rendszerek fejlesztése Magyarországon

Év	Beszédkeltő	Hangadás alapja	Számítógép, (op. rend.)	Felolvasó szoftver	Hangminőség	Tárigény	Fejlesztő
1980	VOX 12 készülék	hangszeletek formánskódolással	PDP-11/34	Hungarovox (magyar)	nagyon robotos, érthető	1kByte	MTA NyI. Fonetikai Laboratórium
1983	MEA 8000 (Philips) chip párhuzamos porton táplálva (külső egység)	hangszeletek formánskódolással	Syster, Commodore 64, PC (DOS)	Skriptovox (magyar, német)	nagyon robotos, érthető	1 kByte	Fonetikai L. és- BME TTT
1985	MEA 8000 (Philips) chip kártyára szerelve	hangszeletek formánskódolással	Syster, Brailab számítógép	Brailab család (magyar)	nagyon robotos	1-2 kByte	KFKI
1986	PCF 8200 (Philips) chip párhuzamos, soros porton táplálva (külső egység)	hangszeletek formánskódolással	PC (DOS, Windows) Sun (Unix)	MultiVox (12 nyelvre)	robotos, érthető	2 kByte	Fonetikai L. és BME-TTT
1988	Szoftver	emberi hangból készített hangszeletek	PC (DOS)	PC talker (magyar)	robotos, érthető	-	Király József
1994	PCF 8200 PC-be dugható szintetizátorkártya	hangszeletek formánskódolással	PC (DOS, Windows)	PC-ROBOT (magyar)	robotos, érthető	2 KByte	NIKOL Elektronika és TTT
1998	Szintézisszoftver	emberi hangból kivágott diádok	PC (Linux, Windows NT)	„ProfiVox” I.	közel emberi, jól érthető	2 Mbyte	BME TTT
2000	Szintézisszoftver	emberi hangból kivágott diádok + triádok	PC, mobiltelefon	„ProfiVox” II.	közel emberi, jól érthető	90 Mbyte	BME TTT
2001	Szintézisszoftver	hangszeletek formánskódolással	PC	„MultiVox” ingyenesen letölthető (magyar)	gépies, de jól érthető	90 Mbyte	BME TTT
2005	Szintézisszoftver	beszéd-adatbázisból elemkiválasztás	PC, csak kötött témakörre	„ProfiVox-korpusz”	teljesen emberi, jól érthető	500 Mbyte	BME TMIT
2008	Szintézisszoftver	HMM alapú szövegfelolvasó	PC, mobiltelefon	„ProfiVox-HMM”	emberi, jól érthető	3 Mbyte	BME TMIT

szerek működnek a hívásközpontokban és más, egyszerű automatikus információs rendszerekben. A kötött szótáras beszédszintetizálók emberi beszédből kivágott hullámformarészletek (esetleg egész mondatok) összekapcsolásával hozzák létre a kimondandó üzenetet, jelfeldolgozási algoritmusokat nem alkalmaznak. A rendszer fő elemei a következők: elembázis (üzenetek, üzenetrészek hullámformáinak hangelemtára); elemző és kiválasztó algoritmus, ami meghatározza, hogy az elemtárból mely összefűzendő hangelemeket (üzenetrészeket) kell kiválogatni a bejövő parancs (információ) alapján; elemösszefűző és megszólaltató egység (10.9. ábra).



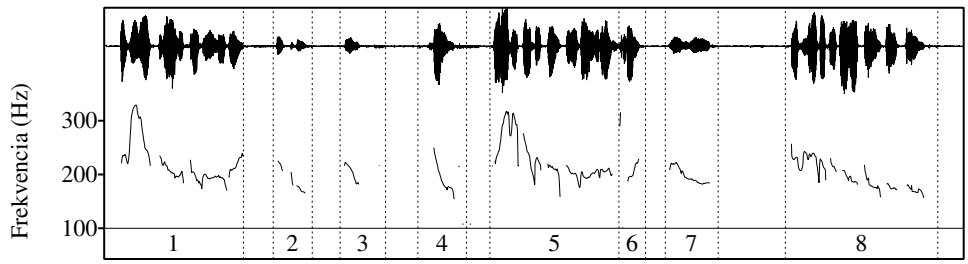
10.9. ábra. Általános kötött szótáras beszédszintetizátor blokksémája

10.2.1. Hangminőségi skála

Hangminőség szempontjából két csoportra oszthatjuk a közvetítendő üzenetet: egyszerű üzenetek, illetve összetett, több részből álló üzenetek. Az egyszerű üzenet csak egyetlen mondatból áll, változó elemet nem tartalmaz, minden esetben ugyanaz a mondat szólal meg. Az ilyen rendszerek hangminősége az elvből fakadóan nagyon jó, hiszen az ember által bemondott mondatot szólaltatjuk meg (például: *A Kossuth tér következik.*). Abban az esetben, ha az üzenetben változó rész is van, a tervezés gondossága erősen befolyásolja a kimondott (összeállított, összefűzött) üzenet hangminőségét (természetesen hangzik, vagy szaggatott, döcögő, huppogó a beszéde). Negatív példaként egy ilyen üzenetre mutatunk be példát.

Az Ön aktuális egyenlege: ***74*** *pont. A legutóbbi pontjóváírás a* ***január*** *havi adatok alapján történt.*

A változó üzenetrészeket félkövéren jelöltük. Amennyiben a változó rész eleme nem úgynevezett vivőmondatból lett kivágva, akkor akusztikailag kicsi az esély arra, hogy illeszkedni fog az őt megelőző és követő hangrészhez (10.10. ábra), szakadozott, kiegyensúlyozatlan, torz lesz a beszéd. Az ábrán látható szöveg a számozás szerint a következő: 1 – *Az Ön aktuális egyenlege;* 2 – *hetven;* 3 – *négy;* 4 – *pont;* 5 – *A legutóbbi pontjóváírás;* 6 – *a;* 7 – *január;* 8 – *havi adatok alapján történt.* Az elemek közé beiktatott hosszú szüneteket a jelölés nélküli szakaszok mutatják. Mi olvasható le az ábra diagramjaiból? Az összeállított üzenet 8 különálló hullámformarészt



10.10. ábra. A példamondat rezgésképe (fent) és dallammenete (lent) annak bemutatására, hogy ez a hangzás milyen távol van a természetes beszédtől.

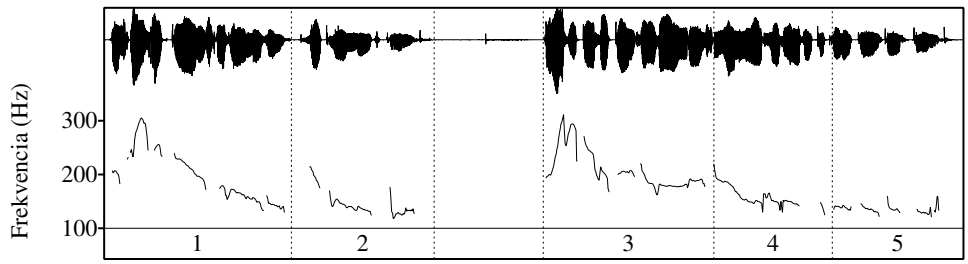
tartalmaz, mindegyik közé szünetet iktatott be a fejlesztő. A szünetek felesleges alkalmazása már eleve természetellenessé teszi a hangzást. Természetes ejtésben egy ilyen üzenetnél csak a 2. az 5. és esetleg a 6. rész előtt szoktak szünetet tartani. A jel amplitúdóját vizsgálva (felső rész) az látható, hogy a fejlesztő egymástól nagy szinteltérésű elemeket tárolt el az üzenetek hullámformáját tartalmazó elembázisban. A változó részek (2, 3, 4, 6, 7) amplitúdója mintegy háromszor kisebb, mint az üzenet állandó részéé. Ez természetes beszédben nem fordul elő, ott az amplitúdók nem térnek el tendenciájukban ennyire egymástól (a hangerő nem változik ilyen hirtelen). Az alapfrekvencia vonulata is természetellenes képet mutat. Egy ilyen közlésben az alapfrekvencia a közlés teljes tartama alatt fokozatosan, de lassan változik, nincsenek benne ugrások. A gondatlan tervezés tehát eredményezheti a nagyon rossz hangminőséget. A jó tervezéshez alapvető fonetikai, beszédakusztikai tudás szükséges. A gondosan megtervezett ilyen rendszerek hangminősége annyira jó is lehet, hogy a bemondó egyéni hangszínezete felismerhető (lásd a 8.2.1. fejezben).

10.2.2. Tervezési tanácsok a jó hangminőség elérésére

A kötött szótáras rendszerek építőelemeit (például egy számfelolvasóét) úgy kell megtervezni, hogy a lehető legjobban biztosítsuk a természetes ejtésre jellemző folyamatosan változó (törések, ugrások nélküli) spektrumképet (formánsmozgásokat). Ugyanezt kell biztosítani a prozódiai szerkezetekre is (hangintenzitás, dallam, alaphangmagasság, időszerkezet). A fenti két folyamatossági kritérium főleg az összekapcsolási pontokra és azok közvetlen környezetére vonatkozik. Ezt gondos elembázis-tervezéssel lehet biztosítani. Az egyes üzenetelemek csatlakozó hangjainak (adott elem utolsó és a hozzá csatlakozó elem első hangja) spektrális és prozódiai szerkezetét kell minél közelebb hozni a csatlakozási ponton, hogy az összekapcsoláskor ne legyen ugrás, csuklás a hangzásban. Ez gyakorlatilag azt jelenti, hogy az elembázisba eltárolandó üzenetelemeket úgynevezett vivőmondatok felolvastatásá-

ból származtatott hullámformából kell kivágni. A vivőmondat megfelelő kialakítása biztosítja a természetes hangzási környezetet az adott üzenetelemnek. A fenti példát követve a vivőmondat(ok) elemei a következők lehetnek (a félkövér részt kell az üzenet változó részeként kivágni):
Az Ön aktuális egyenlege:_ ***74 pont*** .
Az Ön aktuális egyenlege:_ ***75 pont*** .
Az Ön aktuális egyenlege:_ ***76 pont*** .
Annyi ilyen vivőmondatot kell felolvastatni, ahány szám előfordulhat. Az üzenetet minden esetben két részből kell összefűzni, a számtól függően. A második részhez is hasonló módszert kell alkalmazni. A vivőmondat itt a következő.
*A legutóbbi pontjóváírás_* ***a január havi adatok_*** *alapján történt.*
*A legutóbbi pontjóváírás_* ***a február havi adatok_*** *alapján történt.* Ilyen vivőmondatból 12-őt kell felvenni. Itt az üzenetet három részből kell összeállítani, a középső rész előtt és után lehet rövid szünetet beiktatni, ez nem fogja zavarni a kiejtés folyamatosságát. A fenti példából látható, hogy a változó üzenetrészeknél nem kell ragaszkodni az íráskép adta szerveződéshez (lásd fent, az eredeti mondatnál), inkább a hangzó forma alapján kell a stratégiát kialakítani. A válogató szoftvert a hangzáshoz kell igazítani, nem pedig az írásképhez (még akkor is, ha az üzenet aktuális változatát például elektronikusan kapja meg egy felolvasó automata, tehát azt emberi kéz nem érinti). A bonyolultabb üzenetszerkezettel rendelkező kötött szótáras rendszerek elembázisának tervezéséhez már mélyebb beszédakusztikai ismeretek is szükségesek (lásd a 8.2.1. fejezetben).

Akusztikai tanácsok az üzenetrészek vágási pontjainak megtervezéséhez. Erre már a szövegtervezés szintjén kell gondolni. A beszédhangok határai – mint már láttuk – csak bizonyos esetekben határozhatók meg egyértelműen. Ilyenek, amikor gerjesztésváltás van a két hang határán. Ha a vágási ponton zöngétlen zárhang következik a zöngés, kivágandó rész után, a zárhang néma fázisa miatt ideális vágási pont jön létre. Az ilyen pontoknál a beszédhullám intenzitása kicsi, tehát a vágás nem visz be torzítást az akusztikai tartalomba. Ezért a vivőmondatokat úgy kell megtervezni, hogy a változó üzenetrészek kezdete és vége olyan hangszerkezetű legyen, ami biztosítja az egyértelmű elvágási pont kijelölését. Nem lehet egyértelműen elvágni az olyan hullámformát, amelyben nagy energiájú zöngés hangok csatlakoznak egymáshoz a szóhatáron. A fenti példamondatot tekintve tehát nem célszerű úgy tervezni az összefűzendő elemeket, ahogy a 10.10. ábrán szerepel, hogy a hónap előtti névelő külön elemként van felvéve. A [j] hang magánhangzóhoz hasonló tulajdonságokkal rendelkezik és folyamatos ejtésben amplitúdó módosulás nélkül kapcsolódik a magánhangzóhoz. Ezért inkább a névelővel együtt kell kivágni az üzenetrészt a vivőmondatból (ahogy javasoltuk). A gondos tervezés eredménye, hogy a természetes ejtéshez közel álló üzenetek hangozhatnak fel az ilyen beszédszintetizátorok kimenetén. Erre mutat példát a 10.11. ábra. A helyes megoldás a következő. 1 – *Az Ön aktuális egyenlege;* 2 – *hetvennégy pont. ;* 3 – *A legutóbbi pontjóváírás;* 4 – *a ja-*



10.11. ábra. A példamondat rezgésképe (fent) és dallammenete (lent) fonetikai elvű tervezéssel megvalósítva

nuár havi adatok 5 – *alapján történt.* Az elemek közé beiktatott három ritmikai szünetet pontosvesszővel választottuk el. Az ábrából egyértelműen látható, hogy mi a különbség a rossz megoldás és a gondos tervezés eredménye között.

10.2.3. A kötött szótáras rendszerek tervezési folyamata

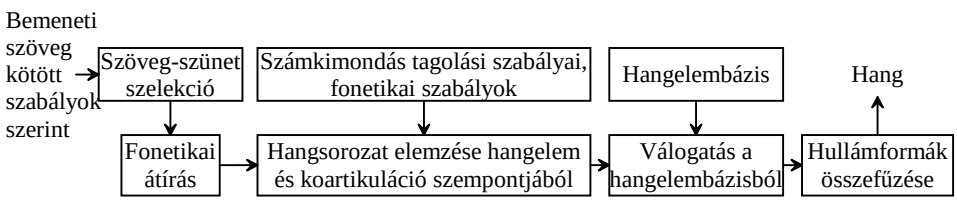
A tervezés menete:
1 – A tematika meghatározása;
2 – Az elmondandó üzenetek meghatározása, optimalizálás a változó tartalom szempontjából (lehetőleg ne az üzenet közepén legyen, hanem a végén);
3 – Az elmondandó üzenetek megtervezése (a megszólaló üzenetelem nem egyenlő a felolvasandó szöveggel, az üzenetelemet vivőmondatba kell elhelyezni és azt a mondatot kell felolvastatni);
4 – A hullámforma-elemtár tartalmának megtervezése. Azt kell meghatározni, hogy a kimondandó üzenetet mely üzenetrészekből állítjuk össze. Ez a munka szorosan kapcsolódik a 2-es ponthoz;
5 – A felolvasandó szöveg megtervezése;
6 – A bemondó kiválasztása;
7 – Hangfelvétel készítése;
8 – Digitalizálás, az üzenetelemek kivágása a vivőmondatokból;
9 – A kivágott hullámformákat tartalmazó adatbázis elkészítése;
10 – Összefűzési algoritmus tervezése, készítése, tesztelése.

A kötött szótáras rendszerek előnye, hogy egyszerű szerkezetűek, különösen, ha nincs változó üzeneteleme; jó hangminősége lehet, ha gondosan tervezik. Hátrányai: kötött témakörhöz lehet csak alkalmazni; nagy üzenetszám esetén már nehéz karbantartani; új üzenet létrehozása bonyolult (ugyanaz a bemondó kell, stúdió, hangfelvétel, biztosítani kell ugyanazt a hangszínt és hanghordozást); ha sok változó üzenetrész van az üzenetekben, nehéz megtervezni a beszédelemek tartalmazó elembázist,

valamint a válogató és összefűző algoritmust; nagy elemszámú rendszereknél szakértő kell a tervezéshez, egyébként rossz hangminőségű lesz a rendszer.

10.2.4. Fonetikai elvű modell szám-, dátum-, időpont-, pénzösszeg-felolvasáshoz

A számok fontos szerepet töltenek be az infokommunikációban. Az itt ismertetett beszélő program természetes, emberi hangon képes számokat, időpontokat, dátumokat, pénzösszegeket felolvasni olyan hangminőséggel, hogy az összetéveszthető egy természetes személy beszédével. A hang hordozza a beszélő hangszínezetét is. Bárkinek a hangjára elkészíthető. A programot a BME Távközlési és Médiainformatikai Tanszékén fejlesztették ki. Az eljárás magyar szabadalom (Olaszy–Németh 1996), lajstromszáma P 9601427. Működési elvét a 10.12. ábra mutatja. A felolvasórend-



10.12. ábra. A számokat tartalmazó üzenetek gépi felolvasására kialakított, kötött szótáras, fonetikai elvű beszédszintetizátor működési egységei

szer szókészlete a következő:

- szám (1 milliárdig),
- euró, dollár, cent, forint,
- egész, tized, század,
- százalék, ezrelék,
- mínusz, plusz,
- óra, perc, másodperc,
- órától, óráig
- perctől, percig
- január, február, március, ….. december
- hétfő, kedd, szerda, csütörtök, péntek, szombat, vasárnap
- szünetelemek (nincs hang), különböző hosszúságban

A felolvasó által adott üzenetet minden esetben kombinálják fix, előre felvett hanggal (dőlten szedve a példákban).

Az ön számlájának egyenlege: 164 523 Ft
Az euró eladási árfolyama: 274,84 Ft

A helység postai irányítószáma: 1221 (de kimondható tagoltan is: 12-21)
A szolgáltatás ára: 13 euró, 24 cent
A változás nagysága és iránya: mínusz 12,32 század százalék
A vásárlás dátuma: 2003. december 23.
Az indulási idő: 16 óra 23 perc
A nyitvatartási idő: Hétfő: 8 órától 16 óráig. Kedd–péntek: 9 órától 17 óráig
A mérés időpontja: 1984. december 29., 13 óra 25 perc, 12 másodperc
A telefonszám: +36-30-123-45-67
A mellék száma: 12
A mai lottószámok: 1, 34, 42, 67, 89
A beruházás teljes ára: 993 456 789 euró

A rendszer a hang előállításához a 8.2.1. fejezetben ismertetett elvek szerint elkészített hangelembázist használja. Ebben kétfajta dallammenetet reprezentáló hangelemek találhatók: az ereszkedő és a szinttartó. Az ereszkedő a kijelentés egészére vonatkozik (nincs benne szünet), a szinttartót pedig az üzenet belsejében lehet használni, olyan helyeken, ahol érzékeltetni kell, hogy a kimondás még nem fejeződött be. Tipikusan ilyen eset, amikor telefonszámokat kell felolvasni, és 3-2-2-es egységekben mondjuk ki a 7 jegyű számot. Az egységek között szinttartó, a kimondás végén ereszkedő hanglejtést kell használni, akkor lesz természetes a hangzás. A felolvasó bemenete szöveg, amelyet előre meghatározott, kötött formában kell a szintetizátor bemenetére adni. A szintézis során hullámforma-elemeket kell kiválogatni a hangelembázisból és ezeket kell összefűzni. Az összefűzés után semmiféle jelfeldolgozást nem alkalmazunk (simítás, intenzitáskorrekció, dallamráültetés stb.), ezért nagyfokú hanghűséggel szól a program. A szöveg-hang konverzió vezérlését a válogató algoritmus végzi el öt lépésben:

1. Fonemikus alakra hozza a bemeneti szöveget és értelmi egységekre bontja az üzenetet. Ez utóbbi azt jelenti, hogy meghatározza az egybefüggően kiejtendő részeket és a szünettartás helyét. Ezzel a hanglejtési formák is definiálódnak.
2. Elkészíti a bemeneti szöveg fonetikai átiratát, azaz hangszimbólumokkal adja meg az egybefüggően ejtendő részeket (fontos, hogy nem a szavakat, hanem az egész részt). Így elkészül a szintetizálandó üzenet elméleti, írott hangképe, amely egybefüggő hangsorozat(ok)ból és szünetekből áll.
3. A hangsorozatok belső szerkezetének elemzése következik. Egyrészről meg kell határozni a számok kimondásából keletkező alapegységeket (Olaszi 2000), valamint azok kapcsolódásai pontjaira jellemző kategóriákat.
4. A számelemekhez hozzárendeljük a dallammeneteket. Ezzel meghatároztuk az elembázisból kiválogatandó elemeket, azok sorrendjét és darabszámát.
5. Kiemeljük az elembázisból a szükséges hullámformákat és összefűzzük őket.

Az öt lépést az alábbi példán keresztül mutatjuk be: 274,84 Ft

1. kettőszázhetvennégyegész; nyolcvannégyszázad; forint

2. kettőszá**szh**etvennégyegész; nyolcvanné**tysz**áza**t; f**orint
3. kett(ősz)á(szh)etvenné(gye)gész; nyolcva(nné)(tysz)ázat; forint
4. kettőszászhetvennégy (kezdettől ereszkedő), egész(szinttartó); nyolcvannéty (belsejében enyhén ereszkedő), százat (ereszkedő); forint (befejező ereszkedő)
5. 1kettő(sz); 2(ö)szász(h); 3(sz)hetven(n); 4(n)néty(e); 5(ty)egész(szinttartó); 6szünetelem; 7nyolcvan(n); 8(n)éty(sz); 9(sz)ázat(sz) (ereszkedő); 10(t)forint

A szám kimondatásához, a szintézishez tehát összesen 10 elemet kell összekapcsolni. A jó hangminőség lelke a precízen összeállított és akusztikailag csiszolt hangelembázis. Ez a rendszer működik számos telekommunikációs szolgáltatásban (telefonszám-alapú tudakozó, árlista-felolvasó, név- és címfelolvasó, fióknyitvatartási tájékoztató).

10.3. Automatikus szövegfelolvasás

Olaszy Gábor

A jó minőségű gépi szövegfelolvasás bonyolult infokommunikációs feladat. Azt kell mondanunk, hogy a szolgáltatók általában rákényszerítik a beszédinformációs rendszereiket (amik a legtöbb esetben komoly minőségi kifogásokkal illethetők) a felhasználókra, nem fordítanak sem energiát, sem pénzt a minőségi megoldásokra. Az ügyfél kénytelen azt használni, amit kap (vásárlási információ, bankszámlaegyenleg, mérőóraállás, hírolvasó stb). Ez abból is fakad, hogy mindannyian tudunk beszélni és azt gondoljuk, hogy ennek gépi megvalósítása is egyszerű. Ennek a könyvnek azonban mégis az a feladata, hogy a minőségi megoldásokat ismertesse és iránymutatást adjon, hogy milyen felkészültség szükséges az ilyen rendszerek tervezéséhez, integrálásához, üzemeltetéséhez.

Az első kérdés a hibaszázalék szintje. A szövegek gépi felolvastatásánál a felhasználónak, szolgáltatónak el kell döntenie, hogy milyen mértékű hibaszázalékot enged meg a kiejtésben és azt is, hogy a hibák mennyire lényeges részei az információközlésnek (például rosszul ejti a címzett nevét). Hibátlan kiejtéssel rendelkező, általános gépi szövegfelolvasó nincs. Ezt a kritériumot az ember sem tudja teljesíteni. Ha gépi felolvasásban például minden tizedik mondatban (üzenetben) egy kiejtési, ritmikai, intonációs hiba van, akkor ez elviselhető a felhasználó számára. Ha minden mondatban 2–3 ilyen hibát vét a gép, akkor kifogásolható a felolvasás minősége.

A második kérdéskör a hangszínezet. A robotos, monoton hangú felolvasókat nehezen viseli el huzamosabb időn keresztül a felhasználó. Idetartozik a hangváltás is (hány hangon tud beszélni a rendszer).

A harmadik kérdéskör a műszaki megvalósítás és a beágyazási kritériumok közötti összhang megteremtése. Egy szerveralapú rendszer jobban elviseli a nagy me-

móriaigényt, mint egy mobiltelefon vagy egy célhardver, amelyben mikrokontroller működik.

Az előbbiekből mind következik, hogy a gépi szövegfelolvasó rendszerek gyakorlati alkalmazásainál sok kompromisszummal kell számolnia mind a megrendelőnek, mind a fejlesztőnek, mind az integrálónak.

10.3.1. A beszéd modellezése szintézishez

A gépi szövegfelolvasás sikeres megoldásához a beszédet valamilyen szinten modellezni kell. Ennek két iránya van. A hagyományos módszernél a modellezés megfigyelésekre, mérési eredményekre alapul. Ezek a modellek szabályrendszereket tartalmaznak, ez alapján közelítik a természetes jelenség tulajdonságait. Az ilyen modellek belső szerkezete ismert, belső állapotaikat a fejlesztő a szabályok megváltoztatásával módosítani tudja. Hasonló a helyzet a hibakeresést illetően, ha a modell rosszul teljesít, akkor általában pontosan meghatározható, hogy mely szabály okozza a hibát. A hibaelhárítás tehát egyértelműen megoldható. Ha jó modelleket készítünk, akkor a géppel előállított beszéd jól fogja megközelíteni a természetes beszédet. A modellezés másik iránya statisztikai eljárásokon alapul. A gépi tanuló és statisztikai elvű válogató algoritmusokkal nagy beszédkorpuszokon végzett mérésekkel kialakíthatók modellek. Az ilyen statisztikai modellekre az jellemző, hogy a belső működésüket csak bonyolult módon lehet ellenőrizni, követni, hangolni, a modellből eredő hibák oka nem határozható meg egyértelműen. Mindkét eljárásra találunk példákat a következő fejezetekben.

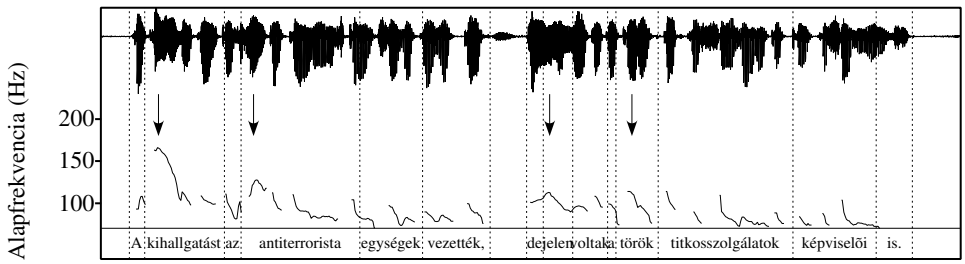
10.3.1.1. Hangsúlymeghatározás a szöveg alapján

A gépi szövegfelolvasás során a mondatok prozódiája akkor lehet természetes, ha képesek vagyunk egyrészről a mondat dallamának, másrészről a szavak hangsúlyozásának, azaz a prozódiát alkotó két leglényegesebb összetevőnek a szövegbe való jelzésszintű beillesztésére (Koutny 2008). Ebben a fejezetben a hangsúlyokkal, a következőben pedig a dallamelemekkel foglalkozunk. Az utóbbi évtizedek leíró jellegű mondattani, fonológiai, valamint kísérleti fonetikai eredményei alapján (É. Kiss et al. 1998, Hunyadi 1995b, Varga 2002, Olaszy 2002) jó eséllyel meg tudjuk jósolni, hogy egy mondatban milyen lesz – vagy lehet – a hangsúlyok eloszlása. A gépi, szövegalapú hangsúlyelemzés két modelljét ismertetjük a magyarra vonatkozóan, azzal a kitétellel, hogy egyik sem ad teljes körű megoldást (ez a jövő feladatai közé tartozik). A modellek gyökeresen eltérnek egymástól, az első (és korábbi) felszíni szövegelemzésen és statisztikai vizsgálatokon alapul (Olaszy 1996), a második

szintaktikai elemzéssel állapítja meg a hangsúlyok helyét (Tamm–Olaszy 2005). Mindkét modellben a hangsúlykijelölés vizsgálati tere a mondat. A két módszert összehasonlítottuk, a hibázási százalékuk egyforma, a vétett hibák helye a mondaton belül módszerfüggő.

Célkitűzés. Meg kell határozni a normál ortografikus szöveg elemzésével, hogy várhatóan mely szavak lesznek a kiejtés során hangsúlyosak, és ezeket jelölni kell a szövegben. Az algoritmizálásból fakadó kiindulási elv, hogy a mondat minden szavára teszünk hangsúlyjelzést (Olaszy 2001a). Így egyértelmű lesz a jelölési rendszer és annak értékelése. Ötféle hangsúly kategóriát használunk a szavak besorolására. Ezek jelzései és tartalmuk a következő: [:F] = fókusz (a legerősebb), [:E] = kiemelt, [:W] = normál, [:N] = neutrális (hangsúlytalan), mínusz jel [:-] = erősen hangsúlytalan, (esetleg redukált). A fenti kategóriákból két hangsúlycsoport következik: azok a szavak, amelyeken van valamilyen hangsúly, és azok, amelyeken nincs hangsúly. A hangsúly jelzését a szó elé tesszük az elemzés során. Így a mondat minden szaván lesz jelzés. A hangsúlyjelekhez a fizikai megvalósítás során konkrét számadatokat rendelünk, amelyek alapján az alapfrekvenciát és hangintenzitást megváltoztatjuk (lásd a ProfiVox rendszert a 10.3.6.1. fejezetben).

Hangsúlykijelölés nyelvészeti elemzés nélkül. Ennél a módszernél a hipotézis az volt, hogy a szóhangsúlyozási fokozatok jó hatásfokkal meghatározhatók és jelölhetők a szövegben szintaktikai elemzésnél egyszerűbb módszerrel is, úgynevezett felszíni szövegelemzéssel. Ez volt az első modell, amely magyar mondatok elemzéséből meghatározta, hogy egy mondatban a szavak milyen hangsúly kategóriába sorolhatók a szöveg függvényében. A felszíni szövegelemzés azt jelenti, hogy nem vizsgáljuk a szófaji és grammatikai viszonyokat, a modell csupán egyszerű, empirikus szabályokra, a szerkezeti tagolásra (például központozás) és listákra (válogatott szavak, szókapcsolatok) támaszkodik. A modell kialakításához korábbi, felolvasott (valós) szövegek hanganyagain végeztünk fonetikai elemzéseket (hírek, ügyfélszolgálati tájékoztatók). Az alapfrekvencia kiemelkedéseit és a szöveg közötti összefüggéseket vizsgáltuk. Külön elemeztük a mondatkezdési szókapcsolatokat, a szövegben előforduló központozási jelek környékét és a mondatok befejezési fázisait. Egy mondat elemzését mutatja a 10.13. ábra. Az ábra alapján az elemzett mondatban a szavakra a következő hangsúlyjelöléseket lehet tenni. [:-]*A* [:W] *kihallgatást* [:-]*az* [:W] *antiterrorista* [:N]*egységek* [:N]*vezették* , [:-]*de* [:W]*jelen* [:N] *voltak* [:-]*a* [:W]*török* [:N] *titkosszolgálatok* [:N]*képviselői* [:N] *is*. Ennél a modellnél két dolgot kell kiemelni. Az egyik, hogy nem törekszünk teljességre. Elvünk az, hogy a hangsúlyos szavak többségét megtaláljuk a mondatban, továbbá az, hogy lehetőleg ne tegyünk hangsúlyt olyan szavakra, amelyek hangsúlytalanok a kiejtésben. A másik fontos szempont, hogy erősen támaszkodunk a gyakorisági adatokra. Ez azt jelenti, hogy a gyakoribb eseteket vesszük szabálynak, a szabály alóli kivételeket



10.13. ábra. A hangsúlyos szavak megjelölése a szövegben az alapfrekvencia-görbe (alul) szövegre való visszavetítésével. A nyilak jelzik a hangsúlyokat a szavak első szótagján, a függőleges vonalak a szóhatárokat

pedig esetlegesen listákban vagy magában a szabályban adjuk meg. Az eljárásban tehát gyakran kell kompromisszumot kötni. A szöveg felszíni vizsgálatában a szavakat két kategóriára osztjuk: hangsúllyal ellátható **tartalmas** szavak, illetve **nem értékes** szavak. Ez utóbbiak azok, amelyekre nem kerülhet hangsúly, vagyis az N, illetve mínusz jellel jelzett szavak (ilyenek például a névelők, a kötőszók). Mindkét kategóriához tartoznak kivételek. Például, amikor mutató névmás és névelő kerül egymás mellé ([:W]*Az* [:-]*a* [:N]*labda* [:N]*kell* [:N]*nekem*. A vizsgálatban fontos szerepet tulajdonítunk a mondatban elhelyezett szeparátoroknak (vessző, pontosvessző, kettőspont stb.). A több szóból álló, de fogalmilag összetartozó kifejezéseket egy szónak tekintettük (csak a jelzett szó kaphat hangsúly jelet, a többi [:N]-t). Ezeket gyűjtött lista alapján detektáltuk (például: *hívásforgalmi tájékoztató*). A hangsúly-meghatározási eljárás három lépcsős. Az első lépcsőben a szavakra egyenként megállapított jelöléseket helyezzük el, a másodikban a szókapcsolatokat is vizsgáljuk, és az összetartozási hangsúlyozási szabályok szerint változtatunk. Így alakul ki a végleges hangsúlytérkép a mondatban. A harmadik lépcsőben az eddig nem jelölt szavakra [:N] jelet teszünk.

A [:F] fókusz jelű hangsúly a legerősebb a hangsúlyozott szavak között. Ezen szavakat lista alapján (ami természetesen erősen korlátozott a valósághoz képes) jelöljük (például: *nem, ne, nagyon, nincs, soha, semmi, senki, jó, szép, minden, meg kell, mikor, milyen stb.).* Fontos szempont, hogy az [:F] jelű szavak után az irtó szabályhoz hasonló műveletet hajtunk végre. Ez azt jelenti, hogy ha esetleg a hangsúlykiosztás első fázisában egy [:F] jelű szó utáni szó [:W] jelet kapott, akkor azt törölni kell és [:N]-re kell változtatni. Ugyanilyen szabály vonatkozik az [:F]-jelű szó előtti szóra is. További szabály, hogy [:F] jelzés két egymást követő szón nem lehet. Ha ilyen előfordul, akkor mindig az első tartja meg az [:F] jelzést, a következő [:N]-re íródik át.

Az [:E] és [:W] jelzésű szavaknál a hangsúlyozást megvalósító F_0 kiemelkedés mértékében van csupán különbség. Az előbbi erősebb hangsúlyt képvisel, mint az utóbbi. A normál szóhangsúlynak a [:W] jelzést tekintjük. Az [:E] jelzést szintén

lista alapján osztjuk ki. A [:W] jelzés meghatározásához listát is és szabályokat is alkalmazunk. Ezen szavak kijelölésének a legbonyolultabb a szabályrendszere, ezek empirikus szabályok. A [:W] szóhangsúly jellel jelöljük meg a listában megadott szavakat, a számok elemeit a *száz, ezer, millió* kivételével, a névelők (*a, az*) utáni szót, a vessző utáni első tartalmas szót, a mondat első tartalmas szavát, a [:-] jelzésű szavak utáni tartalmas szót, bizonyos szóösszetételek meghatározott szavait (listából), a tulajdonneveket (nagy betűvel kezdődnek), a személyek nevét és a mozaikszavak feloldott betűsorait (például MTA).

Az [:N] jelű szavak jelzésének fizikai jelentése az, hogy a hangsúlyozandó szón nem hajtunk végre sem F_0-, sem intenzitásemelést. A szó a kiejtés szempontjából tehát neutrális (nem hangsúlyos), csak a mondatdallamból fakadó alapfrekvencia-változás lehet benne. Az [:N] jelű szavak kijelölését az idetartozó szabályok és lista szerint, valamint a maradék kitöltése elv alapján végezzük. Ezek a szabályok átírásra vonatkoznak. Ez azt jelenti, hogy a korábbi szinten megvalósított jelölést ennek a szintnek a szabálya felülírhatja. Az [:F] jelű szó után következőre [:N] jelzés kerül, ha más volt rajta. A mondat első tartalmas szava [:N] jelzést kap, ha a második szón [:F], vagy [:E] jelzés van. A listák alapján a megadott szókapcsolatra jellemző hangsúlymintát alkalmazzuk. Miután minden jelzést elhelyeztünk a mondatban és a jelzések véglegesítése is megtörtént, akkor az addig nem jelölt szavakra [:N] jelzést teszünk.

A [:-] jelű szavak jelzésének fizikai jelentése az, hogy a hangsúlyozandó szón csökkentjük az F_0 értékét, az intenzitását, valamit a hangok időtartamát. Így egy általános akusztikai redukciót hajtunk végre a szó hangsorban elfoglalt szerepe szempontjából. A [:-] jelzésű szavakat listából jelöljük ki. Talán ezek a szavak rendelkeznek a legstabilabban a [:-] jelzéssel, mint a hangsúlyozási hierarchia legalsó elemei. Kivételek természetesen itt is vannak. A modell minden hangsúly-

10.2. táblázat. Példák a hangsúlyjelzések megadására szólistában

[:F]	[:E]	[:W]	[:N]	[:-]
nem	minden…	bármi…	szerint	a, az, is, és, hogy
nincs	biztos(an)	gondos…	szerű	már, nélkül, …
nagyon	soha	ugyan…	tekintetben	mint, amint, …
senki	szintén	gyakran	ugyanis	ami.. (+ toldalék)
különös(en)	sok, soka(t,n,k)	szép	mindazonáltal	talán, ahogy, ahogyan
jó, jól	mégis, mégsem	gyönyörű	majd	fog, lesz, ha, de, illetve

kategóriájához tartoznak listák. A listák szótárszerűen adják meg a szövegelemet és a rájuk elhelyezendő hangsúly jelzést. A szótár elemei lehetnek szavak (10.2. táblázat) és szófüzérek (10.3. táblázat). A szó- és szófüzérlisták mintegy 1200 tételt tartalmaznak ebben a modellben. Példamondatokon mutatjuk be a címkézés eredményét.

10.3. táblázat. Példák az egybetartozó szófüzér hangsúlyjelzéseinek megadására

szófüzér	szófüzér hangsúlykiosztással
az jellemz..	[:W] az [:N] jellemz...
az (a, is, lesz, volt)	[:W] az [:N](a, is, lesz, volt)
mind (a, az, pedig)	[:W] mind [:N](a, az, pedig)
senki (sem, nem)	[:W] senki [:N](sem, nem)
semi (sem, nem)	[:W] semmi [:N](sem, nem)
sor kerül(t, het)	[:W] sor [:N]kerül(t, het)
vezetékes telefon	[:W] vezetékes [:N telefon
faxszolgáltatás	[:W] fax [:N szolgáltatás
bármely napszakban	[:W] bármely [:N napszakban
hagyatéki eljárás	[:W] hagyatéki [:N eljárás
egymást követő	[:W] egymást [:N követő

[:-]*A* [:W]*szakemberek* [:W] *eddig* [:-]is [:W]*úgy* [:N]*látták:* [:-]*a* [:W]*tankönyvek* [:N] *terén* [:W]*ugyanaz* [:-]*a* [:N] *helyzet*, [:-]*mint* [:W]*huszonöt* [:N]*éve*.

[:-]*A* [:W]*visszaélések* [:N] *kiküszöbölésére* [:W]*ki* [:N]*kellene* [:N]*kényszeríteni* [:-]*az* [:W] *európai* [:N]*sajtóerkölcs* [:N] *érvényesülését*, [:N]*főleg* [:-]*a* [:W]*tisztességes* [:N]*verseny* [:-] *és* [:-]*a* [:W]*kiskorúak* [:N]*védelmében*.

[:-]*A* [:W]*gazdaság* [:N] *fellendülése* [:N]*vonatkozásában* [:W]*az* [:N]*az* [:W]*első* [:N] *kérdés,* [:-]*hogy* [:W]*lesz-e* [:N] *adócsökkentés?*

[:W]*Az* [:-]*a* [:F] *legfontosabb,* [:-]*hogy* [:W]*gyermekeinket* [:N]*megfelelő* [:N]*óvodába és* [:W] *iskolába* [:N]*tudjuk*[:N] *járatni.*

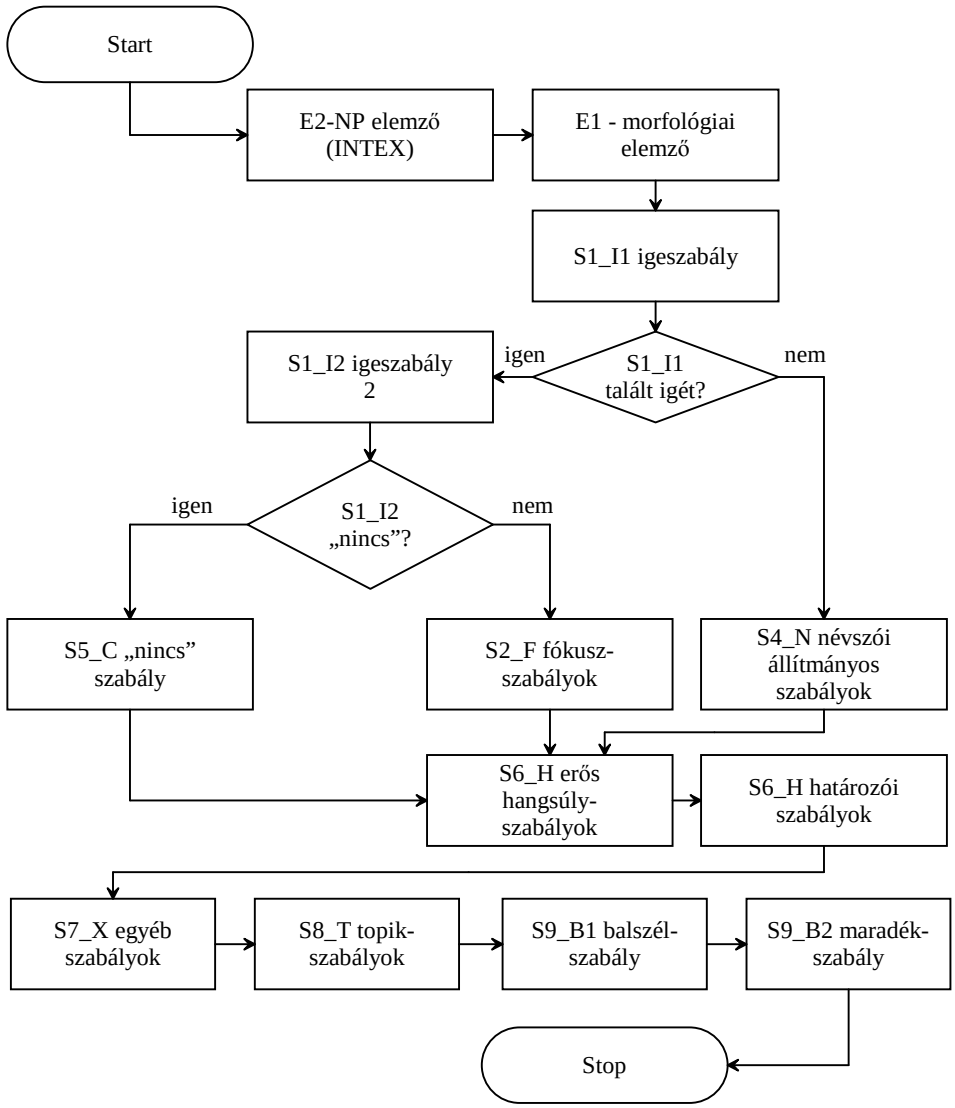
[:F]*Hogyan* [:N]*tervezi* [:N] *a* [:N]*jövőjét* [:N]*az* [:N] *Opera* [:N]*nélkül?*

A szabályok száma: 390. Az alkalmazott listák 1645 elemet tartalmaznak. A szabályok kialakítása során főleg hírolvasást elemeztünk. A modell a ProfiVox magyar szövegfelolvasó szoftverbe került beépítésre a BME Távközlési és Telematikai Tanszékén (Olaszy et al. 2000b). Az elemzés hibázási aránya 12,8 százalékos (Tamm–Olaszy 2005). Ebből mindössze 3 százalékban fordult elő, hogy olyan szóra tett hangsúlyt, amelynek hangsúlytalannak kellett volna lenni. A modell előnye, hogy egyszerű, kis számítástechnikai erőforrást igényel, programba beágyazva gyors a futása. Ezért különösen alkalmas beszédtechnológiai alkalmazásra.

Hangsúlykijelölés szintaktikai elemzés alapján. A klasszikus hangsúlyelemzés nyelvészeti módszerekre épül. Korábbi kutatásokra alapozva (É. Kiss et al. 1998, Koutny et al. 2000), az elméleti megközelítés a következő. A magyar mondatoknak egy lényegében invariáns hierarchikus szerkezetet tulajdonítunk, s ebből próbáljuk levezetni a prozódiai szerkezet hangsúlykomponenseit. A mondat topikrészre és

predikátumrészre oszlik. A topik tetszés szerinti (nulla, egy vagy több) igebővítményt és szabad határozót tartalmaz. Bizonyos típusú összetevők (mondat típusú határozók *szerencsére, valószínűleg, látszólag* stb.) csak a topikrészben állhatnak. A topikrész összetevői mind gyenge hangsúlyt viselnek. A mondat legerősebb hangsúlya a predikátumrész első fő összetevőjére esik. A predikátumrész tetszés szerinti és számú (nulla, egy vagy több) disztributív kvantorral (*mindenki, senki, minden előfizető, a posta is*) kezdődik. Ezek mindegyike főhangsúlyos. Őket követi a szintén főhangsúlyos, közvetlenül az ige előtti összetevő, mely akár fókusz (*A **postás** csengetett be*), akár igekötő (***be**csengetett*), akár névelőtlen főnév (***levelet** hozott*) lehet. Az ezt követő ige hangsúlytalan. Bizonyos mondattípusokban az ige előtti pozíciók üresen maradnak, és maga az ige a predikátumrész kezdete: ilyen esetben az ige főhangsúlyos. A főhangsúlyos elemek hangsúlyának erőssége balról jobbra csökken. Az ige utáni fő összetevők attól függően hangsúlyosak, hogy ismert vagy új információt közölnek-e, és hogy van-e fókusz a mondatban. Az ige utáni disztributív kvantorok akár hangsúlyosak, akár hangsúlytalanok lehetnek. A fenti fő elvek alapján egy humán szakértő minden lehetséges magyar mondathoz hangsúlyszerkezetet tud rendelni. Ugyanakkor az automatikus elemzést megnehezíti, hogy sok elméleti fogalmat (új információ stb.) nem lehet algoritmikusan leírni (ez a jövő feladata lesz). További nehézség, hogy a magyar mondatban lényegében minden mondatpozíció maradhat üresen is, ezenkívül az igét és az ige előtti pozíciót kivéve a szerkezeti pozíciók több fő összetevővel is kitölthetők. A szerkezeti pozíciók azonosításában tehát nem segít a számolás; irreleváns, hogy egy összetevő hányadik helyen áll. Nem mindig könnyű feladat a fő összetevők határainak automatikus felismerése sem. Mindezek tudatában alakítottuk ki a nyelvészeti elemzésre épített hangsúly-meghatározó modellt (Tamm–Olaszy 2005), amely követi a korábban említett hangsúlybesorolás öt kategóriáját. A munka az MTA Nyelvtudományi Intézet és a BME TMIT közös fejlesztése volt. Egyelőre működő rendszerben nem alkalmazzák. A célkitűzés megvalósítására az eddig fejlesztett számítógépes nyelvészeti eszköztárakból elemző algoritmusokat, az elméleti kutatásokból szabályokat és listákat használunk. Algoritmusok: morfológiai elemző (Kiss–Bárkányi 2006); INTEX tartalomelemző (Silberztein 1993). Szabályok: fókuszszabályok (azon belül azonosító szabályok és hangsúlytörlő szabályok); egyéb erős hangsúlyt adó szabályok és környezetük; topikszabályok; határozói szabályok; a szintaktikai egységeken belül működő balszélszabályok és egyéb szabályok (például: szöveg vagy mondat típusától függő szabályok). Listák: 18-féle szólistát határoztunk meg (szavak, kifejezések). Az elemzés algoritmusát a 10.14. ábra mutatja.

Tartalomelemző. Az INTEX magyarra adaptált változata két feladatot lát el: egyrészt megjelöli a mondatban található főnévi csoportok (NP), azaz egy főnévből vagy egy főnévből és több szóból álló mondattani egységek, a melléknévi csoportok (AdjP) és a névutós kifejezések kezdetét és végét (a beágyazottakét is), másrészt



10.14. ábra. Nyelvészeti alapú mondatelemző algoritmus a hangsúlyok megállapítására

megjelöli a mondat- és tagmondathatárokat (Olaszy 2007b). Ha az NP elemző talál egy főnevet (névszót), akkor megállapítja ennek a közvetlen környezetéhez való viszonyát, és ezután dönti el, hogy az NP-hez tartozik ez a környezet. A főnévi csoport határait azért releváns megkeresni, mert a frázisra adott hangsúly egy főnévi csoportban csak az első tartalmas szóra esik. A többi szó az NP-ben semleges hangsúlyt kap vagy azt a hangsúlyjelölést, amelyet a listák alapján előírunk neki. A névelőtlen főnevek helye a ragozott igéhez képest viszont fontos az ige utáni frázisok hangsúlyozásában. Az NP keresés számos esetben bizonytalan lehet. Például függhet a szöveg tartalmától és szerkezetétől. Konkrét méréseket végeztek ügyfélszolgálati szövegek elemzésére (Abari–Olaszy 2007). Ehhez az elemző további tanítására volt szükség. Felsorolunk néhány ilyent. A feldolgozott témakör mondatai nagy arányban tartalmaznak gondolatjeles közbevetéseket, valamint számos, nyílt tokenosztályba tartozó entitást (telefonszámok, egy- vagy többszavas márkanevek, kódok), melyeket nem lehet a szótárban kezelni, így helyes címkézésükről reguláris nyelvtanokkal kell gondoskodni. A nyílt tokenosztályba tartozó entitások kezelésekor a felismerésen túl az NP-nyelvtant adaptálni kell azokhoz (például az idegen szó mint főnévi fej esetén gondoskodni kell a fejhez kötőjellel csatolt esetrag felismeréséről is). Az NP elemzés végeredménye erősen befolyásolja a további modulok helyes döntéseit, ezért a rátanítás fontos.

Morfológiai elemző. Morfológiai elemzőnek a Szószablya (Halácsy et al. 2003) programot használtuk, amely minden szóra megállapítja annak morfológiai alakját. Fontos az ige, a létige, az igekötő, a határozószó, a kötőszó és néhány esetben a szótő megállapítása. A morfológiai elemző elengedhetetlen a fókuszos mondatok elemzésében, ahol a ragozott ige és az igerészek helyétől függően a mondatban több frázison is törlődik a hangsúly.

Igeazonosító szabály. Az egyik legfontosabb szabály az igeszabály, amely egy ragozott igét keres. Kétfajta kimenetet ad (talált ilyet, illetve nem). Ha a program talál egy ragozott igealakot, akkor további vizsgálatokat kell végezni.

Igeszabály. A szabály célja annak megállapítása, hogy az ige egy egyszerű ragozott ige vagy a *nincs, nincsenek* szó.

Tagadásszabályok. A szabály a *nincs, nincsenek* szót [:F] címkével látja el, az őt követő szavakat [:N]-re címkézi a mondat/mondatrész végéig.

Fókuszszabályok. A szabályok feladata a mondatban rejlő fókusz azonosítása és címkézése. A fókusz a legerősebb hangsúly a mondatban. A fókusz lehet egy vagy több szóból álló csoport (egy frázis). A fókuszszabályok két nagy részből

állnak, a fókuszkereső szabályokból (ebből több van) és a hangsúlytörlő szabályból. A fókuszt kereső szabályok összetettek.

Az F-1 fókuszszabály, a névelőtlen főnév, igekötő vagy azzal azonos státuszú igerész előfordulásnál alkalmazható. Gyakran van egy mondatban egy hátravetett igerész, igekötő vagy névelőtlen főnév. Ha névelőtlen főnév, igekötő vagy azzal azonos státusú igerész közvetlenül az ige után helyezkedik el, akkor az a frázis, amelyik az ige előtt van, fókusznak tekinthető. Például:
A gazdagréti bankfiókba valamivel [:F]*10 óra előtt rontott be a símaszkos rabló.*
A példában a szabály megtalálja a ragozott, igekötős ige (*berontott*) igekötőjét az ige után, ezért az a frázis, ami az ige előtt van, fókusz (*valamivel 10 óra előtt*). Az [:F] jelzés helyes elhelyezését majd a későbbi balszélszabály fogja kijelölni. Gyakran nincs a mondatban hátravetett igerész, igekötő vagy névelőtlen főnév, akkor nem tudjuk helyesen megállapítani az ige előtti és utáni hangsúlyeloszlást. Hosszabb, összetett mondatokban ez viszont lényeges.

Az F-2 fókuszszabály szerint, ha a létige bővítménye közvetlen az ige után helyezkedik el, akkor az a frázis, ami az ige előtt helyezkedik el, fókusz.

Az F-3 fókuszszabály szerint, ha a negatívan minősítő határozószó áll közvetlenül az ige előtt, akkor ez a határozószó fókusz. A negatívan minősítő szavak kevésre értékelt számosságú vagy kevésre értékelt mennyiségű dolgot, kis gyakoriságot, kis fokot, mértéket, vagy kevésre értékelt módot jelölnek (vö. É. Kiss et al. 1998, 48. o.). Ilyen elemek például: *rossz, kevés, ritka, kevéssé, ritkán, rosszul*. Ugyanezzel a szabállyal lehet kijelölni a problémás névszói állítmányos mondatokban vagy mondatrészekben is a fókuszt. Például:

A magas hőmérséklet miatt a lehullott csapadék hamar elolvad, így [:F]*kevés az esély a fehér karácsonyra.*

A negyedik fókuszszabály akkor lép be, ha van egy frázis a kvantorok és az ige között és az is fókusz. Például:

A hajnali pára és köd feloszlását követően ma is [:F] *sok lesz a napsütés*.

Az ötödik fókuszszabályt akkor alkalmazzuk, ha egy frázis kezdetén a *csak* szó szerepel.

Ha semelyik fókuszkereső részprogram nem talált fókuszt, akkor fókuszhangsúlyt adó szabályokra nem kerül sor és tovább lehet lépni a nem fókuszos főhangsúly, az úgynevezett erős hangsúly ([:E]) keresésre. Ha megtaláltuk a fókuszt, akkor lehet tovább lépni a hangsúlyirtó szabályokra.

Erőshangsúlyszabályok. Az E-elem hangsúlya abban hasonlít a fókuszhangsúlyhoz, hogy erős. Két szempontból viszont különbözik a fókusztól, az egyik, hogy hangsúlyirtó hatása nincs, a másik, hogy egy mondatban vagy mondatrészben több E-elem is előfordulhat. A fókuszszabályokkal ellentétben itt fontos az alkalmazási sorrend is. Ha a szabályok nem találtak egyetlen E-elemet sem, csak akkor lehet az

igét E-elemként címkézni (azaz az ige lesz az E-elem az utolsó azonosító szabály szerint). Az erős hangsúly keresésére is több szabály vonatkozik.

Az első szabály azt állapítja meg, hogy ha a névelőtlen főnév (*orvos*), az igekötő (*meg-*) vagy azzal azonos státusú igerész (*haza-*) közvetlenül az ige előtt vagy az ige részeként helyezkedik el, akkor ez a frázis E-hangsúlyt kap ([:E]*orvos lett,* [:E]*megcsinálta,* [:E]*hazarendelték,* [:E]*szükség van*).

Az második szabály azt állapítja meg, hogy ha egy disztributív kvantor található a mondatban, akkor ez a frázis E-hangsúlyt kap (*A szomszédom* [:E]*mindig beteg*.).

A harmadik szabály az *is* szóra alapul. Ez disztributív kvantor és erős hangsúlyt kap (*A* [:E]*szomszédom is beteg.)*.

A negyedik szabály a tagadószóra alapul. Ez a szó erős hangsúlyt kap, de magába olvasztja a következő szót, ami azt jelenti, hogy az [:N] jelzést kap.

Az ötödik szabály szerint a létige ige előtti bővítmény kiegészítője is erős hangsúlyt kap ([:E]*szerencsés volt,* [:E]*orvos volt*).

Határozói szabályok. A szabályok feladata a mondatban rejlő határozók azonosítása. Ehhez 3 külön szemantikai-prozódiai leírással rendelkező listát használunk. Mondat- és módhatározókat különböztetünk meg. Ha a határozószó mondathatározó (*tényleg, esetleg, állítólag*, *okvetlenül, feltétlenül, tényleg*), akkor a mondat topikrészében van. Ez azt jelenti, hogy a lista alapján már el lehetne dönteni, hogy a határozószó a mondatkezdeti topikrészhez – nem a predikátumrészhez – tartozó szavak csoportjai között van, és semleges topikhangsúlyt kap vagy nem. Mondathatározó listákból kettő van: hangsúlyosak és hangsúly nélküliek. Ha például egy hangsúly nélküli mondathatározó a mondat első szava, akkor nem kapja meg a topik elejére előírt hangsúlyt. A módhatározók (*nagyon, eléggé, sokszor, állandóan*) és általában a többi határozó [:W] jelet kap.

Egyéb szabályok. Ezek a szabályok a nyelv speciális elemeire konkrétan hivatkoznak, többnyire listákon alapulnak. Számos ilyen szabályt kell alkalmazni a hatásos elemzéshez.

A kötőszavakon (*hogy, ami, de, hanem*) nem lehet hangsúly.

Az *egy* szó hangsúlyos, ha utána mértékegység jön (*fok, másodperc, perc, óra, nap, hét, hónap, év, évtized, évszázad, évezred, milliméter, centiméter, méter-sorozat, kilométer, láb, mérföld, gramm-sorozat, deka stb.* és ezeknek ragozott formái). A mértékegységeken viszont törlődik a hangsúly ([:N] lesz).

Ha van kis fokozatot jelölő összetevő: *néhány, némi, egy kicsi, néha, néhol, egyelőre, enyhén, kissé, némileg, valaki, valahol, valahogyan, valamennyi, némileg*, akkor törlődik a hangsúly ([:N] lesz).

A szemantikailag kiüresedett bővítmények esetében törlődik a hangsúly, [:N] lesz, (*bizonyos, valóságos, szegény, kis)*.

Ha címek és rangok vannak a tulajdonnevek előtt, akkor törlődik a hangsúly, [:N] lesz (*úr, néni, bácsi, út, köz, utca stb.)* .

Páros kötőszók esetén alkalmazható a [:W] jelzés (*nem ... hanem, akár ... akár, vagy .... vagy, mind ... mind*).

Ha a frázisban található egy listázott hangsúlykerülő (*akar, érint, fog, folyik, talál, kell, szabad, szeretnék*), akkor [:-] jelzésű hangsúlyt kell alkalmazni.

Egy vessző utáni mondásige (*mondta, döntött* stb.) a következő hangsúlymintát kapja: az ige és az igemódosító [:-] jelet, a többi frázisba [:W] + [:N] típusú hangsúlyminta kerül (a hangsúlyt a frázis első tartalmas szava kapja a balszélszabály szerint), a többi rész [:N] jelet kap. Ez akkor is érvényes, ha hátravetett igemódosító és főnévi csoportok vagy határozók állnak mögötte. Ha nincs fókusz vagy E-elem a mondatban, akkor is a frázisokon balszélszabály szerinti fenti hangsúlymintázat lesz a topik, a határozók és az erős hangsúlyú ige után.

Az egyéb szabályok alkalmazásánál gyakran számít a sorrendjük. Ezek a szabályok sok esetben korrigálhatják a listaalapú megközelítéssel kapott eredményt.

Topikszabály. A topik alatt technikailag az első E-elem vagy fókusz előtti szövegrészt értjük a mondatrészkezdet-jelzésig. A topikrészben a [:W] alkalmazható a mondat elején, a balszélen, utána [:N] jel kerül a frázisokra vagy a szavakra egészen a predikátumrész kezdetéig.

Balszélszabály. Minden frázison belül megkeresi azt a szót, amelyre a frázisnak adott hangsúly esik. A frázishoz hozzárendelt hangsúlytípus csak a frázis bal szélén marad meg: az első még nem jelölt szó kapja a frázis hangsúlyát, feltéve, hogy az nem névelő.

Default-szabály. A hangsúlyjelölés nélkül maradt szavakon az [:N]-típusú hangsúlyt alkalmazzuk.

Listák. A nyelvészeti hangsúly elemzéshez számos listát is fel kell használni.

- Egy szótagú funkciószók, klitikumok (Varga 2002): *a, az, egy, és, de, vagy, is, ha, én, ő, ez, már, még, csak*
- Mértékegységek: *fok, másodperc, perc, óra, nap, hét, hónap, év, évtized, évszázad, évezred, milliméter, centiméter, méter-sorozat, kilométer, láb, mérföld, gramm-sorozat, deka stb.*
- Számnevek: *egy, kettő stb.*
- Hangsúlykerülő, beférkőző igék: *akar, érint, fog, folyik, talál, kell, szabad, szeretnék stb.*
- Névmások: *én stb.*
- Névutók: *mellett, helyett, után stb.*

- Determinánsok: *a, az, e, ez, egy, eme, ama, ezen, azon, ez a, az a stb.*
- Névmásszerű főnevek, üres szavak: *ország, ügy, kormány, (M/m)agyarország(i) stb.*
- Vonzatszótár (opcionális)
- Negatív határozószók: fok, mód, gyakoriság: *csúnyán, rosszul, ritkán, kevésen, alig stb.*
- Pozitív értelmű határozószók: fok, mód, gyakoriság: *nagyon, eléggé, sokszor, állandóan stb.*
- Mondathatározók ([:N] jellel) *esetleg, állítólag stb.*
- Mondathatározók ([:W] jellel) *okvetlenül, feltétlenül, tényleg stb.*
- Kis fokozatot jelölő összetevők (Varga 2002, 141. o.) szerint, azaz *néhány, némi, egy kicsi, néha, néhol, egyelőre, enyhén, kissé, némileg, valaki, valahol, valahogyan, valamennyi stb.*
- Szemantikailag kiüresedett bővítmények (Varga 2002, 142. o.) szerint, azaz, balszélre kerülő melléknevek: *bizonyos, valóságos, szegény, kis stb.*
- Címek és rangok a tulajdonnevek előtt vagy után: *néni, bácsi, út, köz, utca, doktor stb.*
- Disztributív (univerzális) kvantorok: *mind*-sorozat *(mind, minden, mindenki, mindegyik, valamennyi, az összes, minden alkalommal, mindig stb.*
- Hangsúlyszabályokat befolyásoló konstrukciók, például. a páros kötőszók: *nem... hanem, akár... akár, vagy ... vagy, mind ... mind stb.*

Tesztelési eredmények. Az elkészült nyelvészeti alapú hangsúlyelemző tesztelését 580 mondaton végeztük el manuális módszerrel (Abari–Olaszy 2007). A mondatok összesen 6974 szót tartalmaztak, tehát ennyi hangsúlyjelzést ellenőriztünk. Már a fejlesztés során kiderült, hogy az ilyen elemzők működése erősen függ a szöveg felépítésétől, tartalmától, tehát az ilyen elemzőket hozzá kell igazítani a vizsgált szöveghez. Az elemző az esetek 7,6 százalékában adott hibás döntést. A részletezést a 10.4. táblázat mutatja. A négy hibatípus közül azokat számítjuk nagyobb hibának, amikor

10.4. táblázat. A nyelvészeti hangsúlyelemző hibás döntéseinek eloszlása a vizsgált 580 mondatban

	A hiba típusa	A hibák száma
1.	N kell F helyett	8
2.	N kell E helyett	64
3.	N kell W helyett	230
4.	W kell N helyett	228

az adott szóra F, E, W kerül annak ellenére, hogy a szó hangsúlytalan, tehát a táblázat szerinti 1, 2, 3 kategóriákat. Ezekből összesen 302 esetet találtunk (4,3%). Kevésbé zavaró hibának számítjuk a 4. kategóriát, mivel a hiányzó hangsúly kevésbé zavarja meg a hangzási képet, mint a feleslegesen hangsúlyozott szó. Az eredmények tehát azt mutatják, hogy a súlyosabb hibából több van, mint a kevésbé zavaróból. A vizsgálatokból világossá vált, hogy a nyelvészeti alapú mondatelemzés legjelentősebb

része a főnévi csoportok azonosítója. Egy NP elemzési hiba több hangsúlyhibához is vezethet, mivel a hangsúlystruktúra a mondatszerkezetre épül. Végeredményben azt mondhatjuk, hogy az elemzési eredmények javítására egyrészről a főnévi csoportok azonosítását kell pontosabbá tenni, amivel együtt jár a szövegek tartalmi szempontú vizsgálata is. Általános, nyelvi hangsúlyelemző készítése tehát egyelőre irreális célkitűzés.

10.3.1.2. Az alapfrekvencia változásának szabályalapú modellezése

A beszéd alapfrekvenciája a dallam és a hangsúlyozás létrehozásában vesz részt. A beszédkutatók már számos nyelvre készítettek alapfrekvencia-modelleket, Adriaens (1991) és Möbius (1995) németre, Pasdeloup (1992) franciára, Fujisaki (1992) japánra és más nyelvekre, Taylor (1998) angolra és más nyelvekre, Gronnum (1992) dánra.

A magyarra vonatkozó, szabályalapú alapfrekvencia-modellezés kidolgozásánál az alapkövetelmény az algoritmizálhatóság volt (Olaszy 2002). A modell lineáris alapfrekvencia-változásokkal közelíti a természetes alapfrekvencia-görbét. Percepciós tesztekkel kimutatták, hogy a természetes ejtés dallamformáinak közelítése ilyen egyszerűsített, vonalas dallamelemekkel megengedhető (Collier–Terken 1987), ugyanis hosszú mondatokban a hallgatók az így generált mondatdallamot a természetes beszédhez igen hasonlónak tartják. A modellünk felülről építkezik lefelé, a szöveg szintjétől a hang építőelemig. Az alapfrekvencia-változás végleges időfüggvényét szuperpozíciós elven alakítjuk ki a legmagasabb szintről kiindulva és lefelé haladva a hangig. A modellben alkalmazott lineáris építőkockák helyét (kezdet és végpont) a szöveg felszíni elemzéséből kinyert információk adják meg, azokat címkékkel jelöljük a szövegben. A szövegelemzést szabályok és listák felhasználásával hajtjuk végre (nem nyelvészeti mondatelemzéssel). Az F_0 változását %-ban, azaz relatív viszonyban adjuk meg, így bármely hangfekvéshez használható, csak meg kell adni a hangfekvésre jellemző alapfrekvenciaértékét mint referenciapontot, azaz a 100%-ot (férfi hangra például ez 120 Hz lehet). A fizikai F_0 érték ebből az adatból a mondat minden pontjára meghatározható.

A modell szintjei és tartalmuk.

1. Szövegszint – a mondatok dallammeneteinek kapcsolódási viszonyai.
2. Mondatszint – a mondat általános dallamformájához szükséges építőelemek.
3. Prozódiai egység(ek) – a mondaton belüli egységek definiálása és dallamformáik.
4. Szószint – a mondaton belüli, szószintű dallammenetek meghatározása.
5. Szótagszint – az alapfrekvencia-változások rendszere (hangsúlyok és más esetek (például eldöntendő kérdések).

6. Hangszint – a mikrointonációs változások meghatározása a magánhangzókban, valamint a zönge kváziperiodikusságának megvalósítása.

A modell bemenete ortografikus szöveg, kimenete egy címkékkel ellátott szöveg. A címkék jelzik, hogy hol és milyen alapfrekvencia-változás lesz a mondatban. A legmagasabb (1) és legalacsonyabb szintű mikrointonációs változásokat (6) nem jelöljük címkékkel, azokat szabályok alapján hozza létre az algoritmus. A modell címkekészlete kétszintű, egyrészről a hosszabb egységekre vonatkoztatott (dallam) jelzések rendszeréből áll, másrészről a szótagra ültetett F_0 változásokéból. A szövegszintű kapcsolatrendszer modellszintű leírása a 6.1.1 fejezetben található.

Mondat-, mondatrész- és szószintű dallamcímkék. A hosszabb beszédegységekre ereszkedő, szinttartó és emelkedő lineáris F_0 függvényeket határoztunk meg (10.5. táblázat) különböző frekvenciaátfogásokkal. Ezeket dallamsémának neveztük el. A dallamsémákhoz nincsenek időszakaszok hozzárendelve, így a tényleges, fizikai

10.5. táblázat. Dallamsémák a magyar beszéd dallammodelljéhez. A számadatok %-os változásokat fejeznek ki

	Forma/változat	1	2	3	4	5	6	7
1	Eső	100–85	90–80	90–70	85–75	80–70	70–65	100–70
2	Gyengén eső	100–95	95–90	90–85	85–80	80–75	75–70	–
3	Szinttartó	100–100	95–95	90–90	85–85	80–80	75–75	70–70
4	Gyengén emelkedő	95–100	90–95	85–90	80–85	–	–	–
5	Emelkedő	90–100	85–100	80–90	80–95	75–90	70–80	–
6	Erősen emelkedő	70–100	80–100	–	–	–	–	–

változást tekintve az időtől (a szó hossza, a mondatrész hossza stb.) függően végtelen sok formáció kialakítható, hiszen számtalan meredekség megadható. A dallamséma címkéje a 10.5. táblázat egyes függvényeire való hivatkozást jelenti. A címke az oszlopsor számadatát tartalmazza, jelölése: /fv , ahol f az első oszlop (forma), a v az első sor (változat) számadata. Például a */17 Asztalok.* bemeneti adatsorra a szintetizátor az egyszavas kijelentő mondat dallamát valósítja meg 100%-ról 70%-ra csökkenti az alapfrekvenciát. Ha 120 Hz-es referenciaponttal számolunk, akkor 120 Hz-ről indulva fokozatosan csökken 84 Hz-re. A dallamséma hatásköre addig tart, amíg új ilyen címke másként nem rendelkezik.

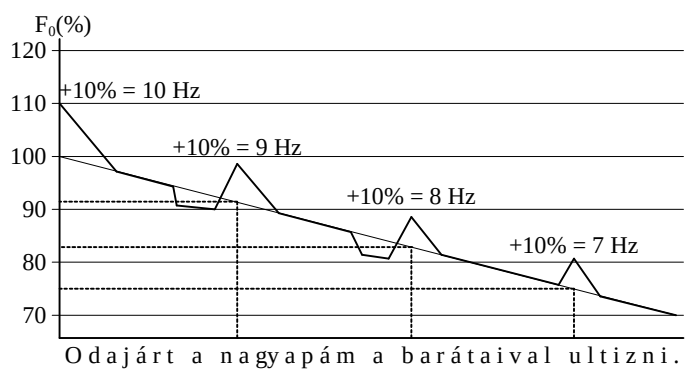
Prozódiai egységek. A modellben a mondaton belüli központozási jelek, illetve a szövegből származtatott határjelzések adják a prozódiai egységek határát. A dallam jellemzően az ilyen egységek határán változhat legmarkánsabban (új dallamrész indul, szünet van). Ennek jelzésére a // jelet használjuk. A // címke hatóköre a következő ilyen címkéig tart.

Szótagcímkék. A szótagszintű jelzéseket a hangsúlyelnevezéssel és különböző fokozatokkal használjuk. A szótagon belüli alapfrekvencia-változásokat egy emelkedő-csökkenő F_0 mozgással modellezzük. A modell számára meghatározott öt hangsúlycímke és a hozzájuk tartozó F_0 csúcsok értéke a következő. F = erős hangsúly (fó-

kusz), 140%; E = kiemelt hangsúly; 130% W = normál hangsúly, 120%; N = neutrális 100% (a szó nem hangsúlyos); mínusz jel = negatív hangsúly, 80% (a neutrálisra jellemző alapfrekvenciánál is mélyebb F_0 érték). Ezeket a címkéket a mondat modalitásától függetlenül, bárhol használhatjuk, ahol egy szótagszintű alapfrekvenciaváltozást jelezni kell. A címkék hatóköre arra a szótagra vonatkozik, ahová elhelyeztük (kivéve a negatív hangsúly címkéjét, ami az egész szóra vonatkozik). A csúcsot a szótagmagon belül változtathatjuk a hangsúlyozási megvalósulások formációi szerint (lásd a 6.2. fejezetben). A szótagcímkék fizikai megvalósítása úgy történik, hogy a szótagon belüli frekvenciamozgást a dallamséma aktuális pontjára vonatkozó F_0 értékből számolva a dallamsémára szuperponáljuk. A hangsúlyozott szó megtalálását a mondaton belül a korábbi fejezetben ismertetett modell segíti.

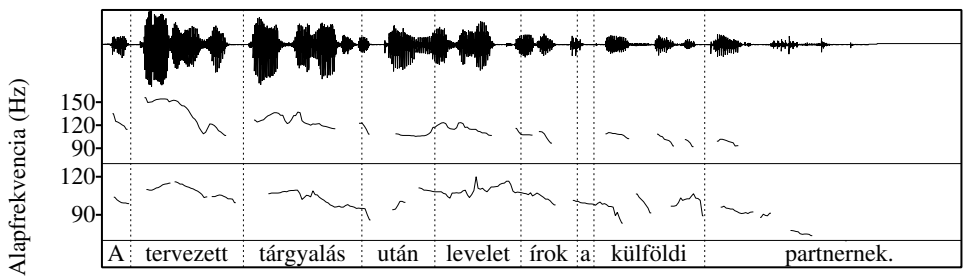
A szuperpozíció eljárása. Az első lépésben a mondatra meghatározott összes dallamséma kezdő- és végpontjait számítjuk ki, és ehhez meghatározzuk a két végpont közötti dallammenet frekvenciaértékeit például 5 ms-os osztással. A második lépésben a szótagszintű alapfrekvencia-változást szuperponáljuk a dallamsémára. A szuperpozíciót gördülő módszerrel, dinamikusan hajtjuk végre, ami annyit jelent, hogy a szótagra meghatározott F_0 mozgások %-értéke alapján kiszámoljuk a változást, mindig a dallamséma adott pontján lévő frekvenciaértékből mint kiindulási alapból, pontról pontra. Ebből az következik, hogy a csúcsra megkapott kiemelkedés a dallamséma különböző pontjain – hacsak az nem lebegő karakterű – más és más lesz. Amennyiben a dallamséma eső jellegű, akkor a csúcsokra számított értékek a dallamséma alapfrekvenciájának csökkenésével arányosan szintén csökkennek (mindig közelebb kerülnek a dallamséma frekvenciaértékéhez). Ez a fokozatos értékcsökkenés (Collier 1990) a beszéd jellegzetessége kijelentéseknél (10.15. ábra). A példamondat az alapfrekvencia jelzésekkel a következő:

//17 [:W] *Odajárt* [:-]*a* [:-W]*nagyapám* [-]*a* [:W]*barátaival* [:W]*ultizni.*



10.15. ábra. Példamondat az //17 jelű dallamsémára, valamint a szuperpozíció elvének alkalmazására és a fokozatos értékcsökkenés létrehozására 10%-os F_0 csúcsok alkalmazásával

Ugyanilyen szuperponálással valósítjuk meg a negatív hangsúllyal jelölt hangsor részek fizikai F_0 tartalmát. Ekkor a dallamséma által reprezentált F_0 érték alá fog csökkenni az alapfrekvencia a jelölt szóban. A fenti elvet alkalmazzuk minden szótagszintű, pozitív irányú alapfrekvenciaváltozás megvalósítására (lásd kérdések dallama). Összetett mondatok esetén a dallamsémák kombinálását alkalmazzuk. A modell működésének eredményét a 10.16 ábrán mutatjuk be.



10.16. ábra. Az alapfrekvencia-modell által generált F_0 menet (lent) összehasonlítva a természetes ejtésből regisztrálttal (középen)

A modell előnyei a következők: a teljes körű prozódiát modellezi; a modell belső szerkezete ismert, tehát új szabályokkal bővíthető, finomítható a működése; egyes moduljai ki-be kapcsolhatók a szövegszintű jelzőkarakterek beírásával-kitörlésével, ezzel támogatva a kísérleti munkát; hangfekvéstől független, a felépített dallamforma bármikor reprodukálható; személyfüggetlen. A modell hátránya, hogy csak magyarra vonatkozik, megvalósításához jelfeldolgozásra van szükség, ezzel romolhat a hangminőség, a személyes jellemzőket elmossa; sok élőmunkát, nagy szaktudást igényel a kialakítása és tesztelése.

A magyar fonológiai hanglejtésmodellt Varga (1994) dolgozta ki. Fonetikai megfeleltetési szándékkal Olaszy (2001a) összekapcsolta a fonológiai modellt a jelen megoldással és bebizonyította Varga rendszerének helyességét. Percepciós teszttel igazolta, hogy a fonológiai egységek alapján a modellel előállított hangzás kifejezi a fonológiai kategóriákat. Ezzel összekapcsolta az elméletet a gyakorlattal. A fenti modellt alkalmazta Koutny (2008) a magyarra kidolgozott, nyelvészeti alapú prozódiai rendszerének ellenőrző kísérleteiben. Ez a modell működik a ProfiVox magyar szövegfelolvasóban (Olaszy et al. 2000a), és alapul szolgált a későbbi, más technológiákkal megvalósított magyar nyelvű gépi szövegfelolvasók tervezéséhez is.

10.3.1.3. A beszéddallam változatosságának statisztikai modellezése

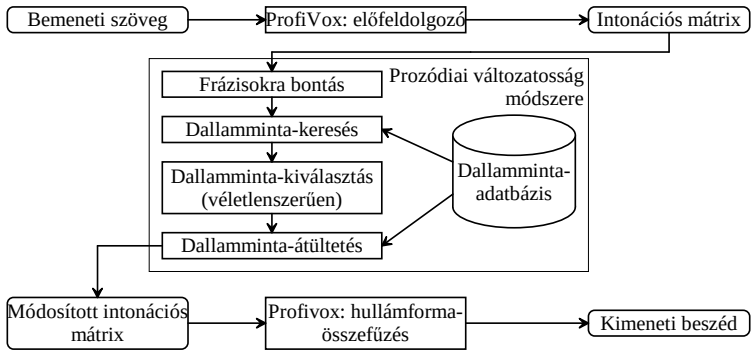
Csapó Tamás Gábor

Számos olyan módszer ismert a szakirodalomban, mely egy szintetizált mondat prozódiáját valamilyen természetes beszédkorpuszból kinyert adatok alapján hozza létre (Dong–Lua 2000, Raux–Black 2003, van Santen et al. 2005). Az ilyen megoldásokban emberihez hasonló dallammenet létrehozása azzal garantálható, hogy a szintetizálandó mondat alapfrekvencia-menetét az adatbázisból vett kisebb-nagyobb elemek (például szótag, szó, frázis) dallammenetéből jósolják statisztikai módszerekkel.

A magyar nyelvre készített statisztikai alapú alapfrekvencia-modell elsődleges célja a természetes beszéd dallamváltozatosságának modellezése volt (Németh et al. 2007c). Az emberi beszédben a prozódia rendkívül változékony paraméter. Egy-egy mondatot még akarattal sem tudunk többször ugyanúgy elmondani, a mindennapi beszédben pedig óriási különbségek tapasztalhatóak dallam, hangsúly és ritmus terén is. A legtöbb beszédszintetizátor-rendszer ezzel szemben determinisztikusan állítja elő a prozódiát, azaz egy-egy bemeneti szöveghez a beszédszintetizátor futása során mindig ugyanaz a dallam tartozik. Ez sokszor ismétlődő, monoton dallamminták túlzott előfordulásához vezet, ami zavaró lehet a hosszabb szövegek (hírek, időjárás-jelentés stb.) hallgatásánál. A prozódiaminták ismétlődése azért fordulhat elő a szövegfelolvasó rendszerekben, mert például egy elemkiválasztásos szintetizátor mindig a legjobb prozódiát próbálja egy-egy mondathoz rendelni. Így az emberi beszéd változatossága lecserélődik a legjobb, leggyakoribb mintára. Ez viszont az emberi fül számára, ami a változékonysághoz szokott, zavaró lehet. Éppen ezért a beszédszintetizátornak sem szükséges mindig a legjobb prozódiát megtalálnia, inkább egy elfogadható tartományt érdemes definiálni, amin belül megfelelőnek tartjuk a minőséget. Az emberi beszéd változatosságának elemzésével Chu et al. (2006), szövegfelolvasóban történő modellezésével Díaz–Banga (2006) foglalkozott korábban. A magyar nyelvhez illeszkedő statisztikai alapú modell kidolgozása során a ProfiVox elemösszefűzésen alapuló beszédszintetizátorból indultunk ki (10.3.6. fejezet). A kidolgozott modellel Németh et al. (2007c) egy kísérlet keretében bemutatta, hogy a változatos dallam megvalósítható szövegfelolvasókban, akár az ugyanolyan kategóriájú mondatokra is. Ezután a módszert automatizálták (Csapó et al. 2008), majd a ProfiVox rendszerbe is beépítették (Csapó 2009). Az alábbiakban röviden összefoglaljuk a módszer lényegi elemeit. A hipotézis az volt, hogy a gépi beszéd változatossága oly módon valósítható meg, hogy egy-egy bemeneti mondathoz a rendszer több különböző alapfrekvencia-menettel rendelkező változatot is előállít, amelyek közül szintéziskor egyet véletlenszerűen választ. Így megoldható az, hogy ugyanazon mondat máshogy szóljon többszöri szintetizálás során, azaz csökkenthető a monotonitás. A különböző dallamminták keresését a statisztikai, korpuszalapú prozódiai modellek működéséhez hasonlóan oldottuk meg. A kutatásban többek között a Fék et al. (2006) által bemutatott időjárás-előrejelzés témájú beszédkorpuszt használtuk fel,

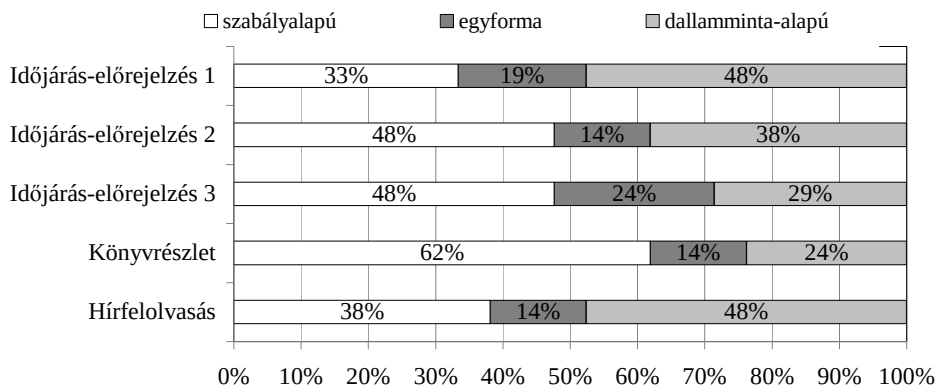
amely 5232 mondatot (mintegy 50 órányi hanganyag) tartalmaz egy professzionális bemondónő felolvasásában (10.3.7. fejezet). A korpuszhoz rendelkezésünkre állt a hullámformák mellett a mondatok szöveges és fonetikus átírása, a hang- és szóhatárok jelölése, valamint a pillanatnyi F_0-értékek. A beszédkorpuszt előzetesen feldolgoztuk, dallamminta-adatbázist hoztunk létre. Első lépésben automatikusan frázisokra bontottuk a korpuszbeli mondatokat, majd eltároltuk ezen kisebb egységek mondaton belüli pozícióját, szótagszerkezetét (az egyes szavak szótagszámai), illetve szótagonként egy-egy átlagos F_0-értéket. A módszer működését a 10.17. ábrán mutatjuk be. A folyamat azt mutatja, hogyan történik a kívánt dallam meghatározása a beszédkorpuszból és a szöveghez rendelése (adatszinten) a ProfiVox beszédszintetizátor segítségével. A bemeneti szöveg alapján a ProfiVox előfeldolgozó része szimbolikus információt (hangsúlyok és szünetek helye, hangok tervezett időtartama) hoz létre, melyet úgynevezett intonációs mátrixban tárol. Ezután a bemeneti mondatot frázisokra bontjuk a tervezett szünetek szerint szétválasztva, melyet a szövegben szereplő írásjelek alapján lehet megoldani. Az egyes frázisokhoz külön-külön dallammintákat keresünk az adatbázisból. A megfelelő minta kiválasztása többféle hasonlósági mérték szerint történhet. Mivel a magyar nyelvben a hangsúly a szavak első szótagján realizálódik, ezért kihasználtuk azt a feltételezést, mely szerint két frázis közötti egyező szótagszerkezet valószínűleg hasonló hangsúlyszerkezetet is jelent. A bemeneti szöveg frázisokra bontott részeihez ennek megfelelően meghatározzuk azok szótagszerkezetét, majd a szótagszerkezet alapján keresünk hasonló frázisokat a dallamminta-adatbázisból. A hasonlósági mértéket különböző kísérleteinkben variáltuk: egyes esetekben figyelembe vettük a frázisok mondaton belüli pozícióját, máskor a ProfiVox előfeldolgozó által meghatározott hangsúlycímkéket is felhasználtuk, illetve a pontosan egyező szótagszerkezet helyett enyhébb kritériumként hasonló szótagszerkezetet is megengedtünk a keresés során. A kiválasztott adatbázisbeli elem dallamát ezután szótagonként a bemeneti szöveghez rendeltük. Az ily módon módosított intonációs mátrix alapján a ProfiVox elemösszefűzés-alapú beszédszintetizátor létrehozza a gépi beszédet. Egy-egy frázishoz nagy valószínűséggel több szerkezetileg hasonló minta is előfordul az adatbázisban. Az ezek közötti választás módszerünkben véletlenszerűen történik. A véletlen választással biztosítható, hogy ha többször egymás után ugyanazt a mondatot vagy akár hasonló szerkezetűt szintetizálunk, ezek különböző dallamúak lesznek.

Így tehát hosszabb mesterségesen előállított beszédben is csökken a monotonitás. A módszer eredményességének ellenőrzésére szubjektív meghallgatásos tesztet állítottunk össze. A teszt célja az volt, hogy hosszabb mesterséges beszéden vizsgáljuk módszerünket. Három témából (időjárás-előrejelzés, hír- és könyvfelolvasás) gyűjtöttünk szövegeket, melyekből először a ProfiVox szabályalapú prozódiamodelljét alkalmazva készítettünk szintetizált beszédet. Ezután az új módszer segítségével is előállítottuk a mondatokat, mindegyiket többféle dallamváltozatban. Az így létrehozott gépibeszéd-részleteket párokba állítottuk, melyek 3-3 egymás után következő



10.17. ábra. A prozódiai változatosságot megvalósító módszer blokkdiagramja

mondatot tartalmaztak mindkét módszerrel előállítva. A fenti három témából összesen öt ilyen párt állítottunk össze. A kísérlet során az volt a tesztelők feladata, hogy az egyes párok közül eldöntsék, melyiknek változatosabb a dallama. A tesztelést a BME TMIT-en létrehozott webes tesztelőrendszerben végeztük. 2008-ban összesen 21 tesztelő végezte el a meghallgatást. A kísérlet eredményeiből az derült ki,



10.18. ábra. A prozódiai változatosságot megvalósító módszer meghallgatásos tesztjének eredményei, a tesztelők válaszainak aránya

hogy a legtöbb esetben a változatosabb dallam észrevehető volt a tesztelők számára is. A 10.18. ábra a tesztelők válaszainak arányát mutatja. Az öt tesztelt beszédrészletpár közül két esetben jelölték egyértelműen változatosabbnak a dallamminta-alapú változatot, a maradék háromban a szabályalapú változat ért el jobb eredményt. Utóbbi döntéseket valószínűleg az okozta, hogy a prozódiai változatosságért felelős modul bizonyos esetekben észrevehető hibákat hoz létre (például hirtelen ugrás a dallamban; vagy mondat végi alapfrekvencia-emelés, amely gondozott, illetve

felolvasásos természetes beszédben nem fordul elő). Ezek kiküszöbölése még fejlesztést igényel. A változatos dallam meghatározására kidolgozott módszer többféle beszédszintetizátor-technológiához is alkalmazható, segítségével természetesebbé tehető a szövegfelolvasók által létrehozott prozódia.

10.3.1.4. A beszéd időszerkezetének szabályalapú modellezése

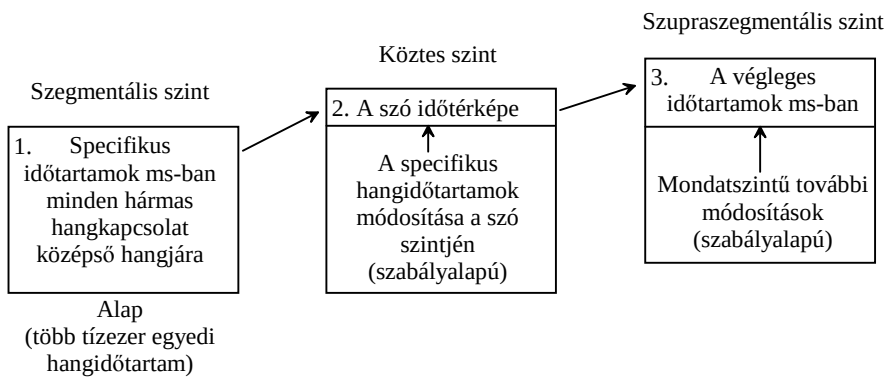
Olaszy Gábor

A beszéd időszerkezetét sok tényező határozza meg. Egyrészről eléggé szabályosnak tekinthető, ami a szegmentális szerkezetet illeti, hiszen itt minden beszédhangnak megvan a rá jellemző specifikus hangidőtartama. Másrészről az időszerkezet erősen függ a kimondandó gondolat (legyen például mondat) hosszától, tagoltságától, tartalmi összetevőitől. Harmadik komponensként jön be a képbe a beszélő egyéni beszédstílusa és a kifejezésmódja (akaratlagos, illetve beidegzett módosítások), valamint az érzelmi tényező (szomorúan mondja, vagy vidáman, örömmel telítve stb.). E három elemet csak elméletileg lehet szétválasztani, a végleges beszédproduktumban mindhárom egyidejűleg fejti ki a hatását, tehát a végeredmény mérhető a hullámformán. Több kérdést is fel lehet tenni a fentiekkel kapcsolatban. Hogyan lehet tehát modellezni a beszéd időszerkezetét? Miért van szükség a modellezésre? Hogyan mérhető a modellezés eredménye?

Lehetséges modellezési formák. Az időszerkezeti modellezés gondolata a beszéd gépi előállításával kapcsolatosan vetődött fel. Kiindulási alapot jelentettek a korábbi fonetikai mérések eredményei, amelyek a 20. század kezdetétől gyűltek (Gombocz 1909, Magdics 1965). A beszédtechnológia kiszolgálására azonban a fonetikai munkákban megadott átlagolt hangidőtartam-adatoknál részletesebb, finomabb jellemzésre volt szükség. Intenzív kutatás folyt az angol nyelvre (Umeda 1975, 1977, Klatt 1976) és más nyelvekre is (O'Shaughnessy 1981, Lehtonen 1970). Angolra készítették az első modelleket (Allen et al. 1987). A későbbi kísérletek arra irányultak, hogy minél mélyebben megtudják, hogy milyen tényezők befolyásolják a beszédben az egyes hangok pillanatnyi időtartamát (van Santen–Olive 1990, Riley 1992, van Santen 1992). A kutatási eredmények azt mutatták, hogy a hangidőtartamok függnek a hangkörnyezettől, a hangsúlyozástól, a hang helyétől a szóban, az adott szó hosszától, a szó helyétől a mondatban stb. Több paraméteres függvénnyel lehet tehát a hangidőtartamokat modellezni. Az utóbbi években előtérbe kerültek a statisztikai alapú időtartammodellek, amelyek nagy méretű, címkézett beszédadatbázisokból jósolják meg, hogy az adott hangkörnyezetben és helyi pozícióban milyen időtartam a legoptimálisabb a hangra. Ezt az tette lehetővé, hogy a beszédkutatásokhoz ma már elkészíthetők olyan beszédkorpuszok, amelyek korábban nem voltak megvalósíthatók. A jövő az ilyen megközelítéseké.

Modell a magyar beszéd hangidőtartamainak meghatározására. Az itt bemutatott modell az 1990-es évek végén készült és célkitűzése az volt, hogy jó közelítéssel meg tudjuk jósolni (adott artikulációs sebességhez) a beszédhangsor minden hangjának végleges, felszíni időtartamát a bemeneti szöveg paramétereiből, mint például a mondat típusa, hossza, összetettsége, a benne foglalt szavak egyedi hossza és hangszerkezete, a szavak hangsúlyozása. Ez volt az első időmodell a magyar beszédre. A modellel kapott hangidőtartamok általánosságban jellemzőek, nem tartalmaznak egyéni, beszélőre vonatkozó sajátosságokat.

A hipotézis az volt, hogy egy szabályalapú modellhez a komplex időszerkezetet három szintre célszerű bontani. Az első szint állandó alapot biztosít a számításokhoz, a másik kettőben vannak a módosító szabályok (10.19. ábra).



10.19. ábra. Háromszintű modell a magyar beszédhangok időtartamának meghatározására folyamatos beszédre a szövegből kiindulva

A modell szerkezete. Az első szint a szegmentális szintű alapidőtartamokat jelenti, amikor lényegében csak az artikuláció befolyásolja a hangidőtartamot (specifikus időtartam). Ezt tekintjük alapnak. A második szint a szóra vonatkozó időtartam-módosulásokat jelenti, a harmadik, felszíni szintet képviseli a szupraszegmentális szerkezet kialakítása, amikor az előbbi szintekre ráépülő összes további tényező, például a hangsúlyozás, az artikulációs sebesség és annak változásai, a mondat típusa és hossza alakítja ki a beszédhang végleges felszíni időtartamát. Az alapérték módosul(hat) a második szinten, az ott kapott érték pedig tovább módosul(hat) a harmadikon. A módosító tényezőket szorzófaktorok formájában valósítjuk meg. A módosított időtartamot úgy kapjuk meg, hogy a szorzófaktorral megszorozzuk a mindenkori aktuális szinthez tartozó időtartamot. A harmadik szinten elvégzett módosítás után kapjuk meg a végeredményt, a felszíni hangidőtartamot. Fontos érzékelnünk, hogy a hallgató ezt a végleges időtartamot hallja.

A specifikus időtartam. Az időmodell legalsó szintjét képezik a specifikus időtartamok (lásd 5.1.1.1. fejezet). A specifikus időtartamok értéke nem változik a modellben, ezek állandó számhalmazként jelentik az alapot. Erre építjük rá a modell további szintjeit. Ez a megkötés biztosítja, hogy ugyanarról az alapról indulva ugyanolyan jó eredménnyel eljuthatunk például az egyszavas kijelentő mondatra jellemző hangidőtartamokhoz, mint a többszörösen összetett kijelentő mondatéihoz. Ugyanezt az eredményt érhetjük el, ha nem kijelentésről, hanem különböző összetettségű kérdésről, felszólításról van szó, sőt, ha különböző beszédstílusokhoz kell időtartamokat meghatároznunk (reklám, hírolvasás, próza stb.). Lényegében tehát egy folyamatos szöveg minden hangjára, mondatról mondatra meghatározhatók a jellemző hangidőtartamok. Amennyiben ezekkel a hangidőtartamokkal számítógépes segítséggel beszédet állítunk elő, és ez a beszéd időszerkezeti szempontból magyar anyanyelvűek számára normatív hangélményt ad (érthető, és közel áll a természetes ejtés időszerkezetéhez), akkor azt mondhatjuk, hogy a modell legalsó szintje jól működik.

A szó időtérképe. A modell második szintjén a szószintű befolyásoló tényezőket határozzuk meg. A korábbi kutatásokban már tettek általános megállapításokat arra hogy mely tényezők határozzák meg egy szón belül a hangidőtartamok alakulását. A legfontosabbak ezek közül: a szó hossza, a szó helyzete a hangsorban és a hangsúly. A hangsúllyal a modellnek ezen a pontján nem foglalkozunk, mivel az a szupraszegmentális szinthez tartozik és hatását a modell következő szintjén vesszük számításba. Úgyszintén nem foglalkozunk a szó mondaton belüli helyzetével, ezt is a modell következő, harmadik szintjén fogjuk figyelembe venni. A szó hossza befolyásolja a hangok időtartamát, minél hosszabb a szó, annál rövidebbek benne a beszédhangok (Gombocz 1909, Tarnóczy 1974). Ezt a kiegyenlítődés törvényével magyarázták, amely szerint a produkció során az a törekvés, hogy a rövidebb és hosszabb hangsorokat nagyjából azonos idő alatt ejtsük ki. Fónagy (1959) kimutatta, hogy a Gombocz által megállapított időtartam-csökkenés versmondás esetén 6 hangnál hosszabb szavakban már nem folytatódik. Ezen általános megállapításokat beépítettük szabályainkba. Ezenfelül saját kísérleteink alapján úgy találtuk, hogy a szó hangszerkezete is lényeges befolyással van a hangidőtartamok alakulására (részletesen lásd Olaszy 2006b). Ez utóbbi hatásokat szabályrendszerbe foglaltuk. Az időmodell második szintjére kialakított fő szabályok a következők:

a) A szó szintjén egyedi, a szóra jellemző időtartam-szerkezeti képet lehet megadni a szó minden hangjára. Ezt a képet a szó hossza, a szóban lévő hangok és azok környezete határozza meg.

b) A szó hosszától függő hangrövidítéseket csak 6 szótagos szóig alkalmazzuk, tovább nem rövidítünk.

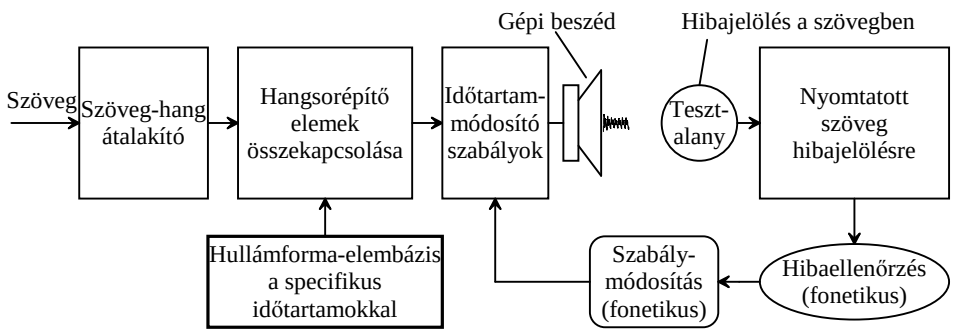
c) Külön kezeljük a szó első és utolsó szótagját, valamint a szó belsejében előfordulókat.

d) A rövid magánhangzók mindegyike külön csoportot képvisel a szabálymeghatározásban.

e) A hosszú magánhangzókat a hangidőtartam-szabályok szempontjából három csoportba soroltuk: az [aː], az [eː] és a rövid magánhangzók hosszú párjai, az [oː], [uː], [yː], [iː], [øː].

f) A hangkörnyezet szempontjából külön kezeltük a magánhangzót követő likvidák, valamint zöngés zárhangok hatását és külön a magánhangzót követő egyéb mássalhangzók hatását.

A szabályokhoz tartozó szorzófaktorok Olaszy (2006b) munkájában megtalálhatók. Az eredmények igazolták, hogy a második szint szabályai a szóban rejlő egyfajta dinamikus időképletet írják le, ami az adott szó kimondásakor jellemző. Ezt neveztük el a szó időtérképének (Olaszy–Abari 2005). Magyar szavak hangidőtartam-térképei tanulmányozhatók a `http://fonetika.nytud.hu/hitint` honlapon. A szószintű szabályok működésének ellenőrzésére az első szinten bevált eljárást alkalmaztuk (10.20. ábra)



10.20. ábra. A modell második szintjén előállított hangidőtartamok teszteléséhez kialakított mérési eljárás és elemei. A tesztelők mondatokat hallgattak meg

Az időmodell harmadik szintje. A beszédhangok a végleges időtartamaikat – a modell szerint – a szupraszegmentális szintű módosításokkal kapják meg. Ezen a szinten tovább változhatnak a hangidőtartamok, valamint szerepet kapnak a szünetek is, mint az időszerkezet fontos elemei. A szupraszegmentális szinten az volt a célkitűzésünk, hogy meghatározzuk a beszédben létrejövő – és általában a szöveg értelmezéséből adódó – kiejtési sebességváltozásokat, lassulásokat, gyorsulásokat, valamint a szünetek jellemző hosszát. A mérést a 10.20. ábra elrendezése szerint folytattuk annyi módosítással, hogy enyhén eső alapfrekvencia-görbére szuperponáltunk egy-egy F_0 csúcsot is (15%-os kiemelkedési csúcsponttal) a hangsúlyos magánhangzókra. Az in-

tenzitásértékeken a hangsúlyos szótagokban nem változtattunk. Mindezekkel némileg megközelítettük azt az állapotot, ami a természetes beszédben is történik a dallam és hangsúlyozás szintjén. Úgy találtuk, hogy a modellezés ezen a legfelsőbb szinten már sokkal bonyolultabb, mint a szó szintjén volt, hiszen itt már többek között a beszélő személy egyéni akarata is érvényesülhet, amely számos kiejtési formát valósíthat meg. A modellben ettől eltekintettünk, csak a legjellemzőbb prozódiai elemekre próbáltunk meg szabályokat megfogalmazni. A hangok időtartamát befolyásolhatja a hangsúly megléte, illetve hiánya, a szó helyzete a mondatban, valamint a hossza, továbbá a szünettartás sajátosságai. E négy elem figyelembevételével elégséges szabályt tudtunk megfogalmazni ahhoz, hogy a mindennapokban általánosan előforduló, felolvasásos beszéd kategóriájába sorolható beszédformák időszerkezeti kialakítására meg tudjuk adni a leglényegesebb időszerkezeti jellemzőket, és ezzel a modellt teljessé és használhatóvá tegyük.

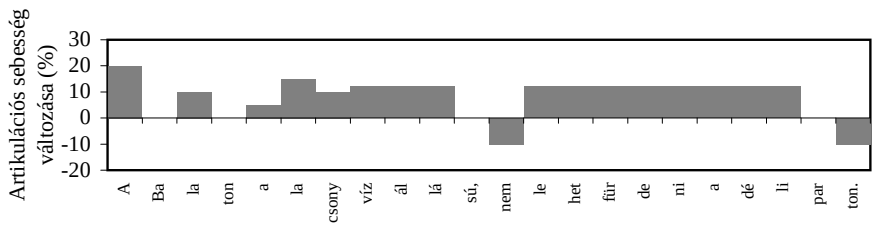
Az időmodell harmadik szintjére kialakított fő szabályok a következők:
a) A hangsúlyos szótagban a hangidőtartamokat nyújthatjuk (meghatározott feltételek teljesülése esetén), a hangsúlytalan részekben viszont rövidítünk (ez a szó belsejében is érvényes). Így tehát egy lassabb-gyorsabb-lassabb váltakozást hozunk létre a mondatban. E váltakozás egy-egy elemének a pillanatnyi hatóköre természetesen a mondat szövegétől erősen függ. Így a fizikailag megvalósított lassúbb szakasz viszonylag rövid, a gyorsabb szakasz hossza pedig attól függ, hogy a hangsúlyos elemek milyen sűrűn követik egymást a szövegben.
b) A legalább három szótagú szavakban a szó utolsó szótagjában fizikailag nem rövidítjük a hangokat, érzékeltetvén, hogy egyfajta relatív lassulás jellemzi a szó végét.
c) A mondatot kezdő tartalmas szót külön kezeljük.
d) A mondat utolsó szavában lassítjuk a tempót, különösen a szó végén.
e) Az egy szótagú szóban, mondat belsejében nem változtatunk az időtartamokon (ez relatív lassításnak felel meg).
f) Szünet után új időszerkezeti frázis indul.
g) A mondat hossza és a hangnyújtás fordított arányban áll. Egy szótagból álló mondatnál (*Ő., Én., Sok.*) erős nyújtást alkalmazunk. Több szótagúaknál fokozatosan csökkentjük az időtartamot a szótagszám növekedésével.
Általános megállapításunk, hogy a szupraszegmentális viselkedés függ a felolvasott szövegtől is. Más időszabályokat kell alkalmazni egy beszédszintetizátorban a regényrészlet felolvasásánál, mást egy számlaegyenleg közlésénél, és megint mást például a hírfelolvasásnál. Tehát ezen a ponton a modell szétágazhat különböző szabályhalmazokra, amelyek egy-egy stílust képviselnek (Olaszy 2005).

A hangsúllyal külön foglalkoztunk a modell harmadik szintjén. A magyarban fonológiai szempontból csak hangsúlyos és hangsúlytalan elemeket különböztetünk meg (Kálmán–Nádasdy 1994). Ezekhez társulhatnak mellékhangsúlyok is (Varga 2000). Fonetikai szempontból a hangsúlyosság tényének különböző fokozatok felelnek meg (gyenge, normál, erős, kiemelt), amelyeknek hatóköre általában a szó első

szótagjára vonatkozik (Elekfi–Wacha 2003). A nem hangsúlyozott szótag a hangsúlytalan kategóriába tartozik. Modellünkben a hangsúly megléte és a pillanatnyi ritmikai szerkezet szoros összefüggésben van. Három fokozatot valósítottunk meg a modellben: kiemelt és normál hangsúly, valamint a hangsúly hiánya (hangsúlytalan fokozat). E három hangsúlyfokozathoz a következő kiejtési sebességeket kapcsoltuk, igazodva Elekfi–Wacha (2003) elnevezéseihez:

közepes sebesség: nem változtatunk a modell 2. szintjén kapott időtartamokon,
gyorsabb: 0,8–0,9 szorzófaktorral rövidítjük a hangot,
lassabb: 1,1–1,3 szorzófaktorral nyújtjuk a hangot.

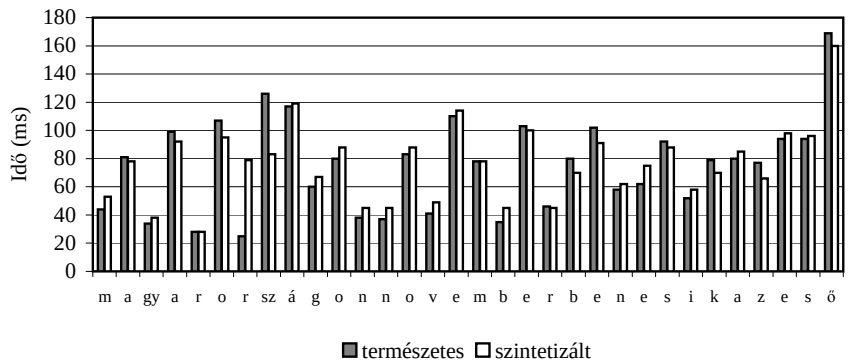
A hangnyújtás mértéke tehát viszonyításon alapszik. A közepes szinthez tartozó hangidőtartamot megnyújtottnak érezzük, ha a környezetében gyorsítottunk. Hogyan használjuk a fenti sebességfokozatokat a beszédritmus megvalósítására? A közepes fokozat a mondat első tartalmas és ugyanakkor kiemelten hangsúlyos szavára jellemző (például: ***Ne*** *menjetek át az úton!*), továbbá egyes szóvégi szótagokra, valamint a tagmondat előtti szó végére (amit a szövegben vessző jelölhet). A gyorsabb fokozat a hangsúlytalan elemeket jellemzi. A lassabb fokozattal a mondat belsejében lévő kiemelten hangsúlyos szótagokat illetjük (itt enyhén nyújtjuk a szótagot), továbbá a mondat végére kialakuló lassulást valósítjuk meg. A kiemelten hangsúlyos szóban a hangokat nem nyújtjuk, de nem is rövidítjük. Ez is relatív nyújtásnak tekinthető csak az egész szó szintjére vonatkozik. Ugyanennél a hangsúlynál, mondat belseji helyzetben viszont 1,1-es szorzóval nyújtjuk az első szótagot, a többi szó belsejit pedig 0,9-es értékkel rövidítjük. Itt tehát a nyújtás fokozottabb, hiszen nagyobb a különbség az első szótag időtartama és az utána következőkéi között. Vegyünk egy példamondatot: *A Balaton alacsony vízállású, nem lehet fürdeni a déli parton.* A mondatra elvégzett hangidőtartam-módosításokat a 10.21. ábra mutatja. Külön szabálycsopor-



10.21. ábra. Az artikulációs tempó változásának bemutatása a modell harmadik szintjén

tot készítettünk a rövid mondatokra. A kategóriák a következők: az egy szótagból álló (*Ő.*, *Lány?*) mondatban a magánhangzóra 1,4-es szorzófaktort alkalmazunk. A két szótagból álló (*Engem? Én is.*) mondatnál az 1,2-est. A folyamatos beszéd szupraszegmentális eleme a szünet is, melynek szerepe túl nyúlik a mondat szintjén. A modellben 3 szünetkategóriát használunk: mondat belsejében lévő szünet, mondatok közötti és a több mondatból álló tematikus egységeket egymástól elválasztó szünet.

A fenti három kategória szerint a modellben a szünetek hossza fokozatosan növekszik. A frázisegységeket 50–250 ms-os szünetek választhatják el egymástól, a mondatok között ennél lényegesen hosszabbakat kell alkalmazni. A mondatok közötti szünet lényeges eleme a folyamatos beszédnek, hiszen a hallgatónak ez a szünet biztosítja az időt arra, hogy feldolgozza, megértse az előző mondatot (Gósy 2000b). A bekezdés végén van a leghosszabb szünet. A szünetek hossza sok tényezőtől függ, a közlés témájától, a beszélő stílusától, valamint a mondat szerkezetétől, összetettségétől és hosszától (Olaszy 2006b). A modell működésének teszteléséből mutatunk be egy ábrát, amelyen összehasonlítjuk a természetes ejtés időtartamait a modellével (10.22. ábra). A szabályok alapján számított hangidőtartamoktól természete-



10.22. ábra. Az időmodell által számított hangidőtartamok összehasonlítása a természetes ejtésből mért adatokkal egy mintamondatban

sen nem várható el, hogy pontosan ugyanazokat az értékeket adják, mint amilyenek a természetes ejtésben szerepelnek, hiszen a modell beszélőfüggetlen. Az elvárásunk az, hogy a kiszámított hangidőtartamok tendenciájukban mutassanak hasonló képet, mint amilyen a természetes ejtésből származik. Ezt a teszteredmények igazolták, amit a mondat időtérképe is mutat. Az ábrán a legnagyobb eltérés az *ország* szó mássalhangzó-kapcsolatában van. Az arányok azonban nem sérültek lényegesen, hiszen ha összeadjuk a két mássalhangzó időtartamát, akkor azt kapjuk, hogy a szintetizált és a természetes változatban az összevont időtartam egyaránt 160 ms körüli érték. Ezt az időmodellt a ProfiVox magyar beszédszintetizátorba építették be, és 1999 óta számos gyakorlati alkalmazásban működik.

10.3.1.5. Komplex prozódiai modell

Az ezt megelőző fejezetekben láthattuk, hogy a beszéd tulajdonságait jól lehet jellemezni szabályalapú szintézistechnológiáknál. A statisztikai elvű gépi beszéd-

előállítóknál ezek a modellek nem használhatók, mivel nagy nehézségekbe ütközik a teljes körű, tanulásra kialakított, nagy méretű beszédadatbázisok előállítása. Jelenleg nincs olyan magyar beszédadatbázis, amelyik minden modalitásból elegendő mondatot tartalmaz és fel van címkézve a modalitás formáival, a mondat összetettségét leíró címkékkel, a dallammenet és hangsúlyozás minden jellemző tulajdonságával. A modellre viszont szükség van, hogy a prozódia legfontosabb komponensei a helyükön legyenek a szintetizált mondatban. Az elemkiválasztás-alapú beszédszintézistechnológiához alakítottak ki a BME TMIT beszédkutató laboratóriumában olyan komplex prozódiai modellt, amely alkalmazható mind a tanulási, mind a szintézisfázisban, és nem igényli a fenti részletezett prozódiai címkézést a tanulási beszédkorpuszban. A modell csak kijelentő mondatokra képes a prozódiai szerkezet mindhárom elemét (alapfrekvencia, intenzitás, időszerkezet) felmérni, osztályozni, majd a szintézisben használni (Németh et al. 2009).

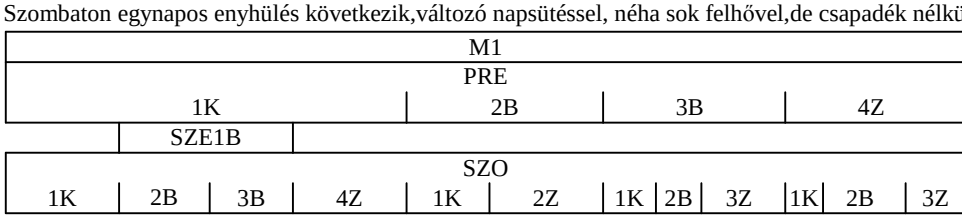
Ebben a modellben a mondatot tekintjük alapegységnek (annak a szöveges formáját). A modell arra alapul, hogy a mondat belsejében létrejövő prozódiai tartalmat visszavezetjük a mondat időskálájára, azon belül pedig a nyelvi szerkezeteire. Ez azt jelenti, hogy hely szerinti pozicionálásból következtetünk a prozódiai tartalomra. A megoldás magában hordozza azt a tényt, hogy az adott ponton lévő beszédrészlet, mint komplex jel, mindhárom paramétert tartalmazza (adott artikulációs sebességet, adott hangidőtartamokat, adott intenzitásszerkezetet – beleértve a hangsúlyt is, ha van –, adott alapfrekvencia-változást, amely magában foglalja a beszéd pillanatnyi dallamát is és a hangsúlyozás hatását is). A modell nem foglalkozik a prozódia egyes elemeinek részletes tartalmával, csak globális fogalmakkal (például: hangsúly van/nincs). Ezt azért tehetjük meg, mert ember olvassa fel a tanító-beszédkorpuszt, és az ember automatikusan alkalmazza a beszéd összes prozódiai elemének megvalósítását. A modellnek csak az a dolga, hogy ezeket az időtengelyen való jellemzéssel jól definiálja, majd használja. A modell nem alkalmaz semmiféle nyelvészeti mondatelemzést. Ha elég nagy beszédadatbázisunk van, és a mondatok belsejében pontosan tudunk helypozíciókat meghatározni (hanghatár, szóhatár, központozás helye, jellemző beszédrészlet), akkor ezekből a pozíciós információkból a modell fel tudja építeni a jellemző prozódiai szerkezetet, majd a szintéziskor ez alapján tud keresni a beszédadatbázisban. Azt is mondhatjuk, hogy a modell valamilyen szinten függetleníti magát a nyelvi tartalomtól. Hasonló felépítésű mondatok prozódiai szerkezete is hasonló (Németh et al. 2007c). Ugyanazt a modellt alkalmazzuk a korpuszban tárolt mondatokban és a szintetizálandó mondatra is. Így a szintetizálandó mondatból adódó prozódiai struktúra – a szegmentális szerkezettel együtt (maga a hangsor) – lépésről lépésre kereshető lesz a korpuszban.

A modell prozódiai elemei és azok hierarchikus szintjei

1. Alapegység – mondat (M). Információk: a mondat belső felépítése és hossza az alábbiak szerint.
2. Prozódiai egység (PRE) – a mondaton belüli legnagyobb egybefüggő jelfolyamat. Egy mondaton belül lehet egyetlen, illetve több ilyen egység. Információk: előtte szünet is lehet, de nem kötelező; innen új prozódiai szerkezet indul. Általában központozási jel jelzi a kezdetét, de meghatározhatók más jelzők is, amelyekhez hozzá köthető a határa. Megállapítási hatékonysága mintegy 80%-os.
3. Szövegrész (SZE) – a prozódiai egységen belüli, prozódiailag egybetartozó jelrészlet (általában szókapcsolatok: *délutáni hőmérséklet; éjszakai órákban, távközlési szolgáltatás, kereskedelmi egység* stb.). Információk: a szövegrésznek jellemző prozódiai szerkezete van, amely a teljes szókapcsolatra vonatkozik (mint ha egyetlen szó lenne a szókapcsolat), tehát egyben kezelendő. Megállapítási hatékonysága 70%-os.
4. Szó (SZO) – a prozódiai egységen belüli legkisebb elem. Információk: hány szótagból áll, hol helyezkedik el az időtengelyen a mondat egészére vonatkoztatva a 2. és 3. pont függvényében.

A modell a fenti kategóriákra bontja a szöveget és minden kategóriának külön címkét ad.

A modell címkerendszere a következő: Mondat (M), prozódiai egység (PRE), szövegegység a prozódiai egységen belül (SZE), szó (SZO). Mindegyiket a saját jele után tett számmal sorszámozzuk annak megfelelően, hogy hányadik az adott egységen belül. A sorszámon kívül minden elemre három pozíció vonatkozhat az adott egységen belül: kezdő elem (K), belső elem (B), záró elem (Z). Ezek a jelek a sorszám után következnek. Egy címkesorozat tehát pontosan megmondja, hogy az adott elem (M, PRE, SZE, SZO) a mondaton belül hol helyezkedik el. A 10.23. ábrán láthatunk egy ilyen címkézést egy példamondatra. A modell működésében a hierar-



10.23. ábra. Komplex prozódiai modell által előállított címkézés

chia fontos szerepet kap. Amennyiben az adott mondat megtalálható, akkor a kereső azt változtatás nélkül veszi ki a mondattárból. Ha az adott címkéjű és tartalmú

PRE megtalálható, akkor azt veszi ki a válogató algoritmus. Ugyanez vonatkozik a szövegegység-, illetve a szószintre. Ezzel a keresési formával nagy valószínűséggel biztosíthatjuk a helyes dallamot, hangsúlyozást, tempót, ritmust.

A modell előnye, hogy egyszerű, nem igényel bonyolult nyelvészeti elemzést, nem alkalmaz jelfeldolgozást, ezzel a teljes hangszínezet megmarad (tisztán felismerhető a beszélő hangja). A modell hátránya, hogy a tanító beszédkorpuszt ezen elvek alapján kell elkészíteni. További hátrány, hogy csak előre meghatározott, szűk témakörre lehet elvárni az optimális működést (például időjárás-jelentés, vasútállomási tájékoztató stb.). A hangzási hibák javítására nincs mód, kivéve, ha új hangfelvételeket csinálunk a kérdéses mondatok felolvastatásával. Ilyenkor ügyelni kell, hogy a bemondó ugyanazon hangfekvésben és hanghordozásban beszéljen, mint a korábbi felvételekkor. Ezt a modellt a 10.3.7. fejezetben ismertetett gépi beszéd-előállító eljárás alkalmazza.

10.3.1.6. Beszélő fej modellezése

Czap László

A beszéd multimodális jelenség. Ha a hangot a szájmozgással és a testbeszéddel együttesen érzékeljük, akkor hatásosabb a beszéd megértése, mint egyébként (Sumby–Pollack 1954). A gépi beszéd természetes kiegészítője lehet a beszédet utánzó fejmodell. A beszéd animációs megvalósítása és az artikuláció modellezése az 1990-es években kezdődött. Az első próbálkozások kétdimenziós megoldások voltak (Cosatto–Grafat 1998). A korszerűbb megoldás a háromdimenziós (3D) modellezés, amit elősegített a számítástechnikai eszközök kapacitásának robbanásszerű növekedése és a természetes artikuláció analízise. A 3D rendszerek életszerű, fotorealisztikus finomságú modellek kidolgozását teszik lehetővé (Cohen–Massaro 1993). A dialógus- és oktatórendszerekben a gépi beszéd érthetőségét és az attraktivitást nagyban javítja a hangzó beszéddel párhuzamos animáció. Multimédiás alkalmazásokban a virtuális bemondó vagy szereplő tágítja a művészi szabadság határait. Hallássérültek beszélni tanítását segítheti a helyesen artikuláló virtuális bemondó, amely esetlegesen átlátszóvá tett arcával a természetes beszélőnél jobban megmutathatja a hangképzés nem látható részleteit is (10.24. ábra).

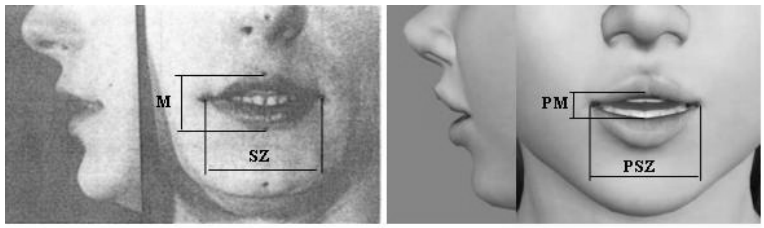
Alapok. A 3D modellek egyik típusa az arcizmok megfeszítésének vezérlésével szimulálja az arckifejezéseket. Az ilyen modellek valósághű eredményt nyújtanak, de a kívánt arckifejezés előállítása rendkívül számításigényes és az analízis során (betanítás) a valóságos izomtónusok nehezen mérhetők. A másik típus a pusztán felületi mozgásokat modellező animáció. Ez utóbbi vezérlési paramétereit megfigyelésből, vagy képfeldolgozási módszerekkel természetes beszélők képeiről gyűjtik össze

10.24. ábra. Fotorealisztikus és transzparens megjelenítés

(Massaro 1998b). Minden modellben a mozgatásnál külön figyelmet kell fordítani a paraméterek összehangolt változtatására, mert könnyen természetellenes hatás alakulhat ki. Az élethű beszédmodellezéshez a következő vizuális elemek megvalósítására van szükség:

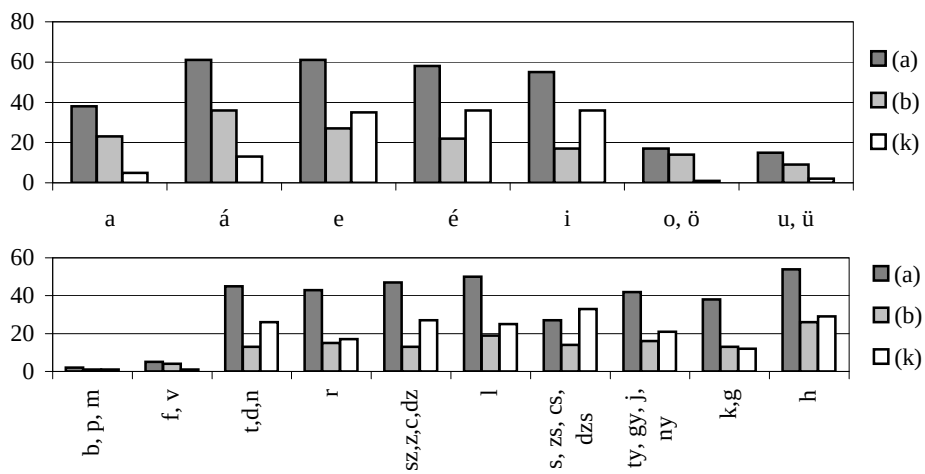
- száj- és nyelvmozgás a beszéd közben
- fejmozgás
- szem- és szemöldökmozgás
- érzelmek kifejezése bizonyos gesztikulálással

A vizuális modellezés alapegysége. A beszédfejmodellek vezérléséhez használt alapegység a *vizéma,* amely az artikulációt reprezentáló statikus képi elem. A vizémákat használhatjuk egyrészt a gépi beszéd-előállítás hangtestének vizualizálására (ezzel az elhangzott beszéd jobb megértését támogathatjuk), másrészt pedig a gépi beszédfelismerési algoritmusok döntéseinek pontosítására azzal, hogy a beszélő személy arcáról gyűjtött pillanatnyi információt összevetjük hangsorozatokhoz tartozó vizémasémákkal (lásd a a 9.12. fejezetet). A vizémák készlete szűkebb, mint a fonémáké. Nem látható például a zöngésség-zöngétlenség, de a képzés helyében megegyező, időtartamban vagy intenzitásban eltérő hangok is azonos artikulációs mozgásokkal jelennek meg az arcon. A hangképző szervek leíró jellegű statikus jellemző helyzetei magyar beszédhangokra alapvető fonetikai munkákban (Bolla 1980, Molnár 1986) megtalálhatók adott hanghelyzetekre (nem folyamatos beszédre). A 10.25. ábrán példát mutatunk be arra, hogy mennyire hasonló egy fényképen látható (Bolla 1995) és egy 3D-s beszélő fejen beállított, ugyanazon [e] hangra jellemző artikuláció (Mátyás 2003). A magyar beszédhangok statikus vizémakészlete a hangalbumokban megadott mintaszavak artikulációs jellemzőiből (Bolla 1980, Molnár 1986) került rögzítésre. A fonetikai leírásoknak elsősorban a képzés helyére vonatkozó adatai se-



10.25. ábra. Egy beszélő személy ajakartikulációja (bal oldalon) és a 3D fejmodell (jobb oldalon) által készített kép

gítik a csoportosítást. A 9.12. fejezetben ismertetett vizuális lényegkiemelési eredmények konkrét adatokkal szolgálnak. Az audiovizuális beszédfelismerést szolgáló adatbázisból kiolvashatók az egyes fonémák kiejtésekor mért ajakszélességek és az ajaknyitás mértéke. A vizémákhoz tartozó szájnyílás belső ajakméreteit és intenzitási tényezőit ábrázolja a 10.26. ábra (Czap 2004). A diagramokon látható harmadik paraméter (*k*) a szájüreg átlagos világosságát mutatja, ami összefüggésben áll a fogak és a nyelv láthatóságával. Maximális a világosság, ha a fogak láthatók, közepes, ha a nyelv elöl van, minimális, ha a nyelv hátul helyezkedik el, ekkor a szájüreg sötét. A magyar nyelvre 15 vizéma elégséges az artikuláció látható elemeinek leírására.



10.26. ábra. A magyar vizémák ajakszélessége (a), ajaknyílása (b) és a szájnyílás átlagos világossága (k). Az a és b méretek pixelben, a k világosság a fekete (0) és fehér (255) világosságértékek közötti aránnyal van megadva

Ezek a vizémák köthetők a beszédhangok artikulációjához. Az egyes vizémákhoz tartozó fonémákat a 10.6. táblázat mutatja.

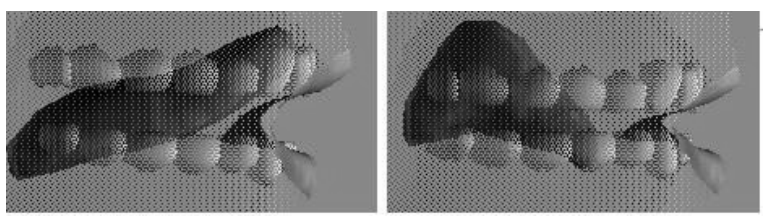
Néhány megjegyzés a vizémák osztályozásához:

10.6. táblázat. A magyar nyelv statikus vizémái sorszámmal ellátva és a hozzájuk tartozó beszédhangok (az írásjelükkel jelölve)

Vizéma sorszáma	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	15
a hang, amihez tartozik	á	a	o	u	i	é	e	b	d	g	gy	h	v	z	zs	r
			ö	ü				p	t	k	ty		f	sz	s	
								m	n		j			dz	dzs	
											ny			c	cs	

- a csoportosítás elsősorban ajakforma alapján történt, a nem látható nyelvállás eltérő lehet (például: o-ö, u-ü)
- a nem jelzett hosszú magánhangzók a rövid párjuknál szűkebb szájnyílással vannak jelen
- az artikuláció előállításához egy vizémaosztályon belül a szinkronizálási szabályok eltérhetnek, például [p, b, m] (lásd később)

A vizémákat számos jellemző írja le. Az artikulációs adatokat a fejmodell vezérlő paramétereivé kell konvertálni. A fejmodellhez tartozó legfontosabb jellemzők az ajkak (modellünkben 8. paraméter) és a nyelv működtetéséhez (7. paraméter) tartoznak. Az alapvető ajakjellemzők a nyitás (tág-szűk), és a szélesség (széles-keskeny). A száj szélességét és nyitottságát befolyásolja, hogy ajakkerekítéssel vagy ajakréssel ejtjük a hangot. A labiális, illetve illabiális jelleg a szájsarkak lekerekítési módját is meghatározza. Ajakkerekítéses hangoknál a száj sarka kerek, ajakréses hangoknál hegyes (a szilvamag alakjához hasonló). Az ajkak nyitása szoros összefüggésben van az állkapocs mozgásával (nyitott-zárt). Az állkapocs helyzete a nyitás mellett a fogak láthatóságával is kapcsolatban áll. A nyelvállást (10.27. ábra) a nyelv függőleges helyzete (fent-lent), vízszintes mozgása (elöl-hátul), hajlítása (domború-homorú) és a nyelvhegy formája (vékony-vastag) határozza meg. A statikus jellemzők



10.27. ábra. Jellemző nyelvállások: baloldalon az [n]-re, jobbra a [g k] hangokra

alapján beállíthatók a beszédhangok állandósult szakaszára jellemző artikulációs paraméterek.

A dinamikus működés diádmodellje A statikus képek között eltelt időt ki kell tölteni köztes adatokkal, vagyis a koartikulációs mozgásokat kell modellezni a kívánt mértékben. Ehhez a beszédszintézisnél megismert diádalapú hangkapcsolódások módszerét vesszük alapul. A vizémák a hangok középső (kvázistacionárius) álla-

potát ábrázolják, ezek közötti interpolációval a hangátmenetek képdiádjai belső állapotainak ábrázolása előállítható. Az eredmény azonban csak néhány képdiádnál elfogadható, általában a valóságosnál intenzívebb mozgás áll elő (túlartikulál a modell). Például a *terhel* és a *korhol* [rh] átmenetének ajakformája eltérő, azt a beszédben a szomszédos magánhangzók határozzák meg. A képdiádos elv ezt nem tudja modellezni.

A dinamikus működés dominanciamodellje. A képdiádos modell hiányosságait próbálja kiküszöbölni a dominanciamodell. Az artikulációs jellemzők koartikulációs alkalmazkodási készségét leíró dominanciamodell engedi, hogy egyes jellemzők a közeli vizémák értékeihez simuljanak. Vannak paraméterek, amelyek a környezettől többé-kevésbé függetlenül felveszik jellegzetes értékeiket, mások a környezetükhöz igazodnak. A vizémák minden egyes jellemzője (ajak- és nyelvállások leírói) osztályokba sorolandó dominanciajellegük alapján. A besorolás alapja az lehet, hogy különböző hangkörnyezetben mennyire változnak az adott hang jellemzői.

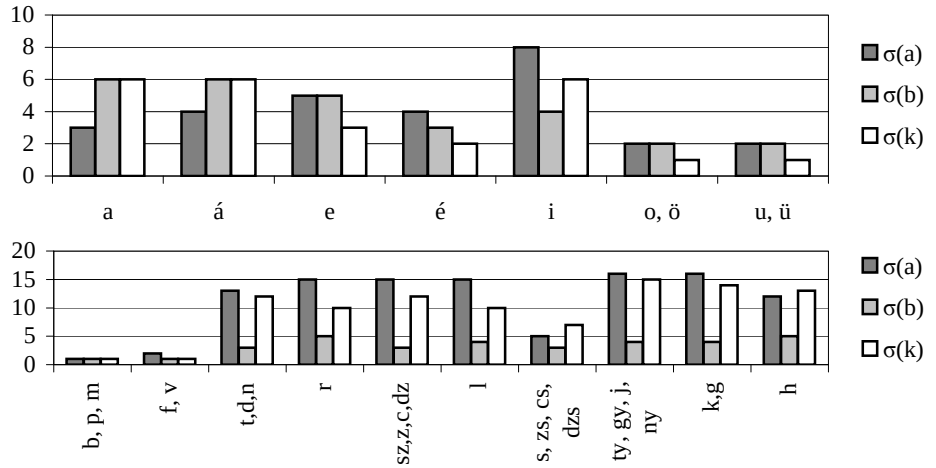
Magyar fejlesztés. A folyamatos magyar beszéd dinamikus artikulációs jellemzőinek átfogó leírása még várat magára. Az analízis során a hangalbumokban található pillanatképekből elkészültek a magyar vizémák, de azok csak a mintaszavak hangkörnyezetére vonatkoztathatók. A munka tapasztalatai alapján készült el egy olyan képi adatbázis, amelyben a beszélő hangján túl az arcát is rögzítették (9.12. fejezet). Az adatbázisból mért jellemzők szórása ad támpontot a dominancia mértékéről (10.28. ábra). Ha a vizéma adatai különböző hangkörnyezetben eltérő értékeket mutatnak, (nagy a szórás), az adott jellemző nem domináns. Ha a szomszédságtól függetlenül hasonló az értéke (kicsi a szórás), az adott jellemző domináns. Ebből származnak az ajkak nyitásának és szélességének, illetve a nyelv és a fogak láthatóságának időbeli változására vonatkozó adatok. A nyelvmozgásra vonatkozó besorolás leginkább önmegfigyelés útján alakult ki. A vizuális lényegkiemelés adatainak szórása alapján a vizémák leírói három kategóriába sorolhatók:

- *domináns* – alig enged koartikulációs hatásoknak
- *rugalmas* – a környezete befolyásolja az adott jellemzőt
- *határozatlan* – a környezete alakítja ki az adott jellemzőt

Példaként a 10.7. táblázat mutatja a vizémák ajakformára, a 10.8. táblázat a nyelv vízszintes helyzetére vonatkozó csoportosítását (betűkarakterekkel). A dominancia beállításai a paraméterek interpolációját határozzák meg. A további módosítások

10.7. táblázat. Dominanciajellemzők az ajakformára nézve

Domináns	magánhangzók, s, zs, cs, dzs
Határozatlan	k, g, r, h
Vegyes	p, b, m, l, j, n, ny, f, v, sz, z, c, dz,, d, t, ty, gy (ajaknyílás domináns, szélesség határozatlan)



10.28. ábra. A vizémák 10.26. ábrán megadott jellemzőinek szórása (ajakszélesség (a), ajaknyílás (b) és a szájnyílás átlagos világossága (k). Az (a) és (b) méretek pixelben, a (k) világosság a fekete (0) és fehér (255) világosságértékek közötti aránnyal van megadva

10.8. táblázat. Dominanciajellemzők a nyelv vízszintes helyzetére nézve

Domináns	t, d, n, r, l, ty, gy, j, ny, s, zs, cs, dzs, sz, z, c, dz
Rugalmas	magánhangzók
Határozatlan	p, b, m, f, v, k, g, h

lehetnek például hosszú magánhangzóknál állandósult szakasz beiktatása. Ezek finomítják az artikulációt. A hangzó beszédhez történő szinkronizálás az akusztikus jel időadatai alapján végezhető el. A fonéma közepéhez rendelt vizémaparaméterek közötti lineáris interpoláció csak ritkán ad kielégítő eredményt. Például az [m] hangnál elejétől végéig össze kell zárni az ajkakat, lineáris interpolációval csak a hang közepén érnének össze, az [m] szinkronizálásához két tartópontra van szükség. A bilabiális zárhangoknál [p, b] a zárfelpattanás előtti néma fázis vagy fojtott zönge is zárt ajkakkal valósul meg. A zárfelpattanáskor az ajkak vagy a nyelv hirtelen mozgásának pontosan egybe kell esni a zörej megszólalásával. A vizémákat tehát több ponton szinkronizálni kell a hozzájuk tartozó beszédhangok jól meghatározott jellegzetes pillanataihoz (az [m]-et két ponton, a [p, b]-t három ponton).

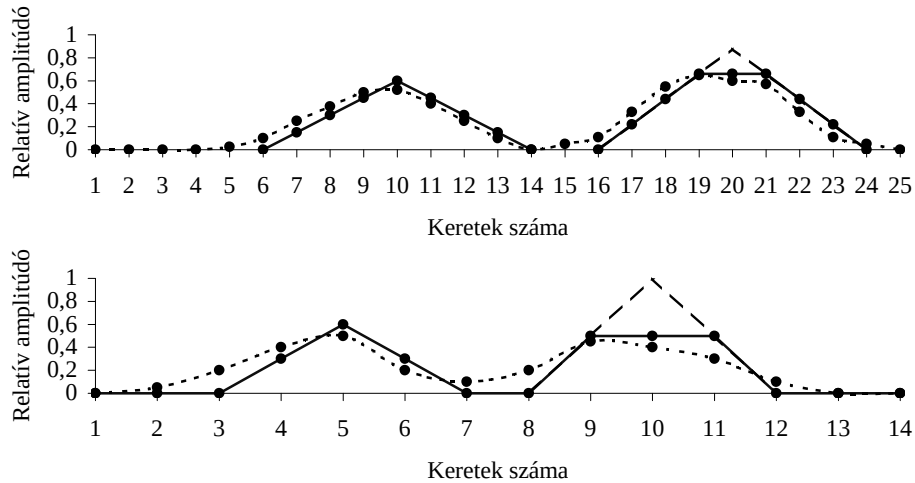
Az előartikuláció modellezése. A kimondás megkezdését mintegy 300 ms időtartamú csend előzi meg. Ez alatt az idő alatt a levegővételt az ajkak megnyitása imitálja. Ezután az ajkak alaphelyzetéből elkezdődik az első domináns vizéma kialakítása. Ezzel a kiegészítéssel – amit előartikulációnak nevezhetünk – már az első hang megszólalása előtt kialakul az ajakforma, hasonlóan a természetes kiejtéshez.

A beszédsebesség modellezése. A természetes vagy szintetizált beszédhez történő szinkronizálás folyamán különböző sebességű beszéddel szembesülhetünk. Lassú beszédnél a vizémák jellemzői megközelítik névleges értéküket, gyors beszédnél az artikuláció elnagyoltabb. A rugalmas csoportba sorolt jellemzőkre is igaz, hogy gyors beszédnél a lekerekítés erősebb. A rugalmas jellemzők kialakítására alkalmas a mediánszűrés: A szűrésben résztvevő mintákat nagyság szerint sorba rendezve a középső lesz a szűrt érték. A szűrés három mintára történik. Egy jellemző időfüggvénye három lépésben alakul ki:

- A domináns és rugalmas vizémák kitüntetett pontjai között – a határozatlanok kihagyásával – lineáris interpoláció.
- A rugalmas vizémák környezetében mediánszűrés. Ez kevesebb minta – gyors beszéd – esetén nagyobb csúcslevágást okoz.
- Az így kapott értékeken még egy simítást végzünk, amely az aktuális, a két megelőző és a követő mintákat érinti. A szűrt érték a négy minta súlyozott összege. A súlyozás állandó, nem függ a beszéd sebességétől. A simító szűrés egyrészt finomítja a mozgást, másrészt gyors beszédnél jobban lekerekíti a csúcsokat. A szintetizált beszéd analízise alapján a szűrés hatása előre erősebb (két keret), mint hátra (egy keret).

A 10.29. ábrán gyors és lassú beszédnél követhetjük a mediánszűrés és a simítás hatását a nyelv vízszintes helyzetére. A példában a lassú beszéd kétszer annyi keretből áll, mint a gyors kimondás. Az ábrán jól követhető a gyors beszédnél érvényesülő lekerekítés, a mediánszűrés és a simítás hatása egyaránt. A vízszintes tengelyen a képkeret sorszáma, a függőleges tengelyen a kitérés mértéke (1 maximum, 0 alaphelyzet) látható.

A modell ellenőrzése. A modellben a beszélő természetes fejmozgását vezérlő adatok televíziós híradókból felvett hírolvasó bemondók felvételeinek mérési eredményeiből származnak. Ennek nyomán álvéletlen mozgások, például visszafogott bólogatás, a fej enyhe oldalra billentése és az átlag körül szóródó pislogási periódus természetesebbé teszi a modell mozgását. A prozódia tükröződése a fejmozgásban, illetve az arcmimikában nehezen algoritmizálható. Így például a mondathangsúly kifejezése nehézségekbe ütközik. Az alapfrekvencia változása azonban felhasználható a szemöldök mozgatásának vezérlésére. Kijelentő mondatnál a szemöldök is enyhén ereszkedő, a mondathangsúlynál emelhető a szemöldök. A gépies, periodikusan ismétlődő mozgások elkerülésére a szabályalapú tendenciákat a valószínűségi alapon megvalósított mozgás teszi változatossá. Az akusztikus jel energiájával arányos a fej lefelé mozgásának, a szemöldök emelkedésének és a pislogásnak a valószínűsége. A szemmozgás a fejmozgás korrigálását szolgálja, hogy a tekintet mindig a nézőre szegeződjön. Ettől eltérő szemmozgatás – például a papírra pillantás imitálása – kézi beavatkozást igényel. Dialógusrendszerekben



10.29. ábra. Generált példa a domináns (1. csúcs) és rugalmas (2. csúcs) jellemző szűrésének és a lassú (fent), illetve gyors (lent) beszéd simításának bemutatására. A lineáris interpoláció eredménye (pontok), a mediánszűrés (síma vonal) és simítás után (szaggatott)

a szerepváltást segíthetik a gesztusok, az értő figyelést a szemöldök emelésével jelezhetjük, bólogatással is visszaigazolhatjuk figyelmes hallgatásunkat. Ezek a műveletek manuálisan állíthatók be a teljes mozgássor kialakítása után a jellemzők módosítására szolgáló grafikus szerkesztővel.

Érzelmek kifejezése. A beszéd multimodális jellegéhez hozzátartoznak a gesztusok is. A testbeszéddel árnyaljuk mondandónkat, megerősítjük, vagy éppen cáfoljuk verbális üzenetünket. Arcanimációs rendszerekben az arckifejezések érzelmi töltését programozni lehet. Általában az Ekman–Friesen (1978) által meghatározott hét alapérzelem jelenti a választékot, skálázható, és keverhető formában: semleges, haragos, ellenszenves, szorongó, boldog, szomorú, meglepett. Erre láthatunk példát a 10.30. ábrán. Az érzelmek kifejezése vezérlő kódokkal skálázható, és vegyes érzelmek kialakítására is lehetőség nyílik, kellő körültekintés mellett, vigyázva a groteszk hatás elkerülésére. Például 40% meglepetés és 20% öröm kellemes meglepetést szimulál. Jelenleg a szövegfelolvasásra is alkalmas rendszerek ugyan kézi beavatkozás nélkül képesek az artikuláció kialakítására, de a fotorealisztikus minőségű videó képkockáinak számítása időigényes művelet. Csak nagy teljesítményű számítógépek, illetve szerverhálózatok képesek a videó valós idejű előállítására. A megvalósult alkalmazások rendszerint előzetesen rögzített üzeneteket jelenítenek meg. A magyar nyelvű vizuális beszédszintetizátor működésére példák találhatók az alábbi internetcímen:

`http://mazsola.iit.uni-miskolc.hu/~czap/mintak`

10.30. ábra. Ellenszenves és boldog arckifejezés modellje

10.3.1.7. Érzelmi töltetű beszéd modellezése

Zainkó Csaba

A beszéd érzelmi tartalmának felismeréséről a 9.11. fejezetben már szóltunk. A szintetizált beszéd előállításhoz meg kell állapítani, hogy a beszédszintézis paraméterei közül melyek azok, amelyek hatással vannak a beszéd érzelmi töltetére. Az érzelmi megítélést a tartalom nagymértékben befolyásolja, de a szintézis során a bemeneti szöveg sokszor kötött, a beszédszintézis szempontjából egy külső nem változtatható paraméter. A kutatások jelenleg is folynak ezen a téren, a legintenzívebb vizsgálatok arra irányulnak, hogy a milyen paraméterek vannak a tartalmon kívül, amelyek befolyásolják a szintetizált beszéd érzelmének megítélését. A paraméterek meghatározása összetett folyamat, mert maga az érzelmes beszéd felismerése sem megbízható, kb. 60–65% (Tóth et al. 2007), 55–75% (Scherer 2003) sikerességű, függően az érzelem típusától abban az esetben, ha a tartalom nem segít az azonosításban. Scherer (2003) munkája alapján a semleges beszédhez képest következő paraméterek változhatnak meg az érzelmes beszédben:
Alapfrekvencia: átlag, tartomány, változatosság, menet, eltolódás.
Formánsok: F1 és F2 átlagai és sávszélességük, formánspontosság.
Intenzitás: átlag, dinamika, változatosság.
Spektrális paraméterek: frekvenciatartomány, magas komponensek energiája, a zaj erőssége és spektrális eloszlása.
Időtartam: beszédsebesség és annak változatossága.

A kutatások még nem tudták egyértelműen igazolni minden érzelemre minden paraméter hatását. A legpontosabb eredmények a prozódia 3 alapelemére, az alapfrekvencia átlagára, az intenzitás átlagára és a beszéd sebességére születtek.

A 10.9. táblázatban látható néhány példa arra, hogy különböző érzelmek estén a paraméterek hogyan változnak a semlegeshez képest. Az érzelmi töltetű beszéd

10.9. táblázat. A paraméterek változása érzelmes beszédben (Scherer 2003, Zainkó et al. 2008)

	Öröm	Undor	Szomorúság	Félelem	Harag
Az alapfrekvencia átlaga	↗	↗	↘	↗↗	↗↘
Intenzitás	↗	↗	↘↘	↗	↗↗
A beszéd sebessége	↗	-	↘	↗↗	↗
F1	↘	↗	↗	↗	↗

↗: növekszik, ↘: csökken, ↗↗: nagyon növekszik, ↘↘: nagyon csökken, ↗↘: biztosan nő v. csökken

előállításához a beszéd módosításán kívül egyéb nem verbális elemek (nevetés, sóhajtás, morgás stb.) is felhasználhatóak. Például az öröm kifejezését és azonosítását segíti a beszédben elhelyezett nevetés. A szomorúság érzetét például növeli egy megfelelően elhelyezet sóhajtás a mondatban.

Az érzelmi töltetű beszédkeltés szövegfelolvasókban történő megvalósításáról még a 10.3.9. fejezetben adunk meg adatokat.

10.3.2. Az ortografikus magyar szöveg fonetikai átírásának gépi módszere

Kiss Géza

Fonetikai átíráson a hangzó beszéd diszkrét jelekkel való kódolását értjük, vagy mérnöki értelemben annak megállapítását, hogy egy, általában a nyelv helyesírási szabályainak megfelelő szöveghez a beszédhangoknak mely sorozata tartozik. Az így definiált jelsorozatot nevezik fonetikai átiratnak. A beszédtechnológiában a beszédszintézisnél, az adatbázisok készítésénél és a gépi beszédfelismerésnél használnak fonetikai átiratokat. A hétköznapi gyakorlatban az idegen szavak szótáraiban is ilyen átiratok segítik a beszélőt a helyes kiejtés megvalósításában. A szöveghez tartozó hangsor meghatározása első látásra talán nem tűnik nehéznek, hiszen a nyelvet beszélő, iskolázott ember számára ez ritkán okoz különösebb gondot. Rendszerint képesek vagyunk helyesen felolvasni újságokból, könyvekből. De ahogy nemsokára látni fogjuk, a pontos gépi megvalósításhoz összetett módszerekre van szükség, amelyekkel a nyelv számos összetevőjét kell modelleznünk.

10.3.2.1. A fonetikai átírás során kezelendő nyelvi jelenségek

A nyelv elemkészletét több nyelvi szinttel tudjuk leírni: a legmagasabb szinten vannak a mondatok (vagy a beszédben elhangzó, mondatszerű egységek, a megnyilatkozások), ezen belül a szavak, majd morfémák (más néven szóelemek), végül írásban a betűk, illetve beszédben a fonémák. A szavakat leíró betűk, betűkapcsolatok a ma-

gyar nyelv esetén általában a kiejtendő fonémákat jelölik. Ha ez minden esetben így lenne, akkor a fonetikai átírás feladata meglehetősen egyszerű volna. De ahogy már a 4.1. fejezetben is láttuk, ez gyakran nem teljesül, emellett a különböző nyelvi szintek közötti interakciók is további megoldandó kérdéseket vetnek fel. Az alábbiakban példákon keresztül áttekintünk néhány nyelvi jelenséget és a helyesírás olyan aspektusait, amelyekre általában nem figyelünk tudatosan, amikor hangosan olvasunk, de a fonetikai átírás algoritmikus megvalósítása során figyelmet kell fordítanunk rájuk. Mielőtt ennek nekilátunk, tisztázzuk, hogy mit értünk a „szó" fogalmán, amelynek az alábbiakban központi szerepe lesz.

A szó fogalmát több módon is meg lehet határozni. Beszélnek például ortografikus, szociológiai, lexikális, szemantikus, fonológiai, morfológiai, szintaktikai és pszicholingvisztikai szóról; a jelentésükről részletesen ír Packard (2000). A célunknak most leginkább Taylor (2009) megközelítése a legmegfelelőbb. A szavakat ortografikus átírásuk (más néven szóalak), szófajuk és kiejtésük együttesével jellemezzük, két szót pedig akkor veszünk különbözőnek, ha az előbbi paraméterek bármelyikében eltérnek egymástól. Ez a definíció szándékosan nem utal a jelentésre, mivel itt csak a felszíni formák (írás és beszéd) közötti kapcsolat megteremtése a célunk.

A mondatok egyértelműsége. A szövegtől a kiejtett forma felé haladva, először az ortografikus átirat alapján meg kell állapítunk, hogy hol helyezkednek el a szövegben a mondatok, és ezen belül a szavak. A mondatok végét írásban az erre szolgáló jelekkel szoktuk jelölni; a magyarban ez a pont, kérdőjel vagy a felkiáltójel. Mivel a pontot más célokra is használjuk (rövidítésekben, évszámokban stb.) és a többi jel is előfordulhat mondaton belül (párbeszédek leírásában az elbeszélő megjegyzései előtt stb.), ezért szükséges e jelek adott környezetbeli használatának megállapítása a mondathatárok helyes megállapításához.

Szemiotikus rendszerek. A szavak szóalakjának megállapítása banális feladatnak tűnhet, de közelebbről megvizsgálva itt is több problémával találkozunk. Számos írásrendszerben nem használnak szóközt az ortografikus szavak elkülönítésére, ilyenkor ezek megállapítása is külön feladat. Emellett írásjeleinket számos jelölésrendszer céljaira használjuk vegyesen, amelyek közül a szavak hangalakjának jelölése csak az egyik. Ezek a jelentést kódoló rendszerek, más néven szemiotikus rendszerek, ugyanazokkal a karakterekkel dolgoznak, mint amelyeket a kiejtés jelöléséhez használunk, de azokhoz teljesen más módon rendelnek egyedi jelentést (Taylor 2009). Egy hétköznapi szövegben is rendszerint több ilyen szemiotikus rendszert használunk vegyesen. Ilyenek a

- számnevek: tőszámok, sorszámok – esetleg ezreshatárolóval, toldalékkal –, valamint a törtek (15, 1., 10 000, 5-én, 2-szer, 3,1415927, 1/4),
- rövidítések (USB, UNICEF),
- mértékegységek (100 Ft, 5,2 kg),
- római számok (IV.),

- dátumok (1948. 03. 15),
- időpontok (09:45),
- matematikai kifejezések ($a \cdot x^2 + b \cdot x + c = 0$),
- azonosító jelek (sorozatszámok, bankkártyaszámok),
- telefonszámok (003614631111),
- internetcímek (www.tmit.bme.hu),
- e-levél címek (noreply@bme.hu),
- logografikus szimbólumok (%, €, $),
- és számos további jelölésrendszer.

Az összes ilyen szemiotikus rendszert nem lehetséges felsorolni, mivel ez előbbiek mellett korlátlan számban lehet létrehozni, és hoznak is létre újakat. Például egy új tudományterület megjelenésével rendszerint bevezetnek a leírásához egy új jelölésrendszert, vagy akár egy új szolgáltatás igénybevételéhez használt azonosítók is elterjedt jelölésrendszerré válhatnak, ahogy az e-levél címek pár évtizeddel ezelőtti bevezetése óta történt.

A szemiotikus elemek feloldása. A fenti szemiotikus formákat érdemes betűkarakterekkel átírt szavak sorozatává alakítanunk, amely már csak a szavak szótári alakját, illetve azok ragozott formáit tartalmazza. Ez a feldolgozási lépés a szövegnormalizálás, a végeredménye pedig a fonemikus átirat. Egyes kifejezéstípusokhoz speciális hangsúlyozás és (szünetekkel való) tagolás is tartozhat, sőt beszédstílustól függően is többféle módon mondhatjuk ki, például formálisan, illetve hétköznapi módon. Így például a dátumoknál az évet jelölő számjegyekből tőszámnevet, a hónapokat jelölőből a hónap nevét, a napokat jelölőből birtokos esetű sorszámnevet hozunk létre: *1948. 03. 15.* →*ezerkilencszáznegyvennyolc március tizenötödike*. A *09:45* időpontot mondhatjuk *kilenc óra negyvenöt perc*, vagy informálisan *kilenc negyvenöt* vagy *háromnegyed tíz* formában. A telefonszámokat számcsoportokra osztjuk jelentésüknek megfelelően, illetve a jobb érthetőség kevéért tagoljuk. Például *003614631111* →*nulla nulla harminchat egy, négy hatvanhárom, tizenegy, tizenegy*. Más helyzetben egy ilyen számsorozat akár bankszámlaszám is lehet, amit ilyenkor számjegyenként olvasunk fel. Egyes rövidítéseket kifejtünk (*du.* →*délután*), másokat betűzünk (*USB* →*úesbé*), megint másokat betűszóként olvasunk ki (*UNICEF* →*unicef*). Különböző típusú kifejezések időnként egyforma alakkal, de eltérő kiejtéssel rendelkezhetnek. Így például az *IV.* karaktersor olvasata lehet *negyedik*, de oktatási szövegkörnyezetben akár *ismétlő vizsga* is lehet a jelentése. A szemiotikus elemek feloldásához szükség lehet kiejtésikivétel-szótárra is (8.3. fejezet).

A szavak szófaja. A szavak jellemzésébe belevettük a szófajt, mert erre esetenként szükség van a hangsúlyszerkezet és a kiejtés megállapításához. Például az angol *record* szó kiejtése kétféle hangsúllyal lehetséges, attól függően, hogy igei vagy főnévi értelemben szerepel. A magyarban a ritka példák egyike az *egyek* szóalak, amelynek kétféle szófaja és ezeknek megfelelően két eltérő kiejtése lehetséges: [ɛɟɛk] vagy

[ɛɟːɛk]. Ha már rendelkezünk a mondat szavainak szóalakjával és szófajával, rátérhetünk a szavak (izolált helyzetbeli) kiejtésének megállapítására.

A szó és szókapcsolódás kiejtése. Egy szó szóalakja és kiejtése vonatkozásában a betűk és hangok közötti leképezés adja meg a kapcsolatot. Ez a leképezés minden nyelvre más és más. A hangok hangsorba szerveződnek, és a hangkörnyezetüktől függően változnak, hatnak egymásra. A beszédhangok kiejtése egyes hangkörnyezetekben más lehet, mint önmagukban való ejtésükkor, mivel artikulációs szerveink mozgatási korlátai miatt nem vagyunk képesek bármely beszédhangot bármilyen környezetben kiejteni. Ezek a hatások érvényesülnek szavakon belül is és szavak határán is, amikor két szó kapcsolódik össze a mondatban. Itt ismét utalunk arra a tényre, amit már a 4.4. fejezetben említettünk, hogy a beszédben a szavak nem különülnek el egymástól akusztikai szinten, hanem összeolvadva, egyetlen hangfolyamként valósulnak meg egy-egy kiejtési egységen belül.

A hangok egymásra hatása révén a magyar nyelvben a következő hangmódosulások jöhetnek létre a szó belsejében, illetve a szóhatáron:

- zöngésség szerinti hasonulás. Bizonyos mássalhangzók esetében nem állhat egymás mellett két eltérő gerjesztésű hang (Siptár 2006c). Ez a szabály a magyar nyelvben előrefelé hat (zöngésedésre példa: *népdal* →[neːbdɔl]; zöngétlenedésre példa: *tűzhöz* →[tyːshøz]). Vonatkozik az orális zár- és réshangokra, valamint a zár-rés hangokra. Kivétel a szabály alól a [v], ha a mássalhangzó-kapcsolat második eleme (*ösvény* →[øʃveːɲ]), ilyenkor nincs hasonulás.
- képzés helye szerinti hasonulás (*színpad* →[simpɔd]).
- összeolvadás (*adja* →[ɔɟːɔ], *barátság* →[bɔraːt͡ʃːaːg]).
- hangkiesés (*dombtető* →[domtɛtøː]),
- hiátustöltés: két magánhanzó között megjelenő átkötő rövid, [j]-szerű hang, bizonyos magánhangzók esetén. (*fiú* →[fijuː]). Megjegyezzük, hogy a hiátustöltés jelölésére nem tartjuk megfelelőnek a [j] jelölést, mivel véleményünk szerint ez a hang nem egyenértékű a palatális zöngés réshanggal (lásd az 5.2.1.1. fejezetet).
- rövidülés (*jobbra* →[jobrɔ]),
- nyúlás (*lesz* →[lɛsː]),

Ezek az egymásra hatások nem mindenütt valósulnak meg, ahol a hangsor alapján lehetséges volna, tehát automatikus módosító algoritmus nem készíthető. A morfémaszerkezet ismerete szinte minden szabályhoz hozzátartozik, ugyanis morfémahatáron sok esetben nem érvényes a szabály. Néha egyszerűen a nyelvi szokás ismerete (kivételszótár) szükséges a kiejtés megállapításához. Néhány példa:

- az *átjön* szóban nincs összeolvadás: a [tj] hangokat ejtjük, mivel morfémahatáron van a betűkapcsolat;
- a *kiáltás* szóban általában nincs hiátustöltés (ellentétben a *kiabál* szóval), pedig indokolt lenne;

- a *gólja* szóban nincs hasonulás, mivel a [j] hang birtokos személyragban van (ellentétben a *szóljon* szóval.

A hangok egymásra hatása miatt egyes fonémákat a hangkörnyezettől függően az alapváltozattól eltérő hang (a fonéma másik allofónja) valósíthat meg. Ilyenek például:

- *kámfor* →[kaːmfor]; az [mv], [mf] hangkapcsolatokból keletkezik,
- *ing* →[iŋg]; *ng* és *nk* betűkapcsolatokból jön létre,
- *kapj* →[kɔpç]; a [j] zöngétlenedik, ha mássalhangzó előzi meg és befejező helyzetben van. Ez akkor is érvényes, ha a hangsorban zöngétlen mássalhangzó követi.
- *ihlet* →[ixlɛt]; *doh* →[dox]; a [h] hang veláris változata a *h+C*, illetve a *ch+C* betűkapcsolatokból keletkezik,
- *lehet* →[lɛɦɛt]; a [h] hang zöngésedett változata intervokális helyzetből keletkezik.

A fonetikus átírás többféle mélységben végezhető. Megtehetjük, hogy a kiejtés átírásakor az egyes fonémák allofónjait más-más jelekkel jelöljük, vagy azt is, hogy nem különböztetjük meg egymástól a közös fonémához tartozó allofónokat. Az átírási mélységet attól függően választhatjuk, hogy milyen céllal végezzük az átírást. A beszédszintézis példáját felhasználva az artikulációs- és formánsszintézis-technológiáknál (10. és 10.3.5. fejezetek) mindenképpen szükséges az allofónok pontos meghatározása és ezek szintézise. Hiszen például a zöngétlen palatális réshangot (a [kɔpç] szó végén) a zöngés fonémareprezentációjú [j] hangra felcserélve érthetetlen hangzást kapnánk, mert a két hang akusztikai szerkezete gyökeresen eltér egymástól. Az egyszerű (fonemikus) átírás viszont megfelelő olyankor, amikor a szintézishez használt algoritmus gondoskodik róla, hogy az összefűzendő beszédrészletet ugyanolyan hangkörnyezetből vegyük, mint amilyen környezetben a keresett fonéma található. Így automatikusan a helyes allofónt kapjuk. Példát látunk erre a hangkörnyezetet is figyelembe vevő korpuszalapú szintézisnél (a 10.3.7. fejezet). Diádos és triádos szintézisnél (10.3.6. fejezet), ahol a figyelembe vehető hangkörnyezet erősen korlátos (egy hang az egyik irányban, illetve mindkét irányban), szükség lehet az allofónok megkülönböztetésére, máskülönben a diádelemek összefűzésekor a szó természetellenes hangzást kaphat. Mivel a fonetikus átiratból előállítható a fonemikus átirat is, egyszerűen a különféle allofónoknak a közös fonémaszimbólumukra való átírásával, ezért egy átíró készítésekor célszerű az allofónokat is tartalmazó fonetikus átirat előállítását célul kitűzni.

Minden természetes nyelvben vannak rendhagyó szavak, amelyekre nem érvényesek a szokványos graféma-fonéma megfelelési szabályok; ezek a hagyományos vagy idegen nyelvű írásmódot őrző szóalakok. Magyarra közismert példák az *egy* (rendhagyó módon hosszú [ɟː] hanggal ejtve), *Kossuth* (hagyományos), *New York-ban*, *Einsteinnek* (idegen írásmódú) szavak. Kiemelkedő figyelmet érdemelnek a tulajdonnevek (name pronunciation; lásd például Damper–Soonklang 2007, Tóthfalusi 2006),

amelyek változatos eredetük és a befogadó nyelv helyesírásához való eltérő mértékű igazodásuk miatt önálló kutatási területet jelentenek. Egy további fontos jelenség, hogy egyes szavaknak több kiejtése is lehetséges, például tájnyelvi változatok, vagy a spontán, laza ejtés miatt. Ezzel a kérdéssel is foglalkoznunk kell, ha például hanganyagot szegmentálunk, vagy az eredetitől jelentősen eltérő kiejtésű változatot kell felismernünk beszédfelismeréssel.

Szünetszerkezet. A szüneteket a hangsor részeként szoktuk tekinteni, mint egy speciális kiejtési elemet. A fonetikus átírás részeként tehát ezek helyét, esetleg hosszát is meg kell állapítanunk, mivel fontos szerepet töltenek be a mondatokban. Szünetet tartunk a mondatok, frázisok végén, vagy esetenként a hangsúly érzékeltetésének részeként is. A szüneteknek jelentésmegkülönböztető szerepe is lehet (például *csoki, puding* vagy *csokipuding*). A spontán beszédben ezen kívül előfordulnak szünetek nyelvbotlásoknál és gondolkozási megakadásoknál, a szónoki beszédben ismert a hatásszünet annak segítésére, hogy a hallgató jobban rögzítse az elhangzottak jelentőségét. Idetartozik a beszédszintézisben is alkalmazott tagolás kérdése is, amit szünetek megfelelő helyre történő beiktatásával érünk el (Fék et al. 2004).

10.3.2.2. Eljárások a fonetikai átírás megállapítására

Szöveg fonetikai átírásának megállapításakor a fentebb vázolt nyelvi jelenségeket kezeljük számítástechnikai módszerekkel. A következő lépéseket kell végrehajtani (lásd például Olaszi 2002):

- a bemenő szöveg normalizálása; eredmény: csak betűvel írt értelmes szavak vannak a szövegben;
- a szavak kiejtésének egyenkénti meghatározása; eredmény: hangkódokkal ábrázoljuk a szavakat;
- a hangok egymásra hatásának érvényesítése; eredmény: a szavakon belül és szóhatárokon átívelő hangmódosulások megvalósítása (bizonyos hangkódokat átírunk, kihagyunk, betoldunk);
- a hangkódokat megfeleltethetjük hangszimbólumoknak és ezzel eljutottunk a hangtest végleges képéhez, ami akár olvasható is.

A megoldások során gyakran támaszkodunk a természetes nyelvfeldolgozás (natural language processing, NLP) eszköztárára, ezért először ebből mutatunk be néhány számunkra fontosat: a szófaji címkézést, a morfológiai elemzést a szintaktikai elemzést, és a súlyozott véges állapotú átalakítókat (Weighted Finite State Transducer, WFST). A WFST-k alkalmasak arra, hogy az előbb említetteket, sőt a nyelvfeldolgozási lánc egészét egységes keretbe foglalják.

Szófaji címkézés. A szavak szófajának gépi megállapítását automatikus szófaji címkézésnek nevezzük. Ez sokat kutatott terület, és számos jól használható megvalósítást

nyilvánosan elérhetővé is tettek már. Ilyenek például a TnT (Brants 2000), a Stanford Tagger (Toutanova et al. 2003), az LTAG-spinal (Shen et al. 2007), illetve a magyar fejlesztésű Hunpos (Halácsy et al. 2007). Ezek általában nyelvfüggetlen rendszerek, amelyek szófajokkal felcímkézett szövegkorpuszon betaníthatók tetszőleges nyelvre. A Hunpos elsősorban a magyar nyelvre készült, de más, gazdag morfológiával rendelkező nyelvekre is jól használható.

Szófajon szokták a fő szófaji kategóriákat érteni, mint például főnév, ige, melléknév. Az analitikus nyelvek esetében ennél többre nincs is szükség, mivel ezekben alakjukat nem változtató, nem toldalékolt szavakból épülnek fel nagyon kötött szórendű mondatok. Ilyen analitikus nyelvek például a kínai és az angol (bár utóbbi a szintetikus nyelvekre jellemző tulajdonságokkal is rendelkezik). A magyar nyelvre, lévén szintetikus nyelv, a tág kategória megállapítása nem elégséges, mivel egy szó gyakran több morfémából épül fel (szótő, jelek, ragok stb.), amelyek hozzájárulnak a szó mondatbeli szerepének meghatározásához is. Ezt a különbséget jól illusztrálja, hogy például a *látnálak* magyar szó tartalmát angolul négy szóval fejezhetjük ki: *I would see you*. Az angol kifejezést négy egyszerű szófaji címkével jól jellemezhetjük (személyes névmás, segédige, ige, személyes névmás), míg a magyar megfelelőjéhez egy összetett morfológiai címkére van szükségünk. Ennek az összetett címkének a megállapítását morfoszintaktikai elemzésnek nevezzük. Egy ilyen rendszer megvalósítását írja le magyarra például Halácsy et al. (2007).

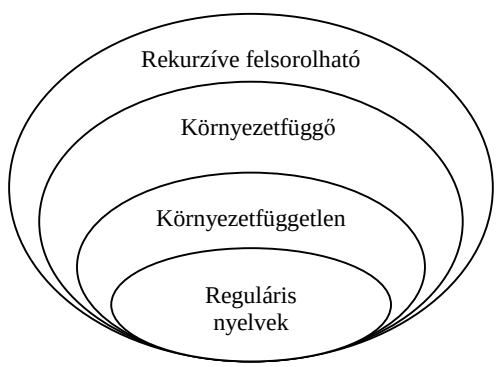
A morfoszintaktikai címkét leírhatjuk például az MSD (Morpho-Syntactic Description) jelölésrendszer (Erjavec 2004) használatával, amelynek létezik a magyar nyelvhez adaptált változata is. Ezt a jelölésrendszert használja például a Szeged korpusz (Csendes et al. 2003), amely többféle szöveget tartalmaz (szépirodalom, fogalmazások, újságcikkek, számítástechnikai szövegek, jogi anyag, gazdasági és pénzügyi rövidhírek).

Morfológiai elemzés. Ahogy korábban láttuk, a morfémák ismeretére szükség lehet a szavak kiejtésének pontos megállapításához is, például hogy eldönthessük, hogy történik-e hangösszeolvadás, vagy megjelenik-e két magánhangzó között a hiátustöltés. Ezek megállapításával a morfológiai elemzés foglalkozik; magyarra lásd például a Humor (Prószéky–Tihanyi 1996, Novák–M. 2006) és a Hunmorph (Trón et al. 2005) rendszereket.

Szintaktikai elemzés. A mondatok szintaktikai szerkezetének megállapítása általában épít a szófaji címkézés eredményére, bár akár egyetlen közös feldolgozási lépést is alkothatnak. A szintaktikai elemzésbe tartozik a mondatok kifejezéseinek és az ezek közötti kapcsolatnak a feltérképezése (parsing). A gyakorlatban környezetfüggetlen, vagy ennél is bonyolultabb nyelvtanokat használnak a szintaktikai elemzéshez. A szintaktikai elemzés segít dönteni egy szó több lehetséges morfológiai címkéje között, ha nem mindegyik illeszkedik egyformán jól az elemzésbe. A szintaktikai elemzés egy megvalósítását magyar nyelvre írja le például Babarczy et al. (2005).

Súlyozott véges állapotú átalakítók használata. A súlyozott véges állapotú átalakítók (WFST) a formális nyelvek egyik csoportjának a megvalósításai, amelynek számtalan felhasználási területe és óriási gyakorlati jelentősége van. Több átfogó leírás található a szöveg-beszéd átalakításban való használatukról (Sproat 1997), illetve általában a nyelv- és beszédfeldolgozásban betöltött szerepükről (Roark–Sproat 2007). Itt nincs módunk az átalakítók elméletét mélyebben tárgyalni, de röviden bemutatjuk őket, hogy a hangátírásban való felhasználásuk jobban érthető legyen. A WFST-k alkalmasak arra is, hogy a fonetikai átírás összes feldolgozási lépését egy egységes keretrendszerbe foglalják, és a többértelműségek feloldására optimális döntést hozzanak. Beszélünk a formális nyelvekről, ezen belül a reguláris nyelvekről, és hogy miért bírnak ezek számunkra kitüntetett szereppel. Ezután bemutatjuk az ezeket megvalósító automatákat, műveleteiket és használatukat a természetes nyelvfeldolgozásban.

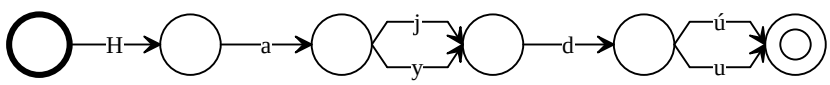
A formális nyelvek a nyelvi struktúrák – például a természetes nyelvek és a programozási nyelvek – leírására használható matematikai modellek. Ebben a modellben úgy értelmezzük a nyelvet, hogy egy véges méretű ábécéből felépíthető végtelen számú, különféle hosszúságú szimbólumsorozatok halmazának egy részhalmaza. (A nyelv ábécéje tartalmazhat például írásjeleket vagy hangszimbólumokat.) A számunkra érdekes nyelvek végtelen számú szimbólumsorozatot tartalmaznak, ezért elemeiket nem tudjuk felsorolni, hanem egy formális nyelvtannal adjuk meg. A formális nyelveknek több csoportját különböztetjük meg aszerint, hogy a nyelvtan milyen típusú szabályokat tartalmazhat. A főbb csoportokat a 10.31. ábrán mutatjuk be. Ezek között tartalmazási kapcsolat van abban az értelemben, hogy például a reguláris nyelvek egyben környezetfüggetlen nyelvek is. Ennek oka, hogy a reguláris nyelvek nyelvtanára szigorúbb korlátozások érvényesek, mint az ábrán látható többi nyelv nyelvtanára; ugyanezért egyszerűbben is kezelhetők. A formális nyelvek részletes tárgyalása megtalálható például Bach (2002) könyvében.



10.31. ábra. A formális nyelvek Chomsky-féle hierarchiája.

A nyelvészek között régóta zajló vita, hogy a természetes nyelvek milyen bonyolultságú formális nyelvekkel írhatók le. A viták fő okát a nyelvi kompetencia és performancia (competence–performance) közötti különbség adja. Erre a különbségre először Chomsky (1965) mutatott rá: Ha egy nyelv nyelvtani struktúráival elvben leírható szerkezeteket nem korlátozzuk bonyolultságukban – például az egymásba ágyazott szerkezetek mélységét engedjük tetszőlegesen nagyra nőni –, akkor ezek elemzéséhez a reguláris nyelveket meghaladó bonyolultságú nyelvtanokra van szükségünk. Viszont ezek az emberek számára is nagyon hamar érthetetlenné válnak. A gyakorlatban előforduló bonyolultságú mondatokra viszont elégségesek a reguláris nyelvek, sőt még ennél egyszerűbb nyelvtanok is. A fonetikai átírás gyakorlati megvalósításához ezért számunkra megfelelőek a reguláris nyelvek, és az azokat megvalósító véges állapotú automaták.

A véges állapotú automaták feladata, hogy bemeneti szimbólumsorozatokat (például természetes nyelvi mondatokat) két csoportba soroljanak: a nyelvhez tartozó, azaz elfogadott, vagy a nyelvhez nem tartozó, azaz elutasított mondatok csoportjába. Az automata által elfogadott összes lehetséges szimbólumsorozat halmazát az automata nyelvének nevezzük. Az automata megjeleníthető egy irányított gráfként, amelyben a gráf egyik csomópontja a kiinduló állapota, legalább egy csomópontja elfogadó állapot, és minden élt az állapotok közötti átmenethez tartozó szimbólum jelöl. A szimbólumok a nyelv ábécéjének elemei lehetnek, valamint a speciális ε (epszilon) szimbólum, amely 0 hosszúságú karakterláncot jelöl. Akkor mondjuk, hogy az automata elfogad egy bemenetet, ha létezik olyan útvonal a gráfon belül, hogy a kiindulási állapotból indulva, majd a bemenet szimbólumaihoz tartozó éleken haladva, valamelyik elfogadó állapotba érkezünk. Egy automatát determinisztikusnak mondunk, ha minden csomópontban egy bemeneti szimbólummal legfeljebb egy kimenő él van megjelölve, tehát minden bemenet esetén az állapotoknak legfeljebb egy sorozatán mehetünk végig; ellenkező esetben azt mondjuk, hogy nem determinisztikus a működése. Az 10.32. ábrán látható determinisztikus automata által elfogadott nyelv a [hɔjduː] kiejtésű vezetéknevek különböző írásmódú változatai. A véges állapotú automaták által elfogadható (reguláris) nyelvek halmaza zárt a nyelvek közötti unió (más néven egyesítés), az összefűzés, a megfordítás, a metszet, a kivonás és a komplementer műveletekre. Egy, a véges automatával rokon, de attól eltérő konstruk-

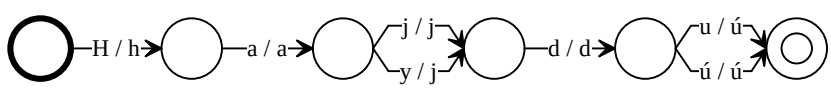


10.32. ábra. Egy véges állapotú automata. Az általa elfogadott nyelv a [hɔjduː] kiejtésű vezetéknév különféle írásmódú változatai. A vastag keretű csomópont a kiinduló, a kettős keretű pedig a végállapotot jelöli

ció a véges átalakító (más néven véges fordító). A véges átalakító feladata, hogy egy bemeneti szimbólumsorozatból egy másikat származtasson. Formailag abban tér el a

véges állapotú automatától, hogy minden éléhez tartozik egy kimeneti szimbólum is. A bemenethez tartozó állapotokon való végighaladás során létrejön a kimeneti szimbólumsorozat, amit akkor tekintünk érvényesnek, ha a bemenet feldolgozásának végén elfogadó állapotba érkezik. A véges átalakítókra példát a 10.33. ábrán láthatunk, ahol a véges automatákra mutatott példánkat módosítottuk úgy, hogy a vezetéknevek kiejtését generáló automatát kapjunk. A véges átalakítók által elfogadott nyelvet – ami reguláris nyelvek közötti leképezés – reguláris kapcsolatnak (regular relation) nevezzük.

A reguláris kapcsolatok zártak a legtöbb műveletre, amelyekre a reguláris nyelvek is zártak, de nem mindegyikre: nem feltétlenül lesz reguláris kapcsolat két reguláris kapcsolat metszete, különbsége, illetve bármelyikük komplementere. Viszont értelmezünk rajtuk néhány fontos műveletet, amelyek a reguláris nyelveken nem értelmezhetők: ezek a kompozíció (composition; jelölése: $W_1 \circ W_2$), a leképezés (projection; $\pi_i(W)$, ahol $i = 1$ a bal- és $i = 2$ a jobb-leképezés; lásd lentebb) és az inverzió (inversion; $inv(W)$). Két automata kompozícióját kapjuk, ha az első kimenetét a második bemeneteként használjuk; ennek használatával egy összetett műveletet több egyszerűbb művelet egymásutánjaként fogalmazhatunk meg. A bal-, illetve jobb-leképezés az átalakító bemeneti, illetve kimeneti nyelvét adja meg. Az inverzió felcseréli az automata élein a be- és kimeneti szimbólumokat. Ha a 10.33. ábrában lévő példánkra alkalmaznánk az inverziót, az utána a [hɔjduː] fonémasorhoz tartozó lehetséges írásmódokat generálná nem determinisztikus módon. Ezeken kívül több, a gyakorlati megvalósításokban fontos műveletet értelmeznek rajtuk, mint például minimalizálás, determinizálás, súlyeltolás (weight pushing) és epsziloneltávolítás (epsilon removal), amelyek az automata méretének, illetve feldolgozási idejének csökkentése miatt fontosak. Megjegyezzük, hogy a nem determinisztikus automaták és átalakítók általában azonos kimenetet produkáló determinisztikus automatákká alakíthatók, kivéve, hogy súlyozott esetben (lásd lentebb) az egyező súlyok előállítása kört tartalmazó gráfok esetén néha nem lehetséges, valamint az átalakítóknál szükség lehet arra, hogy az egy bemenet-több lehetséges kimenet leképezéseket a bemeneti oldalon is megkülönböztessük egymástól jelentéssel nem bíró szimbólumok beszúrásával (Mohri et al. 2002). Súlyozott automaták esetén az automata éleihez



10.33. ábra. Egy véges állapotú átalakító. A bemenetként kapott vezetéknevet annak kiejtett változatává alakítja

egy-egy súlyt is hozzárendelünk. Az állapotokon való végighaladáskor ezeket valamilyen módszerrel kombináljuk, így az elfogadott bemenetekhez egy súlyt is kapunk eredményül. Az élekhez tartozó súlyokon gyakran az élhez tartozó szimbólum adott

helyzetbeli előfordulásának valószínűségét értjük, így az egymást követő élek súlyait összeszorozva az útvonal teljes valószínűségét kapjuk. Nem determinisztikus automata esetben, amikor tehát egy bemenethez több útvonal is tartozhat, gyakran a bemenethez tartozó legvalószínűbb kimenetet, tehát a maximális valószínűséggel rendelkező kimeneti sorozatot választjuk. Gyakorlati megfontolásból ilyenkor a valószínűségek helyett ezek negatív logaritmusával szoktak dolgozni; ennek megfelelően nem szorozzák, hanem összegzik ezeket az értékeket az útvonalon, és a legkisebb súlyú utat választják. Ezt a fajta súlyt „költségnek" is szokták nevezni, mivel a kisebb súly a jobb számunkra, ahogy pénzügyi szempontból is az alacsonyabb költségű megoldást tartjuk jobbnak. A súlyozott automaták egy gyakorlati alkalmazásáról lásd például Mohri et al. (2002) cikkét.

A WFST keretrendszernek a természetes nyelv feldolgozásában való használata számos előnnyel jár a szabálylisták alkalmazásához képest. Először is, gyakorlatilag a teljes feldolgozási lánc megvalósítható egy jól átlátható leíró rendszerben, beleértve a korábban tárgyalt szófaji címkézés és morfológiai elemzés lépéseit is. Másodszor, képes párhuzamosan több alternatívát kiértékelni, amelyekhez eltérő költség rendelhető. Egyes hibás alternatívákat az elemzés előrehaladtával, ahogy új információkat állapítunk meg, ki lehet szűrni; például ha megállapítottuk a szó szófaját, csak az ezzel összhangban lévő kiejtési változatokat tartjuk meg. A megmaradó lehetőségek közül pedig a legkisebb költségű alternatívát választhatjuk. Ez lehetőséget ad az egyértelműsítés (disambiguation) valószínűségekkel való kezelésére. Az egyértelműsítés egy összetettebb, de szintén WFST-k használatán alapuló megoldására példa Dreyer–Eisner (2009). Arra is van lehetőség, hogy a kiejtési változatokat használati módjuk szerint megcímkézzük, például hogy informális vagy formális stílusnak felel-e meg, és megfelelő súlyozással az éppen elvárt beszédmódnak megfelelőket részesítsük előnyben. Harmadszor, alaposan kidolgozott elmélete van a WFST-ken alkalmazott műveleteknek (Roche–Schabes 1997); sőt több szabadon használható szoftvereszközkészlet található hozzá (lásd például Mohri et al. 1997, Beesley–Karttunen 2003, Allauzen et al. 2001). Negyedszer, a determinizált, minimalizált WFST modell nagyon gyors működésre képes: a bemenet hosszában lineáris a végrehajtás ideje. Hátránya, hogy meglehetősen nagy méretű is lehet a modell, ezért nagy rendszerek esetén gyakran nem olvasztják össze előre a modulokat kompozícióval, hanem a feldolgozás során hozzák létre az egyes WFST modulok kimenetének a következő modullal való kompozícióját.

A nem determinizált modellek is gyors működést adhatnak, főképpen ha megadunk egy küszöböt, amely alapján a rendszer lecsökkenti a kiértékelendő alternatívák számát, hogy csak a feldolgozási lépésben legvalószínűbbek maradjanak meg. A küszöb lehet a maximálisan kiértékelendő alternatívák száma, vagy az útvonalak maximális költsége (elvárt minimális valószínűsége). Ilyenkor nem lehetünk biztosak benne, hogy a legjobb megoldást találjuk meg, mivel a feldolgozás egy későbbi szakaszában egy addig kevéssé valószínű megoldás a legvalószínűbbé avanzsálhatna,

de a cserébe kapott gyorsabb működés gyakran kárpótol a pontosságban való valamelyes mértékű visszaesésért.

Szavak kiejtésének megállapítása. Most egy szó kiejtésének a szóalakból való megállapításával foglalkozunk, amelyet a szakirodalomban graféma-fonéma átalakításnak (grapheme-to-phoneme transcription, G2P) szoktak nevezni. Megjegyezzük, hogy kissé más értelemben is szokták használni ezt az elnevezést: vannak, akik a szöveg fonetikus átírásának teljes folyamatát értik rajta, esetleg kiejtési szótár használatának kizárásával, ahogy például Taylor (2009, 219. o.) is említi.

Amikor egy rendszerben szükségünk van az írott szavak kiejtésére, a legkézenfekvőbb megoldásnak tűnhet, hogy a nyelv összes szavát tartalmazó szótárból vegyük a kiejtéseket. Egyes nyelvekre valóban szükséges a köznapi szavakat nagyon nagy százalékban tartalmazó szótárak készítése, mivel a szavak írásmódjában sokkal nagyobb szerepe van a konvenciónak, mint a kiejtéshez való hasonlóság fenntartásának. Ilyen például az angol nyelv, amely szavainak helyesírását viszonylag korán rögzítették szótárakban (az első a XVII. század elején készült), később pedig csak ritkán és kis mértékben változtattak az írásmódjukon, bár a kiejtés időközben jelentősen megváltozott (Carney 1994, 467. o.); továbbá az idegen eredetű szavakat gyakran eredeti helyesírásukkal rögzítették. Az ilyen nyelvek esetében jellemzően nagy méretű szótárakat használnak az alkalmazásokban, melyekben a lexikális hangsúlyok helyét és a szófajokat is tárolni tudják. Az analitikus nyelvek esetén az is hozzájárul a szótárak használhatóságához, hogy a szavaknak kevés alakja van, kisszámú és jól számontartható kivétellel, ezért rendszerint megoldható egy szó összes alakjának tárolása szótárban.

A szótárak használata azonban nem elégséges egyik nyelvre sem. Az egyik oka ennek, hogy a természetes nyelvek szavainak eloszlása LNRE-tulajdonságú (Large Number of Rare Events, nagyszámú ritka eset; lásd például Baayen 2001, 25. o.), ebből fakadóan nagy méretű szövegekben is a szavaknak jelentős százaléka csak egyszer fordul elő (úgynevezett „hapax legomenon”; lásd például Németh–Zainkó 2002). Így egy hosszabb fel nem dolgozott szöveggel való találkozáskor nagy valószínűséggel szerepelni fog benne korábban nem látott szó. A fő oka pedig természetesen az, hogy az élő nyelvek nem statikusak, hanem folyamatosan változnak: a nyelv használói új szavakat hoznak létre, más nyelvekből átvesznek szavakat, új szóösszetételeket képeznek. Emellett számos nyelvre a köznapi szavak jelentős részének felsorolása sem lehetséges. A ragozó nyelvekben, mint a magyar, ugyanis minden szótőhöz nagyszámú toldalékolt alak lehetséges, és ebből számos a gyakorlatban is előfordulhat. Ezért szükség van a szótárokon túlmutató, intelligens megoldásokra.

Egy hagyományosnak nevezhető mód a szavak átírására, hogy a nyelvet jól ismerő szakértő, rendszerint nyelvész, összeállít egy graféma-fonéma átalakítási szabályrendszert, amely a nyelv számos szavához helyes kiejtést képes rendelni. Egyes nyelvekre, mint amilyen a spanyol, finn vagy a magyar, viszonylag egyszerű egy ilyen szabályrendszer összeállítása e nyelvek fonematikus írásmódja miatt. Az ezzel

nem jól kezelt, irreguláris írásmódú szavak tárolására pedig létrehoznak egy kivételszótárt. A kivételszótárba kerülnek rendszerint a hagyományos vagy idegen írásmódú szavak, a családnevek, cégnevek és jövevényszavak (Németh et al. 2003). Ezekhez ma már gyakran elérhetőek precíz, manuális munkával összeállított kiejtési szótárak; lásd például Tótfalusi (2006) idegen szavak kiejtési szótárkönyvét. A szótárt és a grafféma-fonéma átalakítási szabályokat is meg lehet valósítani egy WFST keretrendszer részeként. Ilyen magyarra tudtunkkal még nem létezik.

Az elmúlt évtized egyik kutatási iránya volt olyan gépi tanuló algoritmusok létrehozása, amelyek emberi közreműködés nélkül képesek kinyerni a szótárakban tárolt fonológiai információt, hogy abból létrehozzanak a kiejtési szabályokat és a szótárt is helyettesíteni képes átírót. Ennek több haszna is van: egyrészt kiváltja az emberi szakértelmet, és a szakértők számára is gyakran időigényes és fárasztó szabályjavítgatást, így az algoritmus tetszőleges új nyelvre betanítható további emberi munka nélkül; másrészt a nagy méretű szótárak helyett jóval tömörebb tárolási módot kínál; harmadrészt ismeretlen szavak kiejtésének megbecslésében rendszerint hatékonyabb a kézileg összeállított szabályoknál, azaz jobb az általánosító képessége (Damper et al. 1999). Számos megközelítés létezik a kiejtési szótárból való gépi tanulásra, bőséges irodalommal. A jelenleg legjobb eredményt felmutatók közül néhányat röviden bemutatunk.

A PbA (Pronunciation by Analogy) módszer alapgondolata Dedina–Nusbaum (1991) nevéhez kapcsolódik. Alapvetően egy szó kiejtését ismert szavakhoz való hasonlóság alapján próbálják megállapítani. A módszert az inspirálta, hogy egyes pszicholingvisztikai kutatások arra engednek következtetni, hogy a mi agyunk is hasonló módon oldja meg ezt a feladatot. Az algoritmus megkeresi az összes lehetséges módot ahhoz, hogy összerakja a szót a szótár szavainak töredékeiből, és ezzel párhuzamosan a kiejtését a szótöredékekhez tartozó kiejtésből, a jelöltek közül pedig valamilyen objektív kritérium alapján választ (Soonklang et al. 2008).

Egy másik megközelítés az úgynevezet „joint-sequence model” (Bisani–Ney 2008), amely azon az ötleten alapszik, hogy a bemenet és kimenet együttesen leírható egy közös elemsorozattal, amely mindkettő szimbólumait tartalmazza. A fonetikus átírás során a bemenet a betűk-betűkapcsolatok, a kimenet a fonémák, az ezeket együttesen tartalmazó elemet így grafonémának (graphoneme) vagy grafónnak (graphone) nevezik. A módszer a lehetséges grafónokat és ezek valószínűségét úgy állapítja meg, hogy kiértékeli egy szótár szavainak különböző szegmentálási módjait. Egy új szó kiejtésének azt fogja választani, amelyhez a legvalószínűbb karakterfelbontás tartozik.

Az összes adatalapú megközelítéshez szükség van egy kiejtési szótárra, amely nagy számú szónak tartalmazza az ortografikus és a fonetikus átírását. Emellett szükség van arra is, hogy a kétfajta átirat részei közötti kapcsolatot ismerjük; ennek automatikus megállapítására ad eljárást például Damper et al. (2005), Bisani–Ney (2008). Számos kiejtési szótár elérhető elektronikus formában, különböző formátumokban és

tartalmakkal: egyesek a szóalakok és a hozzájuk tartozó hangsorok mellett tartalmazzák a kettő közötti kapcsolatot is: azaz hogy melyik betűcsoporthoz milyen hangok tartoznak. A használatuknál vigyázni kell arra, hogy az ugyanazon nyelvhez tartozó szótárak is gyakran eltérő hangszimbólum-készlettel dolgoznak.

Magyar nyelvre is elérhető egy kiejtési szótár az interneten (Abari et al. 2006), amely nagyszámú magyar szónak tartalmazza az ortografikus formáját és az annak megfelelő kiejtést IPA jelekkel (8.4.3. fejezet). Néhány idegen nyelvhez tartozó szótár: Amerikai angolhoz használható például a CMUDict (Weide 1998), a Webster szótár (Merriam–Webster 2003), a Pronlex (Kingsbury et al. 1997), a NETtalk korpusz (Sejnowski–Rosenberg 1993) és a TWB (Thorndike–Lorge 1944). Brit angolhoz használható a BEEP (Robinson 1997), az OALD (Mitton–Street 1992) és a CELEX (Baayen et al. 1993). Franciára elérhető a Brulex (Content et al. 1990), a Lexique (New et al. 2001) és a NOVLEX (Lambert–Chesnet 2001). Német nyelvre a CELEX (Baayen et al. 1993) és a LexDb (Lüngen et al. 1998) használható. Ahogy említettük, a kiejtésben célunk lehet tájnyelvi, illetve a spontán beszédben időnként előforduló redukált, sőt egy egyénre jellemző kiejtési formák megjelenítése is. Ezek rendszerint nem szerepelnek kiejtési szótárakban, ezért beszédtechnológiai fejlesztésekhez vagy kézileg kell felvenni őket szótárba, vagy megfelelően kialakított beszédfelismerő szoftverrel megállapítani. Az utóbbit egy mai általános célú beszédfelismerő rendszerint nem tudja előállítani, mivel csak olyan kiejtési változatokat ismer fel, amelyek korábban be voltak táplálva a szótárába.

A mondat kiejtésének megállapítása. A szavak kiejtésének ismeretében a teljes mondat kiejtését bemutató hangszimbólumsorozat a korábban megadott hang-hang átalakítási szabályok használatával meghatározható. Ezek a szabályok nyelvfüggőek, illetve személyfüggőek is lehetnek. A magyar nyelvben érvényesülő hangmódosulási szabályok kategóriáit már korábban több fejezetben is bemutattuk, ezért itt ezt nem részletezzük.

10.3.2.3. Fonetikai átíró magyar nyelvre

Magyar nyelvre általános célú szöveg-beszédhang átalakító szoftver tudtunkkal még nem létezik, ezért egy ilyen létrehozásának tervét adjuk közre.

Az átíró létrehozásának lépései:

- Megválasztjuk a használandó hangkészletet és annak használati szintjeit (egyszerű vagy részletes átíró).
- Létrehozzuk a szövegnormalizálót (szemiotikus elemek feloldása).
- Morfológiai elemző modult építünk be a szóelemek megállapítására. Fontos szerepe van e felbontásnak abban, hogy a szótárban tárolt, de toldalékokkal ellátott szótövekhez is megfelelő kiejtést rendeljünk.
- Szintaktikai elemzőt használunk a kifejezésszerkezet megállapítására.

- Kiejtési szótárt használunk a régies, idegen eredetű és rendhagyó szavak és szókapcsolódások kiejtésének megállapításához.
- Kiejtési szabályokat használunk a szavak hangalakjának meghatározására.

Fontos megjegyezni, hogy az előbb leírt szintek egymással kombinálódhatnak a végleges kiejtési forma megállapítása során.

A hangkészlet megválasztása. Az első lépésben megválasztjuk a használandó hangkészletet, amelyet következetesen kell használnunk a rendszer minden összetevőjében: a kiejtési szótárakban, a beszédadatbázisok hangkészletében és az ezekkel dolgozó programmodulokban. Egy magyar nyelvű beszédszintetizátor hangkészletének megválasztásához első lépés, hogy megállapítsuk, milyen fonémákkal írható le a magyar nyelv. Amennyiben a fonemikusnál részletesebb hangjellemzést szükséges használnunk, meg kell állapítani a fonémákhoz tartozó allofónokat. Magyar nyelvre egy lehetséges hangkészlet leírását lásd a 4.2. és a 10.3.6.1. fejezetekben.

Szövegnormalizálás. A szövegnormalizálás, tehát a fonemikus szóalakok megállapítása történhet WFST keretrendszerben. Ehhez készítünk egy-egy WFST modult a központozás lehetséges értelmezéseinek címkézéséhez (K; például, hogy a pont mondatvéget jelöl vagy rövidítést), az ismert rövidítések megjelöléséhez (R; például az *USB* olyan címkét kap, ami jelzi, hogy ki kell majd betűzni), az arab számok szöveges átírásához (A; például *1* $\rightarrow$*egy*) és a többi kezelendő szemiotikus rendszerhez (például dátumok, római számok, pénznemek; jelöljük ezeket S_i szimbólumokkal, $i = 1..s$). Ezután kompozícióval összekapcsoljuk a bemeneti szöveg egy szakaszát (B) reprezentáló WFSA-t (Weighted Finite State Automaton) e WFST modulokkal, hogy megkapjuk a normalizált szöveget (N):

$$N = \pi_2(B \circ K \circ R \circ A \circ S_1 \circ ... \circ S_s).$$

A B WFSA-ra úgy használjuk a kompozíció műveletét, hogy olyan WFST-nek tekintjük, amelyben a bemeneti és kimeneti szimbólumok minden élen megegyeznek. A sorrend amiatt is fontos, mivel például a K átíró utáni moduloknak számítaniuk kell a központozás esetén a szerepét jelölő címkére, és az S_i moduloknak nem szükséges kifejteniük a számokat, esetleg a kifejtés elvárt módját (például tőszám, sorszám) jelölhetik meg egy címkével.

A rövidítés esetén van más lehetőség is, mint hogy csak egy előre ismert listában lévőket fejtjük ki. Ismeretlen, de rövidítésnek valószínűsíthető (például ad hoc létrehozott) karaktersorozathoz is megpróbálhatunk hozzárendelni egy valószínű kifejtett alakot. A témát részletesen tárgyalja például Sproat et al. (2001).

A szöveg morfológiai és szintaktikai elemzése. A megvalósításunkban célszerű minél több, korábban elkészített összetevőt alkalmaznunk, mivel természetes nyelv feldolgozásáról lévén szó, egy igazán pontos, a kivételeket is jól kezelő megoldás elkészítése hosszú éveket vehet igénybe még kutatócsoportok számára is. Szerencsére magyar nyelvre elérhető nyílt forráskódú szófaji címkéző, a Hunpos (Halácsy et al.

2007), morfológiai elemző, a Hunmorph (Trón et al. 2005) és szintaktikai elemző, a Hunpars (Babarczy et al. 2005), amelyeket felhasználhatunk. Ha szükséges, a programok által még nem kezelt eseteket belevehetjük az adatbázisukba.

Ezeknek az eszközöknek e könyv írásának időpontjában nincs WFST változata, ám így is megoldható a WFST keretrendszerben való használatuk. Megjegyezzük, hogy míg a morfológiai elemzés hatékonyan elvégezhető WFST keretrendszerben, addig az igazán pontos szintaktikai elemzők WFST-ben való megvalósítása kezelhetetlenül nagy méretű modulokat eredményezne (Roark–Sproat 2007). Ezért a szintaktikai elemzés céljára gyakran a reguláris nyelveknél bonyolultabb környezetfüggetlen nyelvtanokat szoktak használni, amelyek használatára már kidolgoztak elfogadható hatékonyságú módszereket. Az ilyen modulok WFST keretrendszerben való használata úgy történik, hogy az őket megelőző modul kimeneteire lefuttatjuk őket (ennek hatékony megvalósítása a „chart parsing”), majd létrehozunk egy WFSA-t a valószínűségekkel súlyozott kimeneteiből, és ezt kompozíció műveletével bemenetként átadjuk a következő WFST modulnak. Ha rendelkezünk a morfológiai és szintaktikai elemzések WFST megvalósításával, vagy a fentebb leírt módon generált kimenetükkel (amelyeket jelölje M és T), akkor a következő módon kapjuk meg az elemzett szöveget (E):

$$E = \pi_2(N \circ M \circ T).$$

Kiejtési szabályok és a kiejtési szótár. A kiejtési szótár létrehozásának feladata abban áll, hogy készítünk egy szótárat, amely tartalmazza a rendhagyó, hagyományőrző és idegen írásmódú szavakat. Ez is óriási munka, amelyben nagy segítséget nyújt, ha fel tudunk használni már elkészült erőforrásokat, mint egy idegen szavak és kifejezések kiejtését tároló szótár (Tótfalusi 2006), és a tudtunkkal egyetlen, nyelvi igényességgel elkészített magyar kivételszótár elektronikus formája (Abari et al. 2006). Ezeknek a birtokában lehetőségünk van arra is, hogy a kiejtési szabályok kézi összeállítása helyett ezeket egy gépi tanuló algoritmussal állapítsuk meg (Bisani–Ney 2008, Dedina–Nusbaum 1991).

Magyar nyelvre a szabályok kézi összeállítása viszonylag kevés munkát igényel, és előnye, hogy könnyebben kontrollálható eredményt ad, mint általában a gépi tanuló algoritmusok modelljei: a betűket és betűkapcsolatokat (például *z, s, zs*) összerendeljük a hozzájuk tartozó hanggal, vigyázva arra, hogy a több karakterből álló kapcsolatok elsőbbséget élvezzenek (hogy például a *zs* karaktereket ne [zʃ], hanem [ʒ] hangként olvassuk ki). Példaként bemutatjuk az erre kidolgozott szabályrendszer néhány szabályát a 10.10. táblázatban. Hasonló szabályrendszert kell felállítani a többi kettős betűkapcsolat által reprezentált beszédhangokra is. A hasonulásokból eredő hangmódosulásokat is hasonló szabályrendszerben lehet kezelni. Példát adunk arra, hogy a *C+j* betűkapcsolatokból milyen hangsorok keletkeznek a kiejtési szabályok szerint a 10.11. táblázatban. Mivel a morfémák meghatározása sok bizonytalanságot tartalmaz (számos lehetséges, de nem valódi morfémát is felismernek az

10.10. táblázat. Az *s* és *z* betűk kombinációiból és környezetükből alkotott betűsorozatok kiejtési formái, ha az *s* a kombináció kezdő betűje. A kiejtési formákat is a betűjelekkel jelöltük. A megjegyzés oszlop értelmezése: morf. = morfémahatáron érvényes; C = mássalhangzó; V = magánhangzó

Betűsorozat=kiejtés	Példa	Megjegyzés
V+s+z+V=V+sz+V	aszal, veszik,	
V+s+z+V=V+zz+V	vonószenekar, Veres Zoltán	morf.
C+s+z+V=V+cc+V	módszer, látszik, gúnyolódsz	C=d, t
C+s+z+V=C+sz+V	unszol, vadszamár, bál szépe, nyolcszög	C zöngétlenedhet
C1+C2+s+z+V=C1+c+V	rendszer, módszer	C1=n, C2=d
C1+C2+s+z+V=C1+C2+sz+V	kardszúrás	morf., C1=l, r, C2 zöngétlenedhet
V+s+z+V=C+zs+z+V	hallászavar	morf.
C+s+z+V=C+zs+z+V	kortárszene	morf., C zöngésedhet
V+s+s+z+V= V+ssz+V	mésszel, tavasszal	
V+s+z+s+V= V+ss+V	készséges, egészség,	
V+s+z+s+V= V+sz+s+V	buszsoffőr, alkuszság	morf.
V+s+z+s+V=V+zzs+V	búzászsák	morf.
V+s+s+z+s+V=V+sz+s+V	dzsesszsarok, összsúly	morf.
C+s+z+z+s+V=C+z+zs+V	bokszzsák	morf., C zöngésedhet
V+s+z+z+s+V=V+zzs+V	színészzseni	morf.
V+s+s+z+z+V=V+zz+V	dzsesszzene	morf.

elemzők), ezért természetesen nem várhatunk 100%-os pontosságot az erre épülő modulok eredményében sem. A kiejtési szabályok és a szótárak egyszerű módon WFST-vé alakíthatók (jelölje őket P és D), és használhatók egy WFST keretrendszerben a fonetikai átírás (H) előállítására:

$$H = \pi_2(E \circ P \circ \mathrm{d}).$$

Egyértelműsítés. A korábbiakban leírt nyelvi példák alapján egyértelműen látszik, hogy a fentebb leírt modulok gyakran több alternatívát fognak szolgáltatni. Tudjuk, hogy a WFST keretrendszerben lehetőségünk van ezek kezelésére, de eddig nem részleteztük, hogy hogyan rendelünk súlyokat az egyes lehetőségekhez, és hogyan szűrjük ki a valamilyen nyelvi kényszernek nem engedelmeskedő kifejtési eseteket. A súlyok megállapítása történhet egy ad hoc szám hozzárendelésével, de célravezetőbb felcímkézett szövegkorpusz gyakorisági számai alapján beállított valószínűségeket használni. Nehézséget jelent az ilyen címkézett szöveg készítése, mert jelentős munkaráfordítást igényel. A könyv írásának idején nem ismerünk olyan kézileg címkézett szöveget, amely tartalmazza a szövegnormalizálás lépéseinek megfelelő címkéket. A WFST keretrendszer arra is lehetőséget ad, hogy a nyelv szabályai alapján lehetetlen konstrukciókat érvénytelennek jelöljük. Például ha a *BT-ban* kifejezésben a *BT* részt kibetűznénk, akkor nem teljesülne a magánhangzó harmónia szabálya. Kényszerek használatával kizárhatjuk ezt a megoldást, hogy helyette a *Betéti Társaság* vagy *Biztonsági Tanács* kifejtés érvényesüljön (az utóbbi főként az *ENSZ* szó utáni helyzetben). Ezeket a kényszereket az őket megvalósító WFST-knek a feldolgo-

10.11. táblázat. Példa a *C+j* betűsorozatok hasonulási kiejtési formáinak megadására. A kiejtett hangokat a betűjelükkel jelöltük. A megjegyzés oszlop értelmezése: kiv. = ha birtokos személyragban van a j hang, akkor az eset ehhez a szabályhoz tartozik; morf. = morfémahatáron érvényes; C = mássalhangzó

Szöveg=kiejtés	Példa	Megjegyzés
d(dd)+j=ggy	adja, védje, feddje	
d(dd)+j=d+j	beszédjel, családjogi, kardjelenet, kedd jó lesz,	morf.
C+d+j=C+gy	áldja, hordja, smaragdja	C nem lehet nazális
C+d+j=C+d+j	biliárdjáték, akkordjavító, rakd jól	morf., C nem lehet nazális
n+d+j=ny+gy	vakondja, kalandja, trendje, csöndjébe	
n+d+j=n+d+j	mindjobban, rendjel, a rend javítása	morf.
t (tt)+j= tty	látja, botja, áldottjai	
t (tt)+j= t+j	pontjóváírás, aszfaltjavító, tengeralattjáró, adott jel,	morf.
C+t+j= C+ty	boltja, sejtje, partján	C nem lehet nazális
C+t+j= C+t+j	partjelző, boltjelentés, aszfaltjavító, aszfalt járda	morf., C nem lehet nazális
n+t+j=ny+ty	bántja, ontja, frontja	
n+t+j=n+t+j	frontjelentés,	morf.
gy (ggy)+j= ggy	rogyjon, hagyjon,	
gy (ggy)+j= gy+j	egyjegyű, higgy Jézusnak	morf.
C+gy+j= C+gy	(C+d+j-vel le van fedve)	
C+gy+j= C+gy+j	hölgyjátékos, a völgy járható	morf., C nem lehet nazális
n+gy+j=ny+gy	(n+d+j-vel le van fedve)	
n+gy+j=ny+gy+j	rongyjelenet, gyöngyjáték, a rongy jó lesz	morf.
ty (tty)+j=tty	bátyját, atyja	
ty (tty)+j=ty+j	pöttyjelölés	morf.
C+ty+j= C+ty	(C+t+j-vel le van fedve)	
C+ty+j= C+ty+j	kortyjelenet	morf., C nem lehet nazális
n+ty+j=ny+ty	(n+t+j-vel le van fedve)	
n+ty+j=ny+ty+j	a konty jól áll, a poronty jár	morf.
n (nn)+j= nny	unja, pattanjon, dzsinnje	
n (nn)+j=n+j	bennjár, kölcsönjegyző, színjáték, kinn jártak	morf.
C+n+j= C+ny	konszernje	C nem lehet nazális
C+n+j= C+n+j	konszernjogok	morf., C nem lehet nazális
ny (nny)+j=nny	anyjukat, hányj	
ny (nny)+j=ny+j	állományjavító, kötvényjegyzés, bizony jól	morf.
C+ny+j= C+ny	-	C nem lehet nazális
C+ny+j= C+n+j	szárnyjavítás, árnyjáték, a szörny jár	morf., C nem lehet nazális
l(ll)+j=jj	teljes, feljebb, hallja	
l (ll)+j=l+j	feljavít, céljellegű , nem hall jól	morf., kiv.
C+l+j= C+jj	ajánljon	
C+l+j= C+l+j	fájljait, görlje	C nem lehet nazális

zási láncba való beiktatásával tudjuk érvényesíteni, célszerűen a legkorábbi ponton, amikor már elérhető a hozzájuk szükséges információ, hogy a rossz alternatívákkal ne dolgozzunk tovább. Megjegyezzük, hogy előállhat olyan helyzet, hogy a rendszerben nincs alternatívája a rossznak látszó kifejtésnek: ha a nyelvtanunk nem kezel egy esetet, vagy a bemenetben lévő elírás miatt jött létre a kényszer megsértése. Ha a rendszer a bemenet minden elemzési alternatíváját kizárja, akkor nem jön létre semmilyen kimenet, ami viszont általában elkerülendő. Ezért kényszerek helyett célszerűbb lehet az alternatívák súlyozásával végezni az egyértelműsítést. További lehetőség, hogy (elírást feltételezve) létrehozzuk a bemenet feltételezett javításának elemzését is kis valószínűséggel – de ezzel rossz elemzéseket is létrehozhatunk – vagy pedig a Karttunen (1998) által bevezetett úgynevezet „lenient composition" nevű konstrukciót használjuk a kényszerek alkalmazására, amely nem hajtódik végre, ha nem jönne létre eredmény az átalakító kimenetén.

10.3.3. Ékezetek gépi helyreállítása

Zainkó Csaba–Németh Géza

A gépi beszédfeldolgozásban sokszor előfordul, hogy olyan magyar szövegeket kell kezelni, amelyek részben vagy teljesen ékezetmentesek. Ilyen szövegek előállhatnak úgy, hogy a felhasználó nem használ ékezetes betűket, mert vagy kényelmetlen neki, vagy az adott eszköz, konfiguráció nem biztosít számára beviteli lehetőséget. A másik eset az, amikor a tárolás vagy adatátvitel során lép fel veszteség, bizonyos ékezetes betűk a konverziók során ékezet nélkülivé válnak.

Példa:

```
Agyunk a beszedet nem onmagaban dolgozza fel, hanem az
osszes erzekszervunkbol kapott informaciot kombinalja
es ertelmezi.
```

Az ember számára az ékezet nélküli szövegek elolvasása nem jelent különösebb nehézséget, az olvasott szöveg eredeti tartalmát szinte teljesen vissza tudja állítani az agy. Az eredmény, hogy néha nem is vesszük észre, hogy nem megfelelő karakter van a szövegben. Ilyenkor a vizuális észlésünk a nyelvi tudásunkra alapozva kikövetkezteti a hiányzó ékezeteket. Az ékezet nélküli, géppel felolvasott magyar szövegeket viszont hallás után nem képes könnyen feldolgozni és megérteni a percepciós rendszerünk, mivel az ékezetek elhagyása a legtöbb esetben hangváltozást is eredményez. A torzított hangtestet nem értjük meg vagy félreértjük. Például a *mögött* szó helyett a *mogott* hangzása értelmetlen, vagy az *ágyát* szó helyett az *agyat* szó mást jelent, lehet, hogy nem illik a mondatba. Ezért szükséges az ilyen szövegek felolvasás elötti újraékezetesítése. A probléma kezelése többfajta gépi módszerrel is megvalósítható.

Az ékezetek és a hangzás kapcsolata. A magyar helyesírásban ékezetet csak a magánhangzókon használunk. Az ékezetek az esetek egy részében főleg csak hosszúságbeli különbséget jelentenek a magánhangzók kiejtésében, míg másik részükben az ékezet hiánya más hangot eredményez. A kiejtésben főleg csak hosszúságban különböző betűpárok az *i-í*, *o-ó*, *ö-ő*, *u-ú*, *ü-ű*, míg más hangot jelölők az *a-á*, *e-é*, *o/ó-ö/ő*, *u/ú-ü/ű*. A megértésnél a hosszúságban való eltérés kevésbé okoz problémát, mint amikor más hang hallható az ékezetesítési hiba miatt. A hibák értékelésekor tehát a hiba típusát is figyelembe kell venni.

Az ékezetek és a jelentés. A magyar nyelvben vannak olyan szóalakok, amelynél az ékezetek nélküli szóalakhoz, több helyes ékezetes szóalak tartozik vagy maga az ékezet nélküli forma is helyes. A *szeket* ékezet nélküli alak értelmes ékezetes variációi a *székét* és a *széket* alakok. Utóbbira példa a *rak-rák* páros. Előfordulhat, hogy a szóalaknak az ékezet nélküli és minden ékezetes formája is értelmes szó (a 10.12. táblázat). Látható a táblázatban, hogy 4 lehetséges alak is előfordulhat egy két szótagú szó esetén.

10.12. táblázat. Az *arat* szó különböző alakjai és előfordulási gyakoriságuk

arat	*árát*	*árat*	*arát*
5,19%	52,67%	41,10%	0,05%

Az ékezetes formák gyakorisága A 10.12. táblázatban látható, hogy az előfordulási valószínűségekben nagyok lehetnek a különbségek. Vannak olyan szavak, ahol egyenletesebbek az eloszlások, ilyen például a *meg-még* pár, ahol majdnem fele-fele arányban fordulnak elő az alakok. Az újraékezetesítési döntéseknél a gyakorisági adatok felhasználhatók.

10.3.3.1. Ékezetesítő eljárások

Három módszert ismertetünk, a gépi tanuláson, a szótáron és a szövegelemzésen alapuló megoldást.

A gépi tanulás lényege, hogy az ékezetek megléte és helye a betűkörnyezetből jó eséllyel jósolható. A módszer lényege, hogy tanító szövegadatbázis segítségével a modellt betanítjuk a különböző magánhangzókra. Külön modellt képzünk az *a, e, i, o, u* betűkre. A tanítás során megadott méretű betűkörnyezetet vizsgál a rendszer, és ennek alapján határozza majd meg, hogy melyik ékezetes forma a legvalószínűbb (Zainkó et al. 2000, Mihalcea–Nastase 2002). A módszer előnye, hogy általánosító képességgel bír, tehát olyan karaktersorozatokra is működik, amelyek a tanító-adatbázisban nem szerepeltek. Hátránya, hogy hibás, a nyelvhez nem tartozó karakteres alakokat is elő tud állítani, illetve hogy a szó pontossága csak 95% körüli.

A szótáralapú megoldás lényege, hogy nagy szövegadatbázis alapján szótárakat építünk, amelyek tartalmazzák a különböző ékezet nélküli szóalakokat és a hozzájuk tartozó, nyelvileg korrekt ékezetes formákat. Ennél az eljárásnál használjuk a gyakorisági adatokat is, ahol több lehetséges szóalak is előfordulhat, ott a statisztikailag legvalószínűbbre dönt az algoritmus. A módszerrel 98%-os pontosság érhető el (Zainkó et al. 2000). Az eljárás magyar szabadalom (Zainkó et al. 2000), lajstromszáma: 226740 P 00 03443. A megoldás előnye, hogy a statisztikában szereplő esetekre nyelvtanilag helyes alakokat ad. Az előkészítő fázisban nagy méretű szövegadatbázist használ az algoritmus, de a működéshez kialakított tudástár (szótár) már kis méretű lesz, és gyors keresést biztosít. Hátránya, hogy nincs általánosító képessége, a szótáron kívüli elemeket nem képes ékezetesíteni. Ha olyan szót kell ékezetesíteni, amely a statisztikák készítésénél használt szövegadatbázisban nem szerepelt, akkor adat hiányában az ékezet nélküli verziót fogja meghagyni, ami értelmetlen szó is lehet.

A szövegelemzésen alapuló módszer nyelvi elemző algoritmusokra épít (Németh et al. 2000). A konkrét megoldásokban használhatunk helyesírásellenőrzőt, morfológiai elemzőt stb. A módszer általában lassabb futású mint az előzőek, minősége a használt elemzők tudásától nagymértékben függ.

Kiinduló adatbázisok. Az első két módszerhez szükséges egy olyan szöveges adatbázis, amelyből az ékezetesítéshez felhasználható információk származtathatók. A gépi tanulásos módszer általánosító képessége miatt, kisebb adatbázisból is tanítható, ez néhány 10 millió szövegszó méretű. A szótáralapú megoldáshoz nagyobb szöveganyag kell, mintegy 100 millió szövegszót tartalmazó szövegadatbázis szükséges a hatásos működéshez. Az adatbázisok nagy mennyiségű írott szövegből készíthetők, és követelmény, hogy helyesen tartalmazzák az ékezetes szavakat. Az ilyen adatbázisok többnyire valamilyen internetes gyűjtés segítségével készülnek. A fenti módszerek csak kisszámú helyesírási és gépelési hibát viselnek el. Amennyiben a hibás alakok száma túl nagy, akkor az ékezetesítés minősége romlani fog (például nagyobb mennyiségű idegen nyelvű szöveg kerül a tanítóadatok közé). Ilyen jelenség fordult elő a magyar, szótáralapú rendszer fejlesztésekor, ahol a ritkán előforduló *dér* magyar szó statisztikáját a szövegben előforduló német részek eltorzították, mert azokban gyakran előfordult *der* névelő. Az adatbázisok mérete nem engedi meg a kézi ellenőrzést és javítást, így a gyűjtési források gondos megválogatásával és gépi ellenőrző algoritmusokkal lehet a minőségen javítani.

Az ékezetesítés menete. Az ékezetesítés menete függ attól, hogy milyen alkalmazásban használjuk. Az alkalmazás tervezésekor számba kell venni, hogy az ékezetek hiánya melyik betűknél forulhat elő, illetve hogy mindig szükséges-e ékezetesíteni, vagy helyes mondatok is szerepelhetnek a bemenő adatok között. Például hibás karakterkódolás használata esetén csak az *ő* és a *ű* átalakulásával kell számolni. Ha SMS-ek felolvasásáról van szó, akkor előfordulhatnak részben ékezetes szövegek vagy teljesen ékezet nélküliek, ha a felhasználó már eleve úgy írta meg az üzenetét.

Azt, hogy az ékezetek hiánya milyen mértékű, legalább mondatszinten kell eldönteni, vagy ennél nagyobb egységeket kell vizsgálni.

Az ékezetesítés a gépi tanulásos módszernél magánhangzónként történik, statisztikai módszernél szavanként. A részben ékezetes szövegek esetén a meglévő ékezeteket figyelembe kell venni, azokat általában nem szabad megváltoztani. Változtatásokra akkor van szükség például, amikor nem magyar ékezetes betűket használ a felhasználó. Az *è* betűt az *é* betűvel helyettesíti, vagy rövid *ö* betűt használ a hosszú *ő* helyett.

Felhasználás. A szótáralapú ékezetesítő rendszer a gyakorlatban is működik több alkalmazásban. Példaként említjük a 12.3.1. fejezetben ismertetett e-levél felolvasórendszert, illetve a 12.3.2. fejezetben olvasható SMS-küldő szolgáltatást, amelyik lehetővé teszi, hogy vezetékes telefonra is küldhessünk SMS-üzenetet (a gép felolvassa).

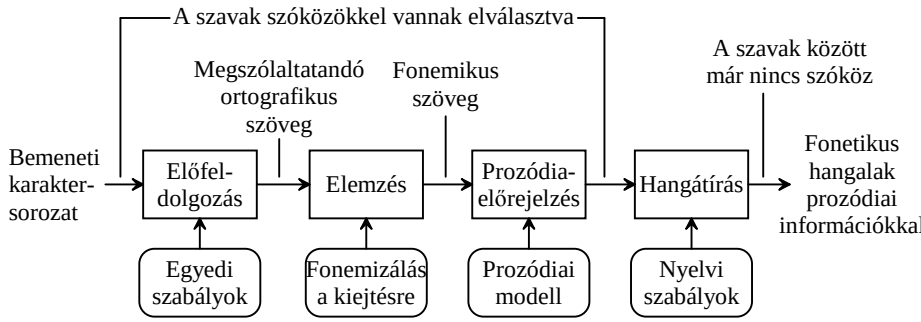
10.3.4. A gépi szövegfelolvasók általános, elvi felépítése

Olaszy Gábor

A gépi szövegfelolvasás általános keretrendszere nyelvfüggetlen. Tartalmaznia kell nyelvi modult és valamilyen fizikai hangelőállítási eljárást. A kettő élesen nem választható el egymástól, egymásra hatással vannak. Ez bármely nyelvre készített ilyen felolvasórendszerre igaz. A részletekben természetesen lényeges eltérések vannak nyelvenként. A szövegfelolvasó fontos része az illesztő egysége is, ami a rendszerintegrálás szempontjából fontos. A beszédtechnológiai szakembernek részletes ismeretekkel kell rendelkezni mindhárom szerkezeti részt illetően, hogy ilyen rendszereket fejleszteni, üzemeltetni, tesztelni tudjon. Ez interdiszciplináris felkészültséget jelent (akusztika, jelfeldolgozás, nyelvészet, fonetika, távközlés stb.). Tisztában kell lenni a kompromisszumokkal, azok kezelésével is. A gépi szövegfelolvasó rendszerek beszédkimenete mindig tartalmaz(hat) hibákat, de nem mindegy, hogy azok milyenek és milyen gyakran fordulnak elő. A hibák minimalizálása gondos előkészítéssel sikeres lehet, de az igényes megoldásoknál sok esetben állandó karbantartást, frissítést, emberi közreműködést kell biztosítani (új készülékek fantázianeveinek kiejtése, a közéleti szereplők változása miatt felmerülő új nevek helyes felolvasása, új márkanevek, vállalkozások, emblémák írott formájának helyes kimondása stb.). Példaként említjük a gépi hírportál megvalósítási gondjait. Miért nincs ilyen műszaki megoldás? Mert a hírek felolvasása az egyik legnehezebb feladat, széles körű műveltséget, nyelvismeretet igényel, amivel a gépi felolvasók egyelőre nem rendelkeznek. A gépi felolvasórendszer minden részelemét valamilyen szinten modellezni kell (példákat láthattunk az előző fejezetekben). A gépi szövegfelolvasók részegységeiben működtetett algoritmusokra a szabály+kivétel elv a jellemző. Lényege, hogy a nyelv működ-

dése csak általánosságban írható le szabályokkal. A szabály lefedi a nyelvi megvalósulások nagy részét, azonban minden esetben lesznek kiejtési kivételek, amikor a nyelvi gyakorlat ellentmond a logikusnak tűnő szabálynak.

A szintézis nyelvi modulja. A nyelvi modul feladata, hogy a bemenetére kerülő karakterláncot (jó esetben ortografikus szöveg) áttranszformálja olyan hangkódsorozattá, ami alapja lehet a kiejtésnek, vagyis a hullámforma előállításának. A nyelvi modul elemzéseket végez, esetleg értelmez(het) is. Az általános nyelvi tudás alapján az írott szövegből minden olyan információt kinyer és jósol, ami elősegíti a hatásos hangelőállítást. A jóslás alapja, hogy az írott szöveg hangzó formáját formalizálva visszavetíti az írásra (például a dallammenetet, a hangsúlyozást, a ritmust). A nyelvi modul alkotórészeit gépi szövegfelolvasáshoz a 10.34. ábra mutatja. A nyelvi mo-



10.34. ábra. A gépi szövegfelolvasó általános nyelvi modulja

dulnak gondoskodni kell mind a szegmentális, mind a szupraszegmentális tulajdonságok meghatározásáról. Vegyük sorra az egyes blokkok szerepét. A bemenetre érkező karakterlánc általában számítógépes rendszerektől származik. Ez a karakterlánc az információt hordozó szövegen kívül tartalmaz(hat) más karaktereket is (vezérlők, azonosítók, soremelések). A nyelvi modul első blokkja tisztítást végez, a bejövő karakterláncból kiválogatja a beszédszintézis szempontjából lényeges elemeket, vagyis a megszólaltatandó szöveget. A tisztításhoz általános és egyedi szabályokat kell meghatározni. Általános szabály lehet például a nyelvdetekció (magyar nyelvű-e a szöveg), az ékezetek ellenőrzése (ékezet nélküli-e a szöveg), a helyesírás ellenőrzése. E három vizsgálat eredményének további feldolgozása bonyolítja a szövegfelolvasó felépítését. Az egyedi szabályok függhetnek a szolgáltatás keretrendszerétől, a kitűzött feladattól, a felolvasandó témakörtől stb. Gondoljunk arra például, amikor elektronikus levelek szövegét kell felolvastatni. A számítógéptől olyan karakterhalmazt kapunk, amelynek csak töredéke az üzenet, a többi adminisztrációs célokat szolgál. A második blokk szövegelemzést végez. Ez abból áll, hogy megvizsgálja az egyes lexikai egységeket és eldönti, hogy azok kiejtése hogyan valósítható meg. Az ortografikus szövegből fonemikus szöveget állít elő (a kiejtésnek megfelelő szavakat

fogja tartalmazni). A legegyszerűbb eset, amikor a lexikai egység egy magyar szó. Ilyenkor azt változatlanul hagyja. Ha nem szót talál, akkor meghatározza az elem (karakterlánc) kiejtési formáját és átírja azt a helyesírás szerinti betűsorozattá. Ezt a műveletet a számok esetében algoritmus végzi, más esetekben úgynevezett kivétellista határozza meg az átírást. Lássunk néhány példát:

13 = tizenhárom; MTA = emtéa; 220 V = kétszázhúsz volt; 8 GB = nyolc gigabájt; EKG = ékágé; BKV = békávé; Rákóczi = rákóci; 1300 mm = ezerháromszáz milliméter; dr. = doktor; stb. = satöbbi; igh. = igazgatóhelyettes; ker. = kerület; SMS = esemes stb.

Belátható, hogy az ilyen átírások fontosak, ugyanakkor az is belátható, hogy nem lehet olyan felolvasórendszert tervezni (ma még) amelyik olyan kiejtésikivétel-listával rendelkezik, amelyik mindenfajta ilyen átírást tartalmaz. Az optimális megoldáshoz három út vezet. Az egyik, amikor az alkalmazó vállalja, hogy rendszeres munkával állandóan bővíti a kivétellistát és megadja az átírási formát (saját kivételszótárt készít és bővít folyamatosan). Ezek az úgynevezett nyílt szerkezetű kiejtésikivétel-listák. A másik út, amikor a fejlesztő határozza meg a kivétellistát. Ebben az esetben ezt a blokkot tanítani kell, mégpedig a leendő alkalmazáshoz tartozó szövegekkel. Sok szöveget kell megvizsgálni, és ezek alapján lehet a leggyakoribb kivételeket meghatározni. Az ilyen listák nem bővíthetők az alkalmazó által, zárt szerkezetűek. A harmadik út, hogy a felhasználó átírja az éppen feldolgozandó szövegben azokat a részeket, amelyek kiejtése kérdéses. Ez a módszer a legmunkaigényesebb és jövőbe mutató eredménye nincs. Egy általános megoldás lehet a nemzeti kiejtési szótár megvalósítása magyarra. Ez független az alkalmazótól is és a fejlesztőtől is. Nemzeti összefogással nagy méretű kiejtési listákat kell gyűjteni, akár témakörönként. Az ilyen gyűjtések elvégzése a jövő feladata. A hírolvasás példájánál maradva azt mondhatjuk, hogy ha az utóbbi 20 évben a hírolvasások minden kiejtési kivételét egy adatbázisba rögzítették volna, akkor sikeresebben lehetne hozzákezdeni egy hírolvasó automata elkészítéséhez, mint ilyen lista nélkül. Az is kérdés, hogy kinek kell ezeket a kivétellistákat biztosítani? Egyes beszédszintetizátor-gyártók az alkalmazóra hárítják az ilyen gondokat. Az alkalmazók (például mobilgyártók, távközlési cégek) pedig nem fordítanak energiát az ilyen részletekre. Mi az eredmény? A felhasználó olyan felolvasást kap, amelyben némely elemek (a kivételek) érthetetlenek. A gépi szövegfelolvasási technológia kialakítói tehát sok esetben rákényszerítik a nyelvileg nem teljes körűen kialakított megoldásaikat az ügyfelekre.

A nyelvi modul harmadik eleme a prozódia előrejelzésével kapcsolatos információkat helyezi el a fonemikus szövegben. Miért van erre szükség? Azért, mert a gépnek már ezen a szinten jelezni kell, hogy milyen prozódiai változásokat kell figyelembe venni a hangsor fizikai megvalósításánál (erre láttunk példákat az előző fejezetekben). A prozódiai modulban lehet állítani a beszédstílust is, annak megfelelően, hogy milyen témájú szöveget akarunk a géppel felolvastatni (hírek, mese, szabályzat, kezelési útmutató, hirdetés stb.).

A negyedik blokk végzi el a fonetikus átírást, vagyis a szövegből beszédhangsorozatot készít kódolási szinten. A kimeneten megjelenik a hangszimbólumok sorozata a prozódiai jelzésekkel. Ennek a blokknak az algoritmusa figyelembe veszi a beszéd és írás közötti különbséget (4.4. és 10.3.2. fejezetek) és hangkódsorozatot állít elő. Felhasználja a magyar nyelv idevonatkozó szabályrendszerét, megvalósítja a hasonulásokat, a hangbetoldásokat és -kimaradásokat (megtartva a prozódiai jeleket). Mindezeken felül minden olyan információt hozzárendel a hangsorozathoz, ami a szintézis technológiájához tartozó modellekből származik (hangidőtartamok, hangintenzitások, az alapfrekvencia változása stb.). Ezen a szinten a kiejtési hangsorozat folyamatos (a hanghullámban a szavak között nincs szünet), a szóhatárok jelzései azonban rejtve megmarad(hat)nak (a prozódiai szerkezet megvalósítása igényli ezt). Elvileg elképzelhető gépi tanulással, közvetlenül a bemeneti vagy ortografikus szövegből a fonetikai átirat kialakítása, azonban magyar nyelvre ilyen megoldást nem ismerünk.

A beszédjel hullámformájának előállítása. A nyelvi modul kimenetén olyan kódsorozat áll elő, amely tartalmazza az előállítandó hangsorozatot, annak minden prozódiai változásával együtt. A hullámforma előállítása függ az alkalmazott szintézismódszertől. A közös ezekben a módszerekben az, hogy mindegyik valamilyen adattárra, illetve elembázisra támaszkodik a beszédépítés során. Az ilyen tárak felépítése, belső szerkezete változó a különböző módszereknél. A beszédhullámra alapozott adattárakra láttunk példát a 8.2. fejezetben. A formánsszintézis és a HMM alapú eljárások adattárai nem hullámformákat tartalmaznak, hanem a beszédre jellemző kivonatolt adatokat, vektorokat (a felépítésüket a módszert leíró fejezetekben tárgyaljuk). A hullámforma előállításával létrehoztuk a bemeneti szövegnek megfelelő szintetizált beszédet.

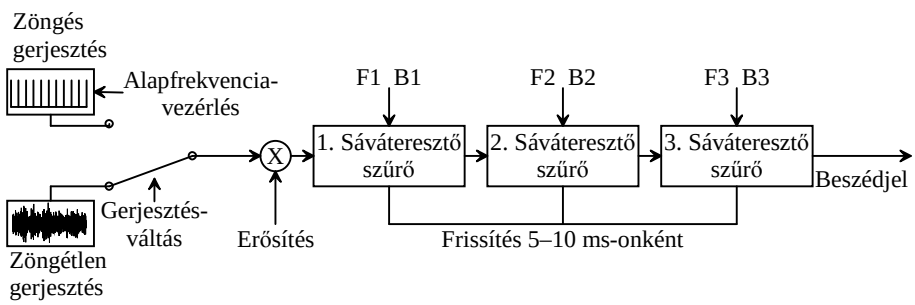
Kódolás, illesztés. A szövegfelolvasó rendszerek gyakorlati használatához azokat már meglévő távközlési, információtechnológiai rendszerekbe kell integrálni. Ennek egyik fontos kérdése, hogy a hullámformaelőállító kimenetén megjelenő hullámformát milyen mintavételezési és kvantálási adatokkal kell megadni. Itt az alkalmazó által használt szabványokat és előírásokat (protokollok) kell figyelembe venni. Ezekről részletesen a beszédinformációs rendszerek tárgyalásánál szólunk (11. fejezet).

10.3.5. Formánsszintézis

Olaszy Gábor

A formánsszintézisnek nevezett eljárás az ember hangképzését utánozza, fonetika elvű megközelítéssel, de a technikai lehetőségekhez igazítva. Ezt úgy kell érteni, hogy egy elektronikus hanggenerátorból és egy ehhez kapcsolt hangolható szűrőrendszerből áll. A vezérlő paramétereket az ismert beszédből, méréssel alakítják ki (míg az

emberi beszédképzésnél fordítva van, az artikuláció alakítja ki a hangzást). Ez volt az első olyan beszédszintézis-technológia, amelynek segítségével egy írott szöveget automatikusan, érthető beszéddé lehetett alakítani (Rabiner 1968). Az 1970-es évekre kialakult a technológia ipari szintű megvalósítása (Liljencrants 1967). A módszer lényege a gége és az artikulációs csatorna pillanatnyi állapotainak elektronikus utánzása. A gége szintjén a hangszalagok által keltett nyomáshullámnak megfeleltethető egy elektronikus (szoftveres) hanggenerátor, amely zöngés hangok esetén periodikus impulzussorozatot ad ki, zöngétlen hangok esetén pedig olyan zajszerű jelet amilyen a gégében keletkezik a fúvó állásnál. A generátor jelének mindenkori amplitúdója adja a hangerősséget, a pillanatnyi frekvenciája (zöngés hangoknál) a beszéd dallamát és egyéb alapfrekvencia-változásait. Ha a generátor jele zaj, akkor zörejhangok állíthatók elő. Ha a generátor jelét nullára állítjuk, akkor a néma fázis hangelemet hozhatjuk létre. Az artikulációs csatornát egy elektronikus szűrőrendszerrel modellezik. A szűrők rezonanciafrekvenciái hangolhatók, és a megadott formánsfrekvencia környezetében erősítik a generátor jelének frekvencia-összetevőit, ezzel modellezve a szájüreg által alkotott rezonátorrendszer módosító hatását. Az első három formánsfrekvencia és formánssávszélesség elégséges egy beszédhang pillanatnyi leírására. A szintetizált beszéd hullámformája tehát egy gerjesztett szűrőrendszer kimeneteként áll elő. A formánsszintézis szűrőrendszere lehet soros (MEA 1982), párhuzamos (Allen et al. 1987) és a kettő kombinációja. Egy soros formánsszintetizátor elvi blokkjait mutatja a 10.35. ábra. A szintézis folyamán beállítandó formánsfrekvenci-



10.35. ábra. Soros formánsszintetizátor blokkvázlata

ákat a szövegből előállított fonetikus hangátiratnak megfelelő hangsorozat határozza meg. A formánsértékek táblázata (10.13). lefedi a beszéd frekvenciatartományát és a szükséges formánsok számát. A frekvenciaadatokból látszik, hogy a tervezők a telefonsávot kissé meghaladó frekvenciatartományra tervezték a rendszert. A kódtáblából leolvashatók az egyes formánsok hatókörei, valamint a formánssávszélességeké is. Érdekes megfigyelni a formánsok közötti frekvencialépések beállítását is. Az alsó frekvenciákon a lépések kicsik, a felsőbbeken egyre nagyobbak. A legnagyobb frekvencia átfogása az első két formánsnak van, a magasabbak egyre szű-

10.13. táblázat. A MEA 8000 formánsszintetizátor-chip formáns- és formánssávszélesség-adatai Hz-ben

kód	F1	F2	F3	F4	F5	B1	B2	B3	B4	B5
0	100	500	1500	2550	3900	3000	3000	3000	3000	3000
1	109	526	1690	2766	4600	600	800	600	700	800
2	119	554	1903	2999		303	433	190	265	335
3	130	583	2143	3253		153	234	60	100	140
4	141	614	2414	3528		77	126			
5	154	646	2719	3826		39	68			
6	168	680	3063	4149		20	37			
7	183	716	3450	4500		10	20			
8	199	754								
9	217	793								
10	237	835								
11	258	879								
12	282	925								
13	307	974								
14	335	1025								
15	365	1079								
16	398	1136								
17	433	1195								
18	472	1258								
19	515	1324								
20	561	1394								
21	612	1467								
22	667	1545								
23	727	1626								
24	793	1711								
25	564	1801								
26	942	1896								
27	1027	1996								
28	1119	2101								
29	1220	2211								
30	1330	2328								
31	1450	2450								

külnek. Egy-egy beszédhang előállításához 5–10 ms-onként kell a szűrőrendszert új formánsadatokkal frissíteni (ezeket hangszeleteknek nevezik). Figyelem! A formánsok változtatásával egyidejűleg az erősítési paramétert is változtatni kell, hogy kompenzáljuk a formánsmozgásból adódó változásokat. A frissítést szabályrendszer végzi. A szűrősor kimenetén előáll a szintetizált beszéd hullámformája. A prozódia megvalósításához egy további szabályrendszer szükséges. Ez megmondja, hogy milyen amplitúdókat, alapfrekvenciát, hangidőtartamokat kell beállítani a vezérlés során hangszeletenként annak érdekében, hogy a természetes beszéd prozódiai szerkezetét utánozzuk. A kimeneti hangintenzitás kialakításánál komplex adatmeghatározásra van szükség, ugyanis az két tényezőtől függ: a gerjesztőjel amplitúdójától és a szűrőrendszer pillanatnyi erősítésétől (ha a formánsok nagyon távol vannak egymástól, az erősítés kicsi, ha közelítjük őket egymáshoz, akkor nagyobb). A formáns-

mozgás tehát befolyásolja a kimeneti hangintenzitást. Ezért a kimeneti jelamplitúdó helyes beállításához az előbbi két hatásból kell keretenként kiszámolni a helyes amplitúdókorrekciót. Az egyes nyelvek beszédhangjaira jellemző formánsmenetek a fonetikai szakirodalomból ismertek, tehát emberi beszéd felhasználása nélkül is tervezhetünk formánsszintetizátort. A finom beállításokat azonban csak percepciós teszt alapján lehet elvégezni. A formánsszintetizátor megfelelő vezérlésével jó minőségű gépi beszéd állítható elő. Ugyanakkor ilyen vezérlő információt általában csak természetes beszédjelből, félautomatikus módszerek segítségével lehet megfelelő finomsággal előállítani. Ez azért van, mert a beszéd változékonyságát szabályrendszerrel nehéz modellezni. A bemeneti szöveg elemzése alapján, szabályrendszerekkel előállított vezérlőinformációk érthető, de erősen gépies hangzású beszéd előállítását teszik csak lehetővé. Ugyanakkor laboratóriumi körülmények között bebizonyították, hogy ezzel az eljárással is mód van az emberi hangot megtévesztésig utánzó beszéd előállítására (Stevens 1992).

A formánsszintézis előnyei. Egyszerű műszaki megoldás, kis memóriát igényel (egy nyelv hangszelettára pár ezer byte-nyi memóriában elhelyezhető). Különböző hangszínezetek könnyen előállíthatók (suttogás, rekedt beszéd, telefonsávi hangzás). Az alapfrekvencia széles határok között állítható (akár ének megszólaltatható). Többnyelvű szintetizátor is készíthető ugyanazzal az alapstruktúrával, csak a vezérlő kódokat kell a másik nyelvre meghatározni. Jól lassítható, gyorsítható. Hátránya, hogy a szövegfelolvasás csak gépies hangzással valósítható meg (bár elviekben előállítható sokkal emberibb hangzású beszéd is, csak a vezérlőparaméterek beállításához kell megfelelőbb modellt készíteni). Elkészítéséhez mély fonetikai tudás szükséges. Magyarországon az első szövegfelolvasó rendszer formánsszintetizátor technológiával készült az MTA Nyelvtudományi Intézetében (Kiss–Olaszy 1984). Hangja meghallgatható a `http://magyarbeszed.tmit.bme.hu` honlapon is. A BME-TMIT jogelődjének számító Híradástechnikai Elektronika Intézetben fejlesztették ki a MultiVox többnyelvű beszédszintetizátort (Olaszy 1989b). Az ingyenesen letölthető magyar beszédszintetizátor is ilyen elven működik (10.3.5.1. fejezet).

10.3.5.1. A MultiVox formánsszintetizátor szövegfelolvasáshoz

A formánsszintézis elvét követi a Magyarországon fejlesztett MultiVox elnevezésű szövegfelolvasó rendszer, amelynek alapjait az 1980-as évek első felében fektették le az MTA Nyelvtudományi Intézetének Fonetikai Osztályán végzett alapkutatások (a magyar beszédhangok és hangkapcsolódások formánsszerkezetének vizsgálatai formájában). A rendszer programozási munkáit a BME Távközlési és Telematikai Tanszékén (jogutódja a BME TMIT) végezték (Olaszy–Gordos 1987). A rendszer először magyar nyelven szólalt meg, majd fokozatosan további nyelvekkel bővült (Olaszy et al. 1989). Az 1990-es évek elejére finn, német, holland, olasz (Olaszy

1991a), portugál (Freitas et al. 1998), spanyol (Gonsalo et al. 1993) és eszperantó (Olaszy et al. 1988) nyelven is beszélt a MultiVox. Akkoriban alapvetően szoftver+hardver elven működtek a beszédszintetizátorok, a vezérlést a számítógép adta, a beszéd pedig egy külső aktív hardveren szólalt meg (10.36. ábra), amit a számítógép párhuzamos portjára csatlakoztattak (Olaszy et al. 1992). A MultiVox magyar változatának hangját percepciós teszttel is minősítették (Gósy–Olaszy 1991). A Mul-

10.36. ábra. A MultiVox, 8 nyelven beszélő formánsszintetizátor hangkeltő egysége

tiVox rendszert több nemzetközi kiállításon is bemutatták, a német nyelvű változat licencét két cég meg is vásárolta (Ausztria, Németország). A MultiVox formánsszintetizátor elvét folytatta az a fejlesztés (MultiVox4), amelynek a végeredménye lett az első ingyenes, kis erőforrás-igényű, magyar nyelvű, bárki által szabadon használható szövegfelolvasó szoftver kialakítása, ami szabványos Microsoft SAPI 4.0 vezérlési felülettel rendelkezik és a legelterjedtebb operációs rendszereken alkalmazható. A tiszta szoftvermegoldás tette lehetővé, hogy a technológia terjedni kezdjen. A rendszer 2002 óta letölthető a `http://alpha.tmit.bme.hu/pub/multivox4/` címről. A fejlesztést a Miniszterelnöki Hivatal Informatikai Kormánybiztossága (IKB) támogatta. A következőkben ezt a rendszert ismertetjük. A MultiVox4 beszédszintetizátor ASCII kódolású szövegből olvas, abból állítja elő a beszédet. A kimondandó szöveget magyar nyelven tiszta hangzású, dallamos, ritmusos köznapi férfi vagy női hangzású beszéddel mondja el. Az MS SAPI 4.0 vezérlési felület biztosítja, hogy más programokhoz kapcsolható, azokat beszédkimenettel lehet ellátni.

Ha a rendszer automatikus szövegolvasást végez (újságcikk, közlemény, máshonnan kapott szöveges anyag, levél stb.), akkor a hangzást nem lehet befolyásolni, azt az eredetileg beépített szabályok határozzák meg, hiszen a szöveg beszéddé való átala-

kítása teljesen automatikusan megy végbe (nem nyúlunk bele a szövegbe). Ha saját magunk készítünk elő kimondásra szövegeket, akkor viszont számos opció áll rendelkezésünkre, hogy saját ízlésünk, kívánságunk szerint programozzuk a kiejtendő beszédet. Ilyen opciók például a beszédsebesség, hangmagasság, hangtípusok, suttogás, rekedt beszéd, tagolási fokozatok, hangerőállítás és mondathangosság. A rendszerben három férfi (Miska, Mátyás, Márton) és három női hang (Magda, Melinda, Manci), valamint azok különböző megszólalási formái (suttogó, rekedt stb.) előre el vannak készítve és fantázianévvel ellátva. Ezek a *v0 - *v23 vezérlő karakterekkel meghatározható hangok. Ezeken nem célszerű változtatni! A *v24 - *v31 jelzésű hangok szabadon változtathatók az installált MultiVox4 rendszer alkönyvtárában található *mv4.cfg* konfigurációs fájl megfelelő sorainak módosításával.

A MultiVox4 konfigurációs fájlja. Ez a konfigurációs fájl tartalmazza a MultiVox4 program számára a kezdeti beszédparamétereket. A rendszer az itt beállított paraméterértékekkel indul el. A fájlban 0–31 közötti indexszel 32 hangkarakter definiálására van lehetőség. A pontosvesszővel (;) kezdődő sorokat figyelmen kívül hagyja a rendszer. Így helyezhetők el megjegyzések is. A konfigurációs fájl paramétereinek jelentése a következő.

AMPLITUDE – hangerősség (0-14); 0 = nincs hang, 3 = nagyon halk, 14 = a leghangosabb

WHISPER – suttogás; 0 = normál hang, 1 = suttogás, 2 = rekedt, 3 = kissé rekedt,

PITCH – hangfekvés Hz-ben (40–500); 40 = rekedt, 100 = férfi, 180 = nő, 400 = kisgyermek

ARTICULATION – a tagolás mértéke; 0 = folyamatos beszéd, 1–5 = egyre fokozottabb tagolás a szavakra vonatkoztatva

INTONATION – intonációs fokozatok; 0 = monoton hangzás, 1 = intonáció bekapcsolva

TIMING – beszédtempó; 0 = nagyon lassú; 1 = nyugodt, 2 = normál, 3 = gyors

VOICE – hangkarakter (1–32); a rendszerben a fent leírt hangkaraktereket lehet megszólaltatni.

Ha a konfigurációs fájlban változtatni akarunk (például meg akarjuk változtatni a kezdő hangmagasságot), akkor egy tetszőleges szövegszerkesztő rendszerrel az *mv4.cfg* kiterjesztésű fájlt behívjuk, a változtatást elvégezzük, majd így tároljuk. A MultiVox4 program ezután minden elindulásnál a konfigurációs fájlból az általunk beállított új értékkel kezd el működni, beszélni. A beállításokat mondatszinten lehet változtatni. A parancsjelző egy csillag (*) karakter. Ennek begépelésével és utána egy betűjel és egy szám megadásával állítjuk be az opciót. A mondatszintű opció csak új mondat kezdésekor alkalmazható, és addig hat, amíg egy másik mondat elején át nem állítjuk.

Példa: *v0Jó napot. Most Márton hangján szól. *v8 Jó napot. Ez Mátyás suttogó hangjával szólal meg.

A MultiVox4 kivételszótárkezelő programja. A szintetizátor használója bővítheti a kiejtésikivétel-szótár egységeit, tehát testre szabhatja a rendszert. Az idegen nevek, kifejezések kiejtését a magyar karakterekkel lehet meghatározni. Ily módon akár idegen nyelvű szavakat is ki tud mondani magyaros kiejtéssel a rendszer.
Példaprogramok. A rendszer felhasználási lehetőségeinek bemutatására néhány felhasználói példaprogramot is tartalmaz a MultiVox4 csomag.

Beszélő óra és dátum felolvasása, hangos emlékeztető. A programmal időközönként a számítógép órája alapján hangos tájékoztatást kaphatunk a pontos időről, az aktuális dátumról, valamint beállíthatunk egy emlékeztető időpontot is. A menüből kiválaszthatunk egy tetszőleges hangot; az óra ezt fogja használni.

A Dátum menüpont kiválasztásával kérhetjük az aktuális dátum felolvasását, a hét napjával és a pontos idővel együtt. A hangos emlékeztető a felhasználó által meghatározott időpontban figyelmeztető beszédüzenetet ad, aminek a szövegét magunk írhatjuk meg.

Folyamatos időtájékoztatás. Az óra a felhasználó által meghatározott időközönként, periodikusan bemondja a pontos időt. A választható időintervallumok: 1, 5, 15, 30, illetve 60 perc.

Felolvasó. Ezzel a programmal szövegfájlok tartalmát szerkeszthetjük és olvastathatjuk fel, soronként, kijelölt részenként illetve egyszerre az egész szöveget. A szöveg szerkesztése során használhatjuk a vágólapot is. A szöveg felolvasásához az Olvasás menüt használhatjuk, melynek részei: egy sor felolvasása (F1), a kijelölt rész felolvasása (F2), a teljes szöveg felolvasása (F3). A felolvasást a Beszéd leállítása menüponttal vagy az ESC billentyű lenyomásával szakíthatjuk meg.

Lapmondó. Az alkalmazással a vágóasztalra helyezett szöveget hallgathatjuk meg magyar nyelven. Ily módon bármilyen alkalmazás (szövegszerkesztő, táblázatkezelő, böngésző stb.) szövegeit kijelölhetjük és megszólaltathatjuk.

A MultiVox4 ingyenesen letölthető a `http://speechlab.tmit.bme.hu` honlapról.

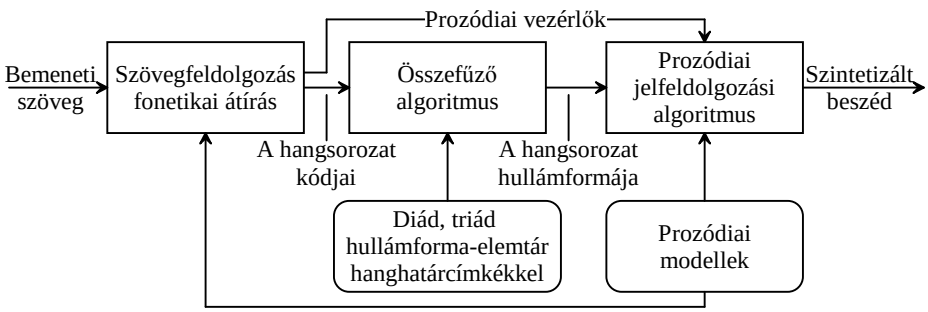
10.3.6. Diád-, triádhullámformák összefűzésén alapuló technológia

Olaszy Gábor–Németh Géza–Kiss Géza

A hullámforma-összefűzés magyar gondolat (lásd a 8.2.2. fejezetet). A beszédszintézistechnológiák fejlődésében (módszertani és technológiai) ez a gondolat radikális változást hozott. Ez jelentette a közeledést az emberi beszédminőséghez. Első elemösszefűzéses rendszerek megvalósítása a 20. század közepére tehető, amikor a számítástechnika már lehetővé tette a megvalósítást (Harris 1953). A fő gondolat az volt, hogy az ember által automatikusan létrehozott beszéd hullámformájában már minden jellemző paraméter megtalálható, tehát nem szükséges a szintézishez a hangképzés finom részleteinek leírásáról gondoskodni hangonként (ahogy ezt a

formánsszintetizátoroknál meg kellett adni). A gondolat nyomán egy teljesen új beszédépítési technológia alakult ki. Az eljárás lényege, hogy emberi beszédből kivágott rövid hullámformarészleteket (diádok, triádok) fűznek össze (8.2.2. fejezet). Az építőelemeket egy hullámforma-elemtár tartalmazza, ebből kell kiválasztani az éppen összefűzendő elemet. A válogatási algoritmus a bemeneti szövegből képzett fonetikai átírás hangsorozatát vizsgálja, és az éppen aktuális hangkapcsolatnak megfelelő hullámformaelemet választja ki az elemtárból, és fűzi hozzá az előzőhöz. Az elemtárról, annak felépítési lehetőségeiről, elkészítési módszereiről már a 8.2.2. fejezetben szóltunk. Az összefűzés után összeáll a bemeneti szövegnek megfeleltetett hullámforma, ami már megszólaltatható. Ezzel a beszéd egyfajta (mondjuk hogy szegmentális) szerkezeti formáját elkészítettük. Kérdés, hogy ez a jel milyen prozódiával szól. Ha zárt rendszerű, kötött szerkezetű hullámforma-elemtárat használunk, akkor minden esetben gondoskodni kell valamilyen utófeldolgozásról, hogy a beszéd dallama, hangsúlyozása, ritmusa is belekerüljön a nyers hullámformába. Ehhez jelfeldolgozási módszereket kell alkalmazni (7.1.4. fejezet), azaz mesterségesen meg kell változtatni esetenként a hangintenzitást, a hangidőtartamot (nyújtás, rövidítés), valamint az alapfrekvenciát (a zöngés hangok periódusidejét). A jelfeldolgozás vezérlését a prozódiai modellel lehet irányítani.

A prozódiai modell fő vezérlési elemeit már a szövegelőkészítés során be kell építeni fonetikus hangátíratba, hogy ezek vezéreljék majd a prozódiai algoritmus működését. Az ilyen beszédépítési technológia általános folyamatábráját a 10.37. ábra mutatja. A gyakorlati megvalósítás lényeges eleme a tesztelés. A prozódiai modul



10.37. ábra. Az elemösszefűzés módszerének fő elemei gépi szövegfelolvasáshoz

szabályrendszerét célzott mondatok felolvastatásával ajánlott tesztelni (sok mondat). Olyan mondatokat kell felolvastatni, amilyenek az alkalmazásban fognak szerepelni. A prozódiai szabályokat és belső paramétereiket addig kell hangolni, amíg az optimális hangzást el nem érjük. A tesztelés fárasztó és nagy szakértelmet kívánó munka. A szövegfelolvasó sikerének a titka a módszeres tesztelés és a precíz javítás, hangolás.

Előnyök és hátrányok. Előny, hogy egyszerű a rendszer szerkezete, emberi hanghoz közel álló hangszínezettel beszél. A beszéd lassítható, gyorsítható, a hangmagasság is állítható, a hangsúlyozásra is különböző mértékek adhatók meg (nagyon hangsúlyozzon, vagy semlegesen, érdektelenül beszéljen), sokféle beszédstílusra készíthető algoritmus, ezek rugalmasan változtathatók, akár mondatról mondatra (szótagolás, magyarázó beszéd, folyamatos felolvasás különböző témakörökre beállítva stb.). A prozódiai szabályok bővíthetők, célzott alkalmazásokhoz bármilyen prozódia előállítható. Az előállított gépi beszéd érthetősége és minősége előre tudható, biztos a végeredmény (kivétel, ha rossz az átírás, például idegen szavak esetén). Több hang előállítása közepes ráfordítással lehetséges (csak a hullámforma-elemtárat kell elkészíteni az új hanghoz). Dialektusok is kialakíthatók kismértékű jelfeldolgozási módosítással. A hibakeresés könnyű, a hibajavítás kis ráfordítással megoldható. Nem igényel nagy memóriakapacitást, a program futása gyors, és nagyon robusztussá tehető. Ezért ideális nagy terhelésű (sokcsatornás), 24 órás információs rendszerekhez. Hátrány, hogy mély szaktudást igényel, csapatmunkában készíthető el. A hullámforma-elemtár címkézése és akusztikai csiszolása sok munkát igényel. A prozódiai teszt és a hangolás munkaigényes. Nem lehet suttogó beszédet előállítani. A prozódia megvalósításához használt jelfeldolgozás rontja a hangminőséget. Egyéni stílusjegyeket nem tartalmaz (a beszélő hangszínezetét csak kissé, a beszédmodorát egyáltalán nem hordozza a megszólaló beszéd). A prozódiai modellezés nehéz.

10.3.6.1. A ProfiVox szövegfelolvasó és fejlesztői rendszere

A ProfiVox volt az első magyar szövegfelolvasó (Olaszy et al. 2000b), amelyik teljesítette azt a négy alapkövetelményt, amellyel egy korszerű beszédszintetizátort jellemezni lehetett 2000-ben.

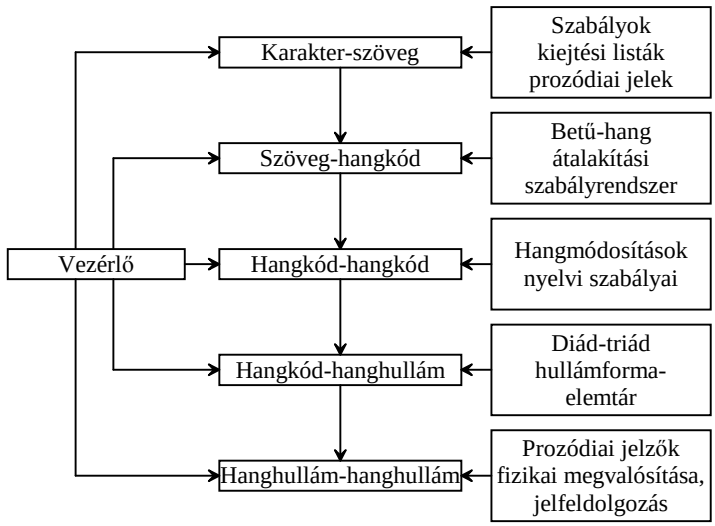
Az első, hogy emberi hangszínezettel beszéljen, amely tiszta, érthető kiejtést hordoz, a szintetizátor beszéde dallamos, és ritmikailag is változatos. Más szóval a beszéd első hallásra is megérthető, és a hosszabb szövegek hallgatása sem fárasztja nagyon a hallgatót.

A második tulajdonság, hogy illeszkedjen a korszerű, általános technikai háttérhez, ezen megszólaltatható legyen, azaz csak szoftvereszközökkel állítsuk elő a beszédet.

A harmadik fontos jellemző, hogy jól kiépített háttértámogatás (szoftvereszközrendszer) álljon rendelkezésre a fejlesztéshez, a módosításokhoz, a rugalmas adaptáláshoz, valamint a rendszerfelügyelethez.

A negyedik a minőségbiztosítás, ami annyit jelent, hogy a rendszer beszédminőségét kialakító algoritmusok helyességét percepciós tesztek jó eredményei támasztják alá.

A szövegfelolvasót a BME TMIT beszédtechnológiai laboratóriumának kutatói fejlesztették ki. A szövegfelolvasó szoftver elkülönített blokkokból áll, amelyek között az adatáramlást a keretrendszer biztosítja. A szöveg-beszéd átalakítási folyamatot a 10.38. ábra mutatja. A modulok függetlenítésével lehetett elérni, hogy a kü-



10.38. ábra. A ProfiVox szövegfelolvasó felépítése és fő moduljai

lönböző feldolgozási szintek helyes működését külön-külön is, csoportokban is és teljes egészében is ellenőrizni lehet a fejlesztés során. Ez a felépítés tette lehetővé, hogy a későbbiekben rugalmasan lehetett a beszédparamétereket, a stílust stb. állítani a felhasználás során, vezérlő karakterek használatával. A rendszer hangkészlete 39 elemből épít fel hangsorokat. A hangkészlet kialakításánál az volt az elv, hogy csak olyan hangokat láttunk el külön kóddal, amelyek akusztikai szerkezetben különböztek egymástól. A magánhangzóknál 9+5 hangkód szerepel (10.14. táblázat), meg kell különböztetni a rövid-hosszú hangokat. A hangsorban előforduló szünetek jelölésére az 1 számkódot használjuk. Az *á* betű oszlopában kétfajta hangjelölés fordul elő, a rövid és a hosszú változat. A *bájt, báj* szavak magánhangzójának korrekt átírásánál rövid, illetve hosszú magánhangzót kell alkalmazni. Ennek hangkódja egységesen a 2-es szám. Mivel spektrális különbség nincs a két hang között, csak időtartami, az alapállapotúnak a hosszú hangot tekintjük (ez fordul elő gyakrabban), ezt rövidíti le egy algoritmus, amikor rövidíteni kell (például a *sztrájk* szó átírt formája: strAjk). A mássalhangzókra 23 kódot használunk (nincs külön kódja a [d͡z] és a [d͡ʒ] hangoknak), viszont külön hangkódja van a [j] hang zöngétlen változatának. A hangkódok az algoritmusban képviselik a hangot szám formájában. A hangjelölés az E1-jelrendszer szerinti. Mindkét jelölést értelmezi a számítógép. A megvalósí-

10.14. táblázat. A magyar magánhangzók ábrázolása a ProfiVox szövegfelolvasóban

betű	á	a	o	u	ü	i	é	ö	e	ó	ú	ű	í	ő
hangkód	2	3	4	5	6	7	8	9	10	34	35	36	37	38
hangjelölés	A:, A	a	o	u	U	I	E:	O	e	O:	u:	U:	i:	O:

tott rendszerben a hangkódot és a hangjelölést az algoritmus különböző, egymástól elkülönülő szinteken használja. A kettőspont a hosszú hangot jelöli. A hangjelöléssel megadott hangsor viszonylag könnyen olvasható is. A rendszer hangkészletében csak rövid mássalhangzók szerepelnek. A mássalhangzók hosszú változatainál nincs spektrális különbség a rövidekhez képest, ezért felesleges külön hangkódot fenntartani számukra. A mássalhangzók nyújtását a gyakorlatban egy külön algoritmus végzi azokon a helyeken, ahol hosszú mássalhangzót kell ejteni. A mássalhangzókat és jelölési rendszerüket a 10.15. táblázat mutatja. A rendszer nem tartalmazza

10.15. táblázat. A magyar mássalhangzók ábrázolása a ProfiVox szövegfelolvasóban. A *j# megfelel a pj#, kj# stb. betűkapcsolatokból keletkező hangnak

betű	b	p	d	t	g	k	gy	ty	m	n	ny	j	h
hangkód	11	12	13	14	15	16	17	18	19	20	21	22	23
hangjelölés	b	p	d	t	g	k	G	T	m	n	N	j	h

betű	v	f	z	sz	zs	s	c	cs	l	r	*j#		
hangkód	24	25	26	27	28	29	30	31	32	33	39		
hangjelölés	v	f	z	s	Z	S	c	C	l	r	J		

a *dz* és *dzs* betűkapcsolatok hangmegfelelőjét, mivel azok a diádelemek megfelelő akusztikai kialakításával létrehozhatók a megfelelő zárhang és réshang belső elemeivel. A hangelembázisban az ilyen kapcsolatok hangzási sajátosságai jól beállíthatók. Ugyanakkor külön kódja van a zöngétlen palatális réshangnak (*kapj*), amelyre azért van szükség, mert a hangzását csak külön hangként lehet biztosítani. A beszédépítés alapelemei diádok és CVC szerkezetű triádok. A ProfiVox rendszerben a fenti hangkódok határozzák meg a diád-, triádhullámformák hangelembázisának felépítését és nagyságát (lásd a 8.2.2. fejezetben). A hullámforma-összefűző algoritmus is erre az elrendezésre került kifejlesztésre, a hangkódok határozzák meg azt, hogy mely elemeket kell kivenni a tárból és összefűzni. Például a *bál* szó beszédhangsorát diádokból összerakva a következő hangkódsorozat képviseli 1-11; 11-2; 2-32; 32-1. Ha triádokat is tartalmaz a hullámforma-elemtár, akkor a kódsorozat így módosul: 1-11; 11-2-32; 32-1.

A ProfiVox rendszer először férfi hangra készült el, majd további férfi és női hangokkal bővült az évek során. A szövegelemző modulja nyelvészeti alapossággal megtervezett algoritmusokat tartalmaz, mind a fonetikai átíró, mind a hangváltozások (hasonulás, hangbetoldás, hangkiesés) tekintetében. A kiejtésikivétel-szótára tartalmazza az általános rövidítéseket, valamint a szövegekben előforduló mértékegységek feloldását. Speciális célalkalmazásokhoz külön kiejtésikivétel-szótárak

állnak rendelkezésre például cégnevek (több ezer), gyógyszernevek és hatóanyagok (több ezer), idegen szavak magyar kiejtése (több tízezer).

Prozódiai modul. A szövegfelolvasó prozódiai modulja a 10.3.1.2. fejezetben ismertetett modellek alapján dolgozik, fel van készítve az alapvető mondatmodalitások, valamint mondathosszak kezelésére, továbbá különböző felolvasási stílusok megvalósítására (hírolvasás, regény felolvasása, név és cím felolvasása, részletező felolvasás szótagolva, magyarázó felolvasás értelmes szövegegységenként). Ez ismereteink szerint az egyetlen magyar szövegfelolvasó, amely a kijelentésen kívül más modalitásoknak megfelelő beszédet is elő tud állítani hullámforma-összefűzéses technológiával. A felsorolásokhoz egyedi prozódiai szabályrendszerrel van ellátva (például nyerőszámok, táblázati számadatok folyamatos olvasása). A hangsúlyok kezelésére kialakított fizikai paraméterváltoztatásokat a 10.16. táblázat tartalmazza. A rendszer alapvető beszédjellemzői hangolhatók külső paraméterként

10.16. táblázat. A hangsúlyok jelzésének fizikai értékei az alapfrekvencia változtatásához

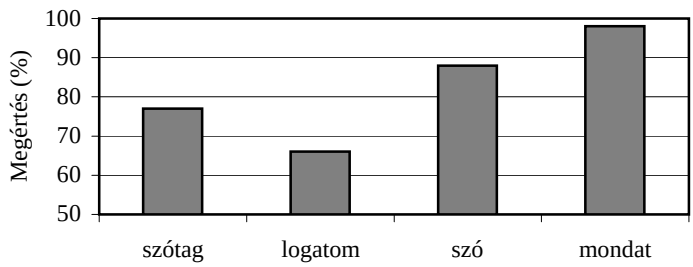
Hangsúlyjelző	Az F_0 csúcs alapfrekvencia-értéke
[:F]	125% a pillanatnyi dallamvonal értékéből számolva
[:E]	120% a pillanatnyi dallamvonal értékéből számolva
[:W]	110% a pillanatnyi dallamvonal értékéből számolva
[:N]	Nincs változás, a pillanatnyi dallamvonal marad meg
[:-]	80% a pillanatnyi dallamvonal értékéből számolva

(hangmagasság, beszédtempó, hangintenzitás, monoton, illetve dallamos beszéd hangsúlyozással, rekedt hangszínezet). A beszéd nagymértékben felgyorsítható torzulás nélkül. A gyorsítás fontos vakok és gyengénlátók részére készített alkalmazásokban (a 12.7. fejezet minden alkalmazásában a Pvox rendszer szól).

Adatmátrix. A ProfiVox szövegfelolvasó adatelőkészítési és fizikai megvalósítási fázisban osztja meg a feldolgozást. A két szint között az adatmátrix teremt kapcsolatot. Az adatmátrix tartalmazza a megszólaltatandó hangsor hangkódjait, a prozódiai jelzéseket, valamint azok fizikai megvalósításának minden részletét számadatok formájában (10.41. ábra). Minden felolvasási stílushoz az adatmátrix belső adatait változtatjuk meg. Így más-más adatmátrix tartozik akár ugyanahhoz a mondathoz is, amennyiben más beszédstílus-szabályrendszert alkalmazunk rá. Az adatmátrixnak lényeges szerepe van a fejlesztésben is, hiszen belőle minden hangsoradatot lát a fejlesztő, sőt a fizikai értékeket reprezentáló számokat meg is tudja változtatni és a változtatás eredményét hangban meg tudja hallgatni. Így a prozódiai szabályok hatásának vizsgálatára, a finomítások, hangolások elvégzésére azonnali visszajelzés alapján tud adatmódosítást elhatározni. Ez az analízis szintézissel eljárás segíti a hatásos szabályfejlesztést, modellfinomítást. A ProfiVox technológia alapján készült lengyel, német és spanyol (Alonso 2004) szövegfelolvasó is.

Akusztikai minőségbiztosítás. A ProfiVox szövegfelolvasó hangminőségét egyrészt percepciós tesztekkel minősítettük, másrészről társadalmi szintű véleménykérés-

sel ellenőriztük (Olaszy et al. 2000a). A percepciós tesztekben 240 egy és két szótagú szót, valamint 48 mondatot (3–6 szó mondatonként) hallgattattunk meg 6 tesztalannyal (4 férfi és 2 nő, életkoruk 30–45 év közötti). Mindkét esetben az volt a feladatuk, hogy írják le, amit hallottak. Az eredmények kiértékelésénél csak azt a szót tekintettük elfogadhatónak, amelyikben nem volt hanghiba. Ezzel a kritériummal 83%-os eredményt kaptunk. Abban az esetben, amikor zárhang tévesztését is elfogadtuk jó azonosításnak (például *Pál* helyett *tál*-t értettek), akkor 93%-os szóazonosítási szintet kaptunk. Mondatok esetében a helyes azonosítás elérte a 98%-ot. Itt utalunk arra, hogy egy sokkal korábbi kísérletben (Gósy–Olaszy 1983) az emberi ejtésű szavak szóértési tesztje 86%-os eredményt hozott. Tehát az ember sem képes hiba nélkül azonosítani rövid, önállóan elhangzó szavakat (10.39. ábra). A kísérletben a szavak egytől három szótagot tartalmaztak, a logatomok két szótagot. A társadalmi szintű értékelést egy mobilszolgáltató végezte a nála bevezetett „Mailmondó” nevű elektronikuslevél-felolvasó szolgáltatásában (Németh et al. 2000). A kérdőíves felmérés során az előfizetőknek 5 fokozatú skálában kellett

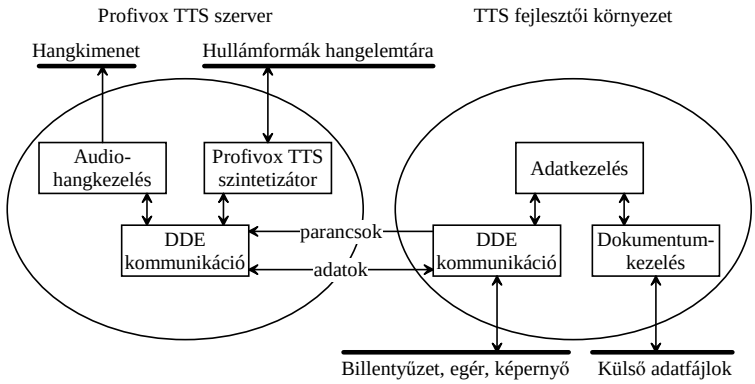


10.39. ábra. Az emberi percepciós rendszer teljesítése különböző hosszúságú és tartalmú, emberi ejtésű magyar hangsorok megítésében (Gósy–Olaszy 1983) alapján

értékelni az elektronikus levélfelolvasó beszédminőségét. A mérés átlageredménye 4,3 volt.

A ProfiVox szövegfelolvasó fejlesztői rendszere. Egy beszédszintetizátor fejlesztése csak akkor végezhető el eredményesen, ha rugalmas fejlesztői környezet támogatja a fejlesztő munkáját. A most bemutatandó fejlesztői környezet elemösszefűzéses technológia támogatására készült a BME TMIT beszédtechnológiai laboratóriumában a 90-es évek közepétől kezdve (Ferenczy et al. 1997, Olaszy et al. 2000b). A ProfiVox beszédmotorját úgy tervezték, hogy fogadni és küldeni is tud adatokat (kliens-szerver). Az adatáramlási diagramot a 10.40. ábrán láthatjuk. A fejlesztői rendszer segítségével a szintetizátor minden szintjéhez hozzáférhetünk. Grafikusan megjeleníthetünk adatokat, időfüggvényeket. Ha megváltoztatjuk az adatokat és azokat visszaküldjük a szintetizátornak, akkor azonnali szintézist kérhetünk az új adatokkal. Egy vezérlőpanel rögzíti a beszédszintetizátor alapparamétereinek értékeit (beszédsebesség, intenzitás, hangfekvéshez definiálandó alapfrekvencia),

valamint kapcsolókat is tartalmaz a különböző modulok ki-be kapcsolására (dallam, hangsúly, hasonulások, szünetek, zönge kváziperiodikussága, amplitúdókorrekciók a magánhangzókban stb.). A rendszer támogatja a hullámforma-elembázis készítését, kezelését, akusztikai csiszolását is. Ez a fejlesztési fázisban elengedhetetlen. Az MDI technológia (Multiple Document Interface) alkalmazása lehetővé tette, hogy különböző típusú dokumentumokat lehet kezelni (hulámforma, szöveg, adat stb.)

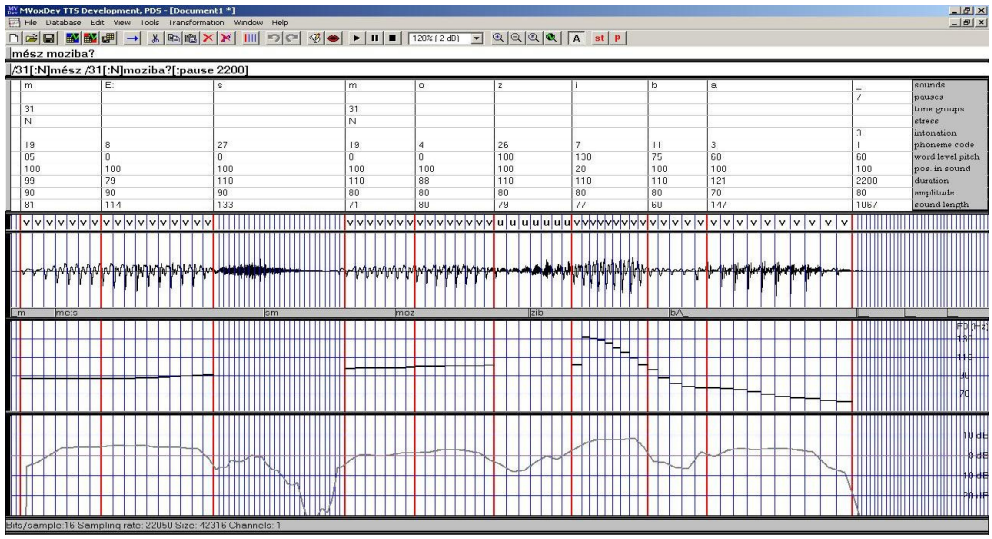


10.40. ábra. A ProfiVox fejlesztői környezete és a szintézismotor kapcsolatrendszere

A fejlesztői rendszer szolgáltatásai a következők:

- Szintézis szövegből
- A szintézis adatmátrixának megjelenítése és dinamikus kezelése
- A szintézis építőelemeinek megjelenítése (diádok, triádok határjelölései)
- Hullámforma-megjelenítés, nagyítással, zöngés-zöngétlen jelölésekkel (v = zöngés periódus)
- Hanghatárok megjelenítése, mozgatása, jelzések elhelyezése a hang belsejében (például egy zöngés zárhangban a zönge és zárfelpattanás határán)
- Intenzitás- és dallamgörbék megjelenítése (a 10.41. ábrán alul és felette)
- A hangelembázis mátrixának megjelenítése
- Meghallgatási formák: teljes szöveg, kijelölt hullámforma, ugyanazon elem sokszori, ismételt lejátszása (akár hangperiódus)

A fejlesztői rendszer egy teljes képernyőjét a 10.41. ábra mutatja.



10.41. ábra. A ProfiVox fejlesztői rendszer képernyője szintéziskor: szöveg, címkézett szöveg, adatmátrix, hullámforma, alapfrekvencia, intenzitás

10.3.7. Elemkiválasztás-alapú szövegfelolvasó

Zainkó Csaba

A rugalmas elemkiválasztás-alapú beszédszintézis (korpuszalapú technológia) adja napjainkban a legjobb minőségű szintetizált beszédet (Schweitzer et al. 2003). A beszédépítéshez nagy méretű, címkézett beszédkorpuszt használunk. Ennél a módszernél a beszélő hangszínezete, kiejtési stílusa egyértelműen felismerhető, szemben a diád-triádrendszerekkel. A diád-triádalapú elemösszefűzéses szintetizátorok által előállított beszéddel érezhető a természetes emberi hang finom részleteinek a hiánya. Ennek egyik oka az, hogy a felhasznált elemkészletben minden diád vagy triád csak egyszer szerepel, tehát ilyenformán sematizálják a beszédhullámot. A másik ok pedig, hogy a beszéd dallamát, hangsúlyozását, ritmikai elemeit utólagos jelfeldolgozással, modellek alapján alakítják ki. Mindkét előbbi feldolgozási forma csökkenti az összeállított beszéd természetes hangzását. A minőségromlás csökkentésére született az a gondolat, hogy a hangelembázist növeljék, az folyamatos beszédet tartalmazzon. Így egy adott beszédépítő hullámformaelemnek több változatát tárolják el és a beszédépítésnél ezeket lehetőleg nem szakítják ki a hangkörnyezetükből (a diád-, triádelemeknél ez nem volt lehetséges), hanem megmaradnak a folyamatos beszédjelben (Sagisaka et al. 1992, Campbell–Black 1995). A különböző változatok más-más akusztikai és prozódiai paraméterekkel fognak rendelkezni. A beszédépítést ezek után egy keresési algoritmussal végezzük el (Hunt–Black 1996), ami kiválogatja a korpuszban eltárolt folyamatos beszédből a megfelelő hullámforma-elemeket

és azokat fűzi egymáshoz. Ennek az algoritmusnak futási időben kell működni, ami erős megkötés. Az eljárás eredménye, hogy a felhasznált elemek közelebb állnak a megvalósítani kívánt prozódiához, mert már az adatbázisok tervezésekor figyelembe vettük ezt a szempontot (lásd 8.2.3. fejezet). Így kevés jelfeldolgozásra van szükség, amely természetesebb hangzást produkál. Minél nagyobb a beszédadatbázis (korpusz), annál nagyobb a valószínűsége, hogy megtaláljuk benne a megfelelő hullámforma-részletet a szintézishez. A válogató algoritmus feltételrendszerének paraméterei, azok súlyozása is meghatározó tényező a sikeres keresésben. A keresés nyitott, bármilyen tartalmú és hosszúságú elem kijelölhető keresésre.

Elemek típusa. Az elemkiválasztásos beszédszintetizátoroknak egyik fontos kérdése, hogy milyen típusú elemekből állítja elő a beszédet.

A következő típusú elemek lehetségesek:

Keretek. Rövid beszédkeretek – például hangperiódusok – amelyeket tetszőleges sorrendben összefűzhetünk (Hirai–Tenpaku 2004).

Állapotok. A beszédhang egy-egy állapota, például egy hangszelet, mint a formánsszintetizátornál (10.3.5. fejezet).

Félhang. Egy hang fele, amely valamilyen kritérium alapján lett elvágva. Lehet a hang közepén, vagy más szabályok alapján, mint például egy zárhang néma fázisa vagy a felpattanása.

Diád. Egy hangkapcsolódás két fél hangja, amely az első hang közepétől a második hang közepéig tart.

Hang. Hanghatárok mentén elvágott beszédhangok.

Félszótag. A szótag fele, amely a szótag közepén történő elvágásából keletkezett. Sokszor a szótag magánhangzójának közepén van a vágási pont.

Dupla félszótag. A diádhoz hasonlóan felépített két félszótag.

Szótag. Szokásos értelemben használt szótagok.

Morfémák. Szótaghoz, a nyelv legkisebb jelentéssel bíró egységéhez hasonló méretű elem.

Szavak. Szokásos értelemben használt szavak.

Frázisok. A prozódiai értelemben használt frázisok.

Ha egy szintetizátorban csak egy típusú elemet használunk, akkor homogén rendszerről beszélünk. Ha különböző típusú elemeket is felhasználunk, akkor heterogén rendszerről beszélünk.

A magyar nyelvre készített első korpuszalapú beszédszintetizátorban a legnagyobb építő elem a szó volt. Ezenkívül hangok keresésére is van mód (Fék et al. 2006). Később elkészült olyan verzió is, amelyik diádokat és hosszabb hangkapcsolatokat használt. A rövidebb elemek használata azért előnyös, mert kisebb lehet a korpusz amelyből a szintetizátor kiválogatja az elemeket (gazdaságosabb számítástechnikai szempontból). Hátránya, hogy az illesztések száma nőhet, ezért az illesztésekkor fellépő torzítások ronthatják a hangminőséget. A kis méretű elemekből több

kell egy adott mondat szintetizáláshoz, mint ha hosszabb elemeket használnánk, ezért a számítási igény nő, a szintetizálás sebessége csökken. A szintetizátor által használt legnagyobb építőelem mérete nem korlátozza azt, hogy mekkora összefüggő részeket tud a korpuszból egy-egy keresésnél használni. Ha tehát a rendszer hangméretű építőelemekből dolgozik, attól még a kész szintetizált mondatba akár szavak és még nagyobb összefüggő részek is bekerülhetnek, mint sok egymás mellett tökéletesen kapcsolódó hang. Az építőelemek mérete tehát hatással van a szintézis sebességére, de ezek mellett közvetetten a szintetizálás minőségére is. Kisebb elemek esetén a prozódiai modell kevésbé működik (lásd a 10.3.1.5. fejezetben).

A cél előírása. A cél a bemeneti szövegből levezetett, egyfajta szimbolikus leírás elkészítése, amelyet a keresés folyamán használunk fel arra, hogy a korpuszban megkeressük a neki legjobban megfelelő beszédrészletet. A cél meghatározásának részletessége függ a keresésben használt elemek típusától. Ha például szó típusú elemeket használunk, akkor elég szószinten meghatározni a célt. Tehát a cél a szavak sorozatából és az egyes szavak tulajdonságaiból fog állni. A tulajdonságok a következőek lehetnek: hangsúlyos-e a szó, milyen prozódiai pozícióban áll (10.3.1.5. fejezet), vagy akár fizikai, akusztikai paraméterek is képezhetik a tulajdonsági adatokat, mint például a szó átlagos alapfrekvenciája vagy időtartama. Ha kisebb elemekből – például hangokból – állítja elő a szintetizátor a beszédet, akkor hangok szintjéig vagy részletesebben meg kell határozni a célt. Itt már szükséges a fonetikai átírás, amely megadja a hangsorozatot, amelyről részletesen a 10.3.2. fejezetben szóltunk. A célként előírt hangsorozat mellé meg kell adnunk a hangok tulajdonságait is, amelyek főleg fizikai paraméterekből állnak. A prozódiageneráló modul állítja elő a legfontosabb ilyen paramétereket, az alapfrekvenciát (10.3.1.2. fejezet) és a hangidőtartamot (10.3.1.4. fejezet). A célfüggvény a korpuszos szintetizátorban hasonló szerepet tölt be, mint a ProfiVox intonációs mátrixa (10.3.6.1. fejezet). A célelőírás szükséges a gépi felolvasók általános modelljében szereplő előfeldolgozó modulhoz is (10.3.4. fejezet). A kivételek, rövidítések, számok és dátumok feloldása nélkül a cél nem lenne meghatározható, vagy annyira eltérne a korpusz tartalmától, hogy a szintetizálás későbbi lépései sikertelenek lennének.

A célköltség és a célfüggvény. A célköltség meghatározásánál különböző információkat használunk. Ezek az elemtípusától függően lehetnek nyelvi-fonetikai információk, illetve fizikai, akusztikai jellemzők. Nyelvi-fonetikai információ lehet például, hogy melyik beszédhangról van szó, hogy hangsúlyos-e az adott szó vagy szótag. A költség meghatározásakor figyelembe vehetjük a mondat modalitását (kijelentő, óhajtó, kérdő), az adott elem mondatbeli vagy frázison belüli pozícióját. Akusztikai információ lehet az adott elem alaphangjának értéke, az elem időtartama, intenzitása vagy például formánsfrekvenciái. A célfüggvény az a függvény, amelyik megadja a cél- és az adott korpuszból kivett elem közötti célköltséget (mennyire hasonlítanak egymásra). Tehát a függvény minden korpuszban szereplő adatbáziselemre meg tudja adni, hogy mennyire hasonló ahhoz az elérendő célhoz, amelyet meghatároz-

tunk a bemeneti szöveg alapján. Ezt a hasonlóságot a célköltséggel fejezzük ki. Azt az elemet amelyiket össze kívánjuk hasonlítani a céllal, jelöltnek nevezzük. A megvalósítástól függ, hogy a célköltségben milyen típusú tulajdonságokat használunk, csak nyelvit, csak akusztikait vagy keverten. A tulajdonságokat lehet logikai függvénnyel kezelni (0 - eltérő, 1 - azonos), vagy valamilyen súlyozó függvénnyel. A súlyozó függvények általában lineárisak, de vannak olyan tulajdonságok – például az F_0 –, ahol a logaritmikus függvény előnyösebb lehet, mert a kisebb különbséget arányaiban kevésbé bünteti, mint a jobban észlelhető nagyobb különbségeket. A cél és a jelölt közötti tulajdonságok közötti hasonlóságot valamilyen távolságfüggvénnyel jellemezzük, tehát a célfüggvényt felírhatjuk a 10.1 egyenlet formájában, ahol távolságfüggvények súlyozott értékeként áll elő a célköltség. Az így kiszámított távolságokat különböző súlyozással (w_p) összesítjük, ami megadja a cél és a jelölt közötti költséget.

$$C^t(t_j,u_k) = \sum_{p=1}^{P} w_p(C_p^t(t_j[p],u_k[p])). \tag{10.1}$$

ahol

t_j a j-edik célelem, amely P tulajdonsággal rendelkezik
u_k a k-edik adatbáziselem (jelölt), amely P tulajdonsággal rendelkezik
$C^t(t_j,u_k)$ célköltsége a j-edik célelem és k-adik adatbázis elem között
$C_p^t(x,y)$ távolságfüggvény az x és az y tulajdonság között.

Másik lehetőség az euklédeszi távolság használata a célköltség kiszámításánál:

$$C^t(t_j,u_k) = \sqrt{\sum_{p=1}^{P} w_p(C_p^t(t_j[p],u_k[p]))^2}. \tag{10.2}$$

A súlyozás összehangolása összetett feladat, kísérletezést igényel. A paraméterek nagy száma miatt (százas nagyságrendet is elérhet), többnyire empirikus úton történik a súlyok meghatározása. Kísérleteznek ilyen tanuló algoritmusok fejlesztésével is (Rutten–Fackrell 2003). A célköltség számításánál tehát minden célhoz meghatározzuk a jelölteket és az azokhoz tartozó célköltséget. A célhoz elvileg bármilyen azonos típusú elemet jelöltnek tekinthetünk, de ez felesleges számítási kapacitást igényelne. A jelölteket első lépésben valamilyen egyszerű szempont szerint szűkítjük, majd csak ezekre a jelöltekre számítjuk ki a célköltséget. Például, ha egy [oː] hangot keresünk, akkor jelentősen szűkíthetjük a keresést, ha csak az [o] és [oː] hangokat tekintjük jelölteknek, mert más hangoknak valószínűleg olyan nagy célköltségük lenne, hogy a szintetizált mondatba végül úgysem kerülnének be.

Az összefűzési függvény és az összefűzési költség. A szintézis során a célhoz kiválasztott jelöltek fognak egymás mellé kerülni. Azt, hogy az egyes jelöltek mennyire illeszkednek egymáshoz, az összefűzési függvény számítja ki és adja meg mint

összefűzési költség. Az összefűzési költség tehát két egymás után álló jelölt viszonyáról ad egy számértéket. Az összefűzési költség kiszámításakor a célköltségekhez hasonlóan felhasználhatóak az akusztikai és a nyelvi információk is. Az összefűzési költség számítása összetettebb feladat, mint a célköltségé, mert itt nem biztosított, hogy azonos típusú elemek között kell kiszámítani az illeszkedés mértékét. Például egy heterogén rendszerben egy szó után következhet egy hang méretű elem is. További különbség az is, hogy az elem típusától függően egyes tulajdonságok különböző értékűek lehetnek, attól függően, hogy az elem bal vagy jobb oldali csatlakozását vizsgáljuk. Például egy [p] hang bal oldalának spektrális tulajdonságai lényegesen eltérnek a jobb oldalának tulajdonságaitól. Az összefűzési költség az 10.3 egyenlet segítségével számítható.

$$C^c(u_k, u_{k+1}) = \sum_{r=1}^{R} w_r(C_r^c(u_k^l[r], u_{k+1}^r[r])), \tag{10.3}$$

ahol

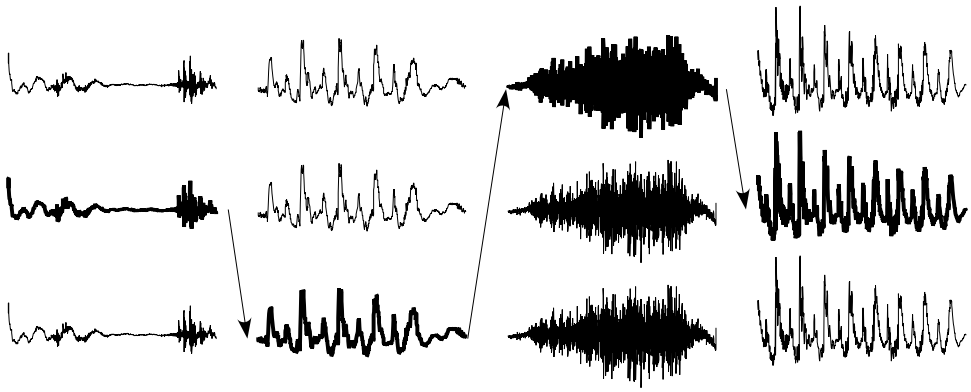
$u_k^{r,l}$ a k-adik adatbáziselem (jelölt), amely R jobb és bal oldali tulajdonsággal rendelkezik

$C^c(u_k^l, u_{k+1}^r)$ az összefűzés (concatenation) költsége a k-adik adatbáziselem (jelölt) és $k+1$-edik (jelölt) között,

$C_r^c(x,y)$ a távolságfüggvény az x és az y tulajdonság között.

Az összefűzési függvényben többnyire akusztikus tulajdonságok szerepelnek, mint például: alapfrekvencia, formánsfrekvenciák, spektrális tulajdonságok, intenzitás. Az összefűzési költséget úgy határozzák meg, hogy a beszédadatbázisban egymást követő elemek összefűzési költsége a legkisebb – általában 0 – legyen, így elősegítve, hogy az adatbázisból minél hosszabb beszédrészletek legyenek egy darabban felhasználva.

A hullámformaelemek rugalmas kiválasztása, a keresés menete. A hullámformaelemek kiválasztásához tehát rendelkezésre áll minden célelemhez egy lista, amiben az adott célhoz szóba jöhető jelöltek vannak. A 10.42. ábrán láthatók egymás felett az adott célhoz tartozó jelöltek. Egy-egy oszlopban szereplő hullámformák egy-egy cél jelöltjeit képviselik. A jelöltekre – a már ismertetett módon – kiszámoltuk a célköltségüket. Az egymás melletti oszlopok jelöltjei között pedig ki tudjuk számítani az összefűzési költséget. A feladat az, hogy megkeressük azt a jelöltekből álló sorozatot, amely esetében a számított költségek összege a legkisebb. Ezt hívjuk a minimális költségű útnak, amelyet az ábrán a nyilakkal és a megvastagított hullámformával jelöltük. Matematika úton leírva, egy adott út költségét az 10.4 egyenlet adja. Minden elemnek venni kell a célköltségét, majd az egymás mögött található elemek összefűzési költségét. Az egyenlet tartalmazza még a szünetelem és a legelső elem, illetve a legutolsó és a szünetelem összefűzési költségét is.



10.42. ábra. Keresés a jelöltek között

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + C^c(S, u_1) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + C^c(u_n, S) \tag{10.4}$$

A legkisebb ilyen költségű útban szereplő elemek adják meg tehát azt a hullámforma sorozatot, amely a szintetizált beszédet fogja adni, tehát a 10.5 egyenlet megoldása szükséges a szintetizáláshoz.

$$U \;=\; argmin\Big\{ \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=1}^{n+1} C^c(u_{i-1}, u_i) \Big\} \tag{10.5}$$

A legkisebb költségű út meghatározása történhet a Viterbi-algoritmus segítségével. Az algoritmus analóg a felismerésnél használttal (lásd 9.5.2. fejezet). A Viterbi-algoritmus megadja a legkisebb költségű utat, de számításigénye nagy. A számításigény csökkenthető úgy is, hogy az utak számolásánál csak a legígéretesebb n utat tartja meg az algoritmus. Ilyen a Beam-search algoritmus.

Utófeldolgozás. A legkisebb költségű út meghatározásával tehát rendelkezésre áll egy olyan elemsorozat, amelynek a hullámforma-részeit egymás után helyezve megkapjuk a szintetizált beszédet. Optimális esetben az így összeálló hullámforma további feldolgozást már nem igényel. A gyakorlati feladatok esetében általában szükséges valamilyen kismértékű jelfeldolgozás. Az elemek összeillesztése – hangadatbázistól függően – lehet egyszerűbb vagy összetettebb megoldású. Ha az elemek azonos időpillanatú hullámformából kerültek kiválasztásra – hasonlóan a ProfiVox diádos adatbázisához (8.2.2. fejezet) –, akkor a hullámformák egyszerűen egymás után pakolhatók. Amennyiben ezek a feltételek nem állnak fent, akkor valamilyen idő- vagy frekvenciatartománybeli összelapolás és simítás szükséges. A költség minimalizálása során – durva esetben – komolyabb alapfrekvenciabeli eltérés is előfordulhat az elemek között, amely már hibás hangzásként érzékelhető a hallgató számára. Mivel

az alapfrekvencia gépi módosítása csak kis részre vonatkozik, a keletkező minőségromlás csak lokális. Ez a fajta kiegyenlítés okozta torzítás kevésbé zavaró, mint maga az F_0 ugrás, tehát érdemes utólagosan ezeket kompenzálni. A felhasznált korpusz- és elemtípus tulajdonságaitól függően, valamilyen mértékű intenzitáskiegyenlítésre szükség van, mert a keresés során nem garantált, hogy minden elem arról a helyről származik, ahol a célnak megfelelő intenzitásviszonyok vannak. Ilyen eltérés lehet például, hogy nem mondat végéről származik egy olyan szó, amelyik a mondat végére került. Fontos megjegyezni, hogy a mondat vége pozíció szerepelhet a célköltségben, így preferált a megfelelő mondat végi részlet kiválasztása, de a teljes költség minimalizálása miatt előfordulhat, hogy egy nem mondat végi elem sokkal jobban illeszkedik a végleges szintetizált mondatba. Az így keletkező intenzitáseltérést korrigálni kell.

Intenzitáskiegyenlítés virtuális szóintenzitással Az kis méretű – például szótag, diád, hang méretű – építőelemek kiegyenlítéséhez szükséges a virtuális szóintenzitást fogalma, amely megadja, hogy az a szó amelyben a hang szerepel, milyen intenzitású lenne, ha minden hangja átlagos intenzitású lenne. Az átlagos intenzitás meghatározása több nagy méretű beszédadatbázison alapszik, ahol az adatbázisban előforduló össze hang átlagát számítjuk ki fonémánként. A virtuális intenzitás kiszámításnál a hangok időtartama szerinti súlyozott általagot számolunk a következő képlet szerint, ahol t_{ph} a hang időtartama, $I_{ph}^{avarage}(ph)$ az adatbázisokból meghatározott átlag és a $ph(i)$ a szó i.-dik hangja:

$$I_{word}^{virtual} = \frac{\sum_{i=1}^{N} t_{ph}(i) I_{ph}^{avarage}(ph(i))}{\sum_{i=1}^{N} t_{ph}(i)}. \tag{10.6}$$

Ezek alapján meghatározható az adott beszédhang módosításának a mértéke:

$$gain_{ph} = \frac{gain_{prosody} \cdot gain_{base} \cdot I_{word}^{virtual}}{I_{word}^{real}}. \tag{10.7}$$

A $gain_{base}$ adja az átlagos intenzitását a mondatnak, a $gain_{prosody}$ pedig a prozódia által meghatározott eltérést.

A korpuszadatbázis előállítása. Az adatbázis elemeinek meghatározásakor figyelembe kell venni, hogy milyen tematikájú szövegeket kell a rendszernek szintetizálnia (8.2.3 fejezet). A korpuszos szintetizátor a legjobb minőséget akkor tudja előállítani, ha a szintetizálandó mondat tartalma a korpusz tematikájába illik. Általános szövegek felolvasása esetén a korpusszal szemben az az elvárás, hogy fonetikailag gazdag adatbázis legyen, tehát a hang és hangkapcsolatok kellően széles választéka álljon rendelkezésre a keresés során mind szegmentális, mind szupraszegmentális szinten.

A hangadatbázisok előállítását részletesen a 8.2.3. fejezet tárgyalja. Korpuszok készültek magyar nyelvre időjárás-előrejelzések felolvasásához (Fék et al. 2006), mobilkészülékek árlistájának szóbeli megadásához (Olaszy et al. 2009), vasútállomási utastájékoztatáshoz és általános célú szövegfelolvasáshoz (Zainkó 2008). Ez a fajta szintézistechnológia több magyar alkalmazásban is felhasználásra került, amelyekről a következő fejezetek szólnak: időjárás előrejelzése (12.4.1. fejezet), árlistafelolvasás (12.3.7. fejezet), vasúti tájékoztató rendszer (12.5.1. fejezet).

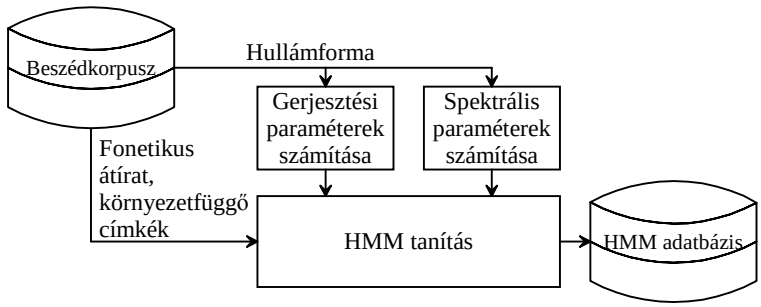
10.3.8. A rejtett Markov-modellen alapuló gépi szövegfelolvasás

Tóth Bálint–Németh Géza

A gépi tanulás lényege, hogy a jellemző információkat (paraméterek formájában) kinyerjük a beszédből, majd a szintéziskor a megszerzett tudás alapján létrehozzuk a gépi beszédet. Az első fázis tehát a tanulás. Az itt ismertetett, gépi tanulásos beszédépítési eljárás rejtett Markov-modell (HMM) alapú megoldás. Nem közvetlenül a hullámformával dolgozik, hanem a hullámformából spektrális és prozódiai jellemzők sokaságát nyeri ki. Tulajdonképpen szétszereljük a beszédjelet apró részletekre, majd ezeket vizsgáljuk, szortírozzuk, rendszerezzük. Mi kell a sikeres tanuláshoz? Három dolog. Mindenekelőtt sok idő, sok beszéd, valamint pontos információk a beszéd szerkezeti és nyelvi tartalmáról. Az, hogy a tanulásra sok idő kell, nem zavaró, mert ezt a háttérben elvégezhetjük, úgynevezett offline módon. A sok beszéd igényét nagy beszédadatbázisokkal elégítik ki, sok órányi beszédet gyűjtenek össze, néhány órától a néhány száz óráig. Ezt a továbbiakban tanító-beszédadatbázisnak fogjuk nevezi. Ezekbe a beszédadatbázisokba kell elhelyezni a beszéd szerkezeti és nyelvi részleteit úgynevezett jelzésekkel, valamint hangjelekkel. A jelzések a beszéd szerkezeti részére vonatkoznak (hanghatárok stb.), a hangjelek a nyelvi tartalmat rögzítik az elhangzás szintjén. Mindezekhez még hozzájön egy további nyelvi elem, a szöveg fonemikus alakja. Ebben benne van a szavak sorozata, a mondatok hossza stb. Ezek az adathalmazok képezik a tanulás alapját. A tanulási folyamat számos részegységből áll, melyek egymás után következnek. A tanítás végén egy viszonylag kis méretű adathalmaz áll elő (2–10 MByte), ezt nevezzük HMM szintézis adatbázisnak. A szintézis fázisban a felolvasandó szövegnek megfelelő paramétereket válogatjuk ki a HMM szintézis adatbázisból, és beszédkódolási eljárással a paramétereket hullámformává alakítjuk. A HMM-alapú beszédszintetizátort csak egyszer kell betanítani, az így előállított HMM szintézis-adatbázist lehet ezek után a szintézishez használni. A HMM-alapú beszédszintézis alkalmas arra, hogy egy általános HMM szintézis-adatbázis használatával több hangkaraktert is felépítsünk. Tehát személyre szabott hang is előállítható.

Tanítás. A tanítás célja, hogy az egyes hangokhoz és környezetükhöz rendelt paraméterek és az adott nyelvi sajátosságok segítségével előállítsunk egy azokat minél pontosabban becslő függvényt (például Gauss-eloszlást). A tanításhoz használt beszédkorpusz tartalma a következő: a hullámforma digitalizált változata, a hang- és szóhatárok pontos pozíciójának és magának a beszédhangnak a megjelölésével, valamint a felolvasott szöveg pontos fonemikus átirata. A hullámforma és a fonemikus szöveg között a jelzések teremtenek kapcsolatot, a két szintnek pontos szinkronban kell lenni. A feldolgozás egysége a mondat. A hullámformát 25 ms-onként ablakoljuk. Egy ablakon belül a hullámforma sok mintából áll össze, ezek száma a mintavételi frekvenciától függ. A mintákból kinyerjük az adott ablakra jellemző paramétereket. Egy-egy HMM tanításához tehát a beszédkorpuszból származtatott paraméterek sokaságára van szükség. Ezek a következők: a hullámforma spektrális tartalmára utaló, úgynevezett MFCC (Mel Frequency Cepstrum Coefficients) adatok, ezek első és második deriváltjai, továbbá a gerjesztési paraméterek (az alapfrekvencia, F_0, zöngés/zöngétlen osztályozó), valamint annak első és második deriváltjai. A deriváltak segítségével lehetséges a szintetizált beszédben a természetes hangzást növelni. Ezen túl még szükség van a beszédkorpusz fonetikus átiratából képzett hangkörnyezetfüggő címkékre. A környezetfüggő címkék írják le egy adott beszédhang környezetét, helyzetét, állapotát (például, hogy előtte és utána milyen hangok találhatók, az adott szótag hangsúlyos/hangsúlytalan, hány szótagú szóban szerepel, a mondat mely részén található stb.). A környezetfüggő címkékről később részletesen szólunk. A fenti paramétereket jelfeldolgozási és matematikai szoftverekkel automatikusan lehet előállítani. A tanító beszédadatbázis címkézési pontosságát célszerű kézi módszerekkel is ellenőrizni (Olaszy–Bartalis 2008). Mindezekből következik, hogy a tanítás előkészítő fázisa meglehetősen bonyolult és időigényes. A tanításból kapott HMM szintézis-adatbázis mérete (kb. 2–10 MByte) nagyságrendekkel kisebb az eredeti beszédkorpusz méreténél. A tanításhoz szükséges beszédkorpusz elkészítési részleteit a 8.2.3. fejezetben ismertettük.

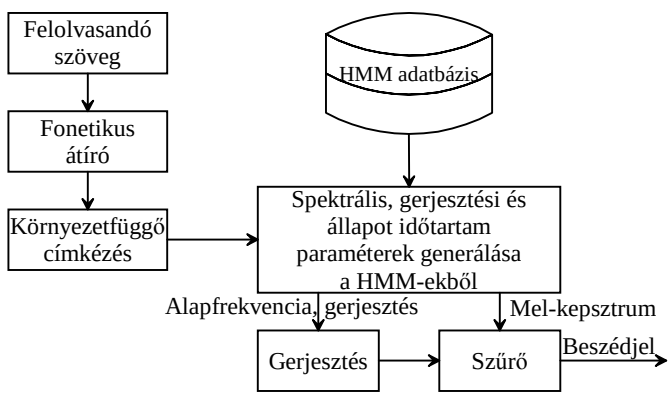
A betanításkor kétfajta megoldást alkalmazhatunk. Betaníthatjuk a HMM-eket egy adott beszélőtől származó 2–4 órás beszédadatbázissal. Ebben az esetben az adott személy hangján fog megszólalni a rendszer. Betaníthatjuk több beszélőtől gyűjtött adatbázissal is (beszélőnként 1–1.5 óra hanganyag, minimum 5–6 különböző személy hangja), melyet tetszőleges beszélő hangkarakteréhez tudunk szabni a későbbiekben rövidebb adaptációs adatbázisokkal (5–10 perc beszéd) (Masuko et al. 1997, Tamura et al. 2001). Ennél a tanítási formánál a több beszélő adatbázisaiból a nyelvre jellemző általános paramétereket megtanulja a rendszer, majd az 5–10 perces adaptációs hanganyagon végzett tanítás után az általános jellemzőket a célszemély hangja felé tolja el, tehát a célszemély hangjához közeli hangszínezettel, stílussal fog megszólalni a HMM alapú beszédszintetizátor. Ezen túl még számos módszer létezik a beszédhang jellemzőinek a megváltoztatására (Yoshimura et al. 1997, Tachibana et al. 2005). A HMM beszédszintetizátor által használt tanítórend-



10.43. ábra. A HMM alapú szövegfelolvasó tanítási folyamata

szer felépítését a 10.43. ábra mutatja. A betanítási folyamat sok időt igényel, akár napokig is fut a program a számítógépen egyetlen tanításkor.

Beszéd-előállítás a HMM-szintézisadatbázisból. A beszédszintézis során már csak a tanítás eredményét, a HMM szintézis adatbázist használjuk. Első lépésként elkészítjük a szintetizálandó szöveg fonetikai átiratát és a környezetfüggő címkéket, majd a várható hangidőtartamokat nyerjük ki az állapot-időtartam valószínűség sűrűségfüggvényekből, ezután pedig a legvalószínűbb spektrális és gerjesztési paramétereket nyerjük ki a HMM adatbázisból (a spektrális jellemzőket, az időtartamokat, a szüneteket és az alapfrekvenciát). Ezekből állítjuk elő a szintetizált hullámformát egy alkalmas beszédkódoló eljárással (Imai 1983, Maia et al. 2007). A HMM alapú beszédszintetizátor felépítését a 10.44. ábra mutatja.



10.44. ábra. A HMM alapú szövegfelolvasó beszéd-előállítási folyamata

A magyar HMM alapú beszédszintetizátor. A magyar nyelvű adaptáció során 14 magánhangzót és 25 mássalhangzót jelöltünk ki érvényes beszédhangoknak (az esetleges allofónokat nem jelöltük külön). A tanításhoz 6 órányi beszédet használtunk,

melynek a fele fonetikailag kiegyenlített mondatokból állt (Vicsi et al. 2005). A mondatokat 16 000 Hz-es mintavételezéssel, 16 bites kvantálással tároltuk. A mondatok fonetikus átiratát, automatikus célszoftverrel, kényszerített felismerést alkalmazva (Mihajlik et al. 2002) készítettük el. Ennek során az eljárás bejelölte a hang- és szóhatárokat is. Az adatbázis jelöléseit félautomatikus módszerrel finomítottuk (Olaszy–Bartalis 2008), hogy elérjük a 99%-os pontosságot.

Környezetfüggő címkézés. Annak érdekében, hogy a HMM-ek a legmegfelelőbb elemeket válasszák majd ki a beszéd-előállítás során, számos fonetikai jellemzőt kell megadni paraméterként. A jellemzőket minden egyes beszédhangra kiszámoljuk. Minden beszédhanghoz külön környezetfüggő címkehalmaz tartozik. Alapvetően 5-ös hangcsoportonként (kvinfon) vizsgáljuk az adott beszédhang környezetét (a hang és előtte kettő, utána kettő). A magyarra meghatározott legfontosabb címkéket a 10.17. táblázatban foglaljuk össze. A 10.17. táblázat alapján az 5 címkehalmazban

10.17. táblázat. A környezetfüggő címkézéshez használt jellemzők

Címkehalmaz	Címkék száma	Jellemzők
Beszédhangok	5	Az aktuális hang, valamint a megelőző és követő két-két hang (kvinfón). A szüneteket is hangként jelöljük.
Szótag	6321	Szótaghangsúlyok jelölése (hangsúlyos / hangsúlytalan) az aktuális, az előző és a követő szótagban. A beszédhangok száma az aktuális / az előző / a következő szótagban. A szótagok száma az előző / következő hangsúlyos szótagtól / szótagig. A szótag magánhangzója.
Szó	32	Szótagok száma az aktuális / előző / következő szóban. Az aktuális szó pozíciója a mondatrészben (elölről és hátulról is számítva). Mondatrésznek tekintünk két központozási jel közötti tagmondatot.
Tagmondat (két írásjel közötti szakasz)	62	A szótagok és szavak száma az aktuális / előző / következő mondatrészben. Az aktuális tagmondat pozíciója a mondatban (elölről és hátulról is számítva).
Mondat	111	A szótagok száma az adott mondatban. A szavak száma az adott mondatban. A tagmondatok száma az adott mondatban.

összesen 33 címkét kell figyelembe venni hangonként. A hangsúly előrejelzésére a ProfiVox beszédszintetizáló rendszer hangsúlymeghatározó algoritmusát használtuk (Olaszy et al. 2000b). Ez a címkéző mintegy 87%-os pontosságú, statisztikai elvű modell. A nyelvészeti alapú modellezési kísérletek sem adnak egyelőre jobb eredményt (Tamm–Olaszy 2005). A szótagokat a szótagmagok alapján keressük, számoljuk és jelöljük, tehát nem a nyelvi szótagolási szabályokat vesszük figyelembe. A szótaghatárok meghatározására alapul vettük a „Magyar helyesírás szabályai” (MTA 1985) című kiadvány „A szótagolás szerinti elválasztás” szabályait, azokat algoritmizáltuk. A 10.17. táblázatban meghatározott jellemzőket is címkéknek nevezzük. Tehát a beszédadatbázisban található összes mondat összes hangjához minden akusztikai és környezetet leíró paramétert ki kell számolni. Egy hanghoz összesen 38 környezetfüggő címkét rendelünk, így például egy 100 hangból álló mondathoz 3800 címkére van szükségünk.

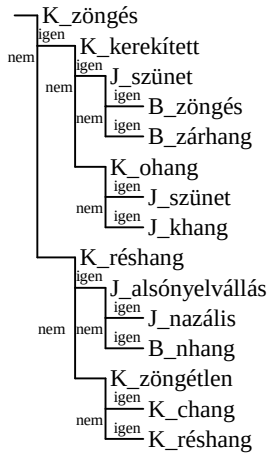
Döntési fák alkalmazása. Láthattuk, hogy számos környezetfüggő tulajdonság létezik, melyek összes lehetséges kombinációja óriási szám. Ha csupán a kvinfónok lehetséges változatait számoljuk meg, az is több mint 160 millió, de ezt a számot a többi környezetfüggő tulajdonság még exponenciálisan növeli. Ezért lehetetlen egy olyan, adott nyelvre jellemző, a tanításhoz szükséges beszédadatbázis-szöveget előállítani, melyben minden lehetséges kombináció szerepel. Tehát, ha a tanító adatbázisban nem szerepel az adott hanghoz kapcsolódó teljes környezet függő leírás, akkor hasonlót kell választanunk. E probléma leküzdése érdekében kellett bevezetni a döntési fa alapú klaszterezést, hasonlóan a beszédfelismerésnél alkalmazott eljáráshoz (Odell 1995, Shinoda–Watanabe 2000). Ez azt szolgálja, hogy a beszédszintézis során a HMM szintézis-adatbázisban az adott környezetfüggő címkékhez legjobban hasonlító elemeket válasszuk ki. Mivel a különböző környezetfüggő címkejellemzők hatnak mind a spektrális, mind az alapfrekvencia-paraméterekre és az állapot-időtartamokra is, ezért ezeket egymástól független három csoportra bontottuk. A döntési fákat a tanítás során számoljuk ki és a HMM szintézis-adatbázisban tároljuk. A döntési fák építésekor felhasznált jellemzőket a 10.18. táblázatban mutatjuk be. Egy példát láthatunk a 10.45. ábrán egy adott beszédhanghoz tartozó általános spektrális paraméterekre vonatkozó döntési fára. Az ábrán látható „K” előtag azt je-

10.18. táblázat. A döntési fák építéséhez használt jellemzők

Beszédhangok	Magánhangzó / mássalhangzó. Zöngés / zöngétlen. Rövid / hosszú. Képzés helye a magánhangzóknál (hátul, középen, elöl). Nyelvállás a magánhangzókra (felső, középső, alsó). Ajakállás a magánhangzókra (kerekített, nem kerekített). Képzés módja mássalhangzóknál (zárhang / réshang / zár-rés hang / pergőhang / nazálisok).
Szótag	Hangsúlyos / hangsúlytalan. Az adott szótagra vonatkozó számszerű adatok (lásd 10.17. táblázat).
Szó	Az adott szóra vonatkozó számszerű adatok (lásd 10.17. táblázat).
Mondatrész	Az adott mondatrészre vonatkozó számszerű adatok (lásd 10.17. táblázat).
Mondat	Az adott mondatra vonatkozó számszerű adatok (lásd 10.17. táblázat).

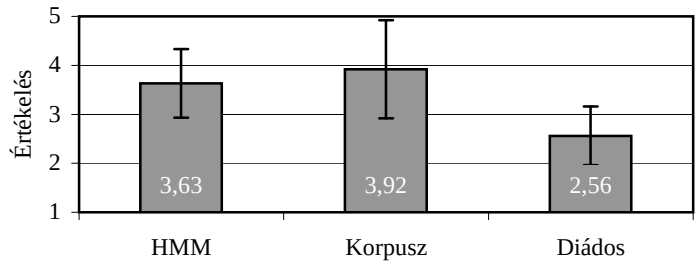
lenti, hogy épp a középső hangot vizsgáljuk a kvinfónból, a „J” előtag a középső hangot követőre (jobbra lévőre), a „B” előtag pedig a középső hangot megelőzőre (tőle balra lévőre) vonatkozik. A bemutatott példában azt láthatjuk, hogy a középső hang zöngés-zöngétlen csoportokra bontása volt a legelőnyösebb (a „K_zöngés” tulajdonság került a döntési fa legfelsőbb szintjére). Ezután a fa következő szintjén lévő „K_kerekített” (középső hang labiális) és „K_réshang” (középső hang réshang) alapján való csoportokra bontás volt a legjobb. A fa további szintjeit ezzel az elvvel lehet tovább bontani. Ezután a fa következő szintjén lévő „K_kerekített” (középső hang labiális) és „K_réshang” (középső hang réshang) alapján való csoportokra bontás volt a legjobb. A fa további szintjeit ezzel az elvvel lehet tovább bontani.

A HMM beszéd-előállítási módszer percepciós vizsgálata. A magyar nyelvű HMM alapú beszédszintézis minőségének értékelésére egy MOS (Mean Opinion Score) meghallgatásos tesztet végeztünk. Három beszédszintetizáló rendszert hasonlítot-



10.45. ábra. Egy döntési fa szerkezete

tunk össze, egy diád-triádadatbázisból építkező hullámforma-összefűzéses rendszert, egy korpuszos technológián alapulót és a HMM módszerű szövegfelolvasót. Ez utóbbinál impulzus-zaj gerjesztésű LSP (Linear Spectral Pairs) kódoló állította elő a hullámformát a megadott paramétersorozatból. A teszt elején mindegyik rendszer hangjával 3-3 mondatot játszottunk le véletlenszerűen, amelyeket a tesztalanyoknak még nem kellett értékelni. Ez azt a célt szolgálta, hogy az alanyok hozzászokjanak a mesterséges hangokhoz, és hallják előre, hogy nagyjából milyen minőségre számíthatnak. Ezután mindegyik rendszer mintáiból 29 mondatot játszottunk le, minden tesztalany esetén más-más sorrendben, így zárva ki az esetleges memóriahatásokat (van Santen 1993). A tesztmondatok tartalma időjárás-jelentés volt. A teszt körülményeihez az is hozzátartozik, hogy a diád-triádalapú rendszer kötetlen témakör szintézisére készült, míg a másik kettő csak kötött témakörre, az időjárás-jelentés felolvasására. Mindhárom szintetizátorral ugyanazt a 29 mondatot generáltuk (ezek a mondatok nem szerepeltek a tanító-adatbázisokban). A tesztalanyok a mondatokat 1–5-ös skálán értékelhették (egy volt a legrosszabb, öt a legjobb). A meghallgatásos tesztet 12 személy végezte el. Az eredményt a 10.46. ábra mutatja. Néhány HMM alapú szintézissel készített mondatok meghallgathatók a következő internetcímen: `http://speechlab.tmit.bme.hu/hmm/`



10.46. ábra. A meghallgatásos teszt eredménye a MOS átlag- és szórásértékekkel

10.3.9. Érzelmes szövegfelolvasás

Zainkó Csaba

Az érzelmi töltetű beszéd modellezése során ismertetett paraméterek (10.3.1.7. fejezet) két nagy csoportba sorolhatók be, az egyik részük a prozódiát érinti, a másik – a spektrális paramétereken keresztül – a legtöbb esetben a beszédadatbázist. Az érzelmi állapotot kifejező beszéd gépi előállításához a szövegfelolvasó működési elve korlátot szabhat (a paraméterek csak korlátozottan módosíthatók). A következőkben az egyes beszédelőállítási módszerekhez alkalmazható technológiákat ismertetjük.

Formánsszintézis. A formánsszintézis lehetőséget biztosít a legtöbb paraméter számszerű módosítására, így az érzelemi paraméterek elvileg jól átültethetők az ilyen típusú rendszerekbe. A problémát az jelentheti, hogy a formánsszintézis alapját képező műszaki modell nem elég finom a természetes hangzás előállításához, így az érzelmi jelleg elsikkadhat. A beszéd ilyen jellegű hangzásának módosítására (Olaszy–Kiss 1982) kísérleteiben találunk példákat.

Diád-, triádhullámformák összefűzésén alapuló technika. A hullámforma-összefűzéses szintetizátorok esetében az érzelmes beszéd előállításához magának a diád-triádelembázisnak az akusztikai módosítása szükséges, azaz minden diád- és triádelemnek az adott érzelmet kell tartalmaznia hangzásban. A diád-tirádelembázis előállításához kialakított szövegfelolvasási eljárás nehezen alkalmazható érzelmi töltetű elembázis készítésére (Zainkó et al. 2008), ezért célszerűbb egy már meglévő elembázist módosítani. Ebből következik az, hogy ahány érzelem előállítása szükséges, annyi számú módosított elembázist kell készíteni. Az eredeti elembázison csak a spektrális változtatásokat kell elvégezni, mert prozódiát érintő változtatások a szövegfelolvasó prozódiai moduljának finomhangolásával valósíthatók meg. A módosításokhoz használható például a Přibilová–Přibil (2009) munkáján alapuló algoritmus, amelyet Zainkó et al. (2010) dolgozott át hullámforma-összefűzésés szövegfelolvasókhoz.

Elemkiválasztás-alapú szövegfelolvasó. A korpuszalapú szintetizátor beszédadatbázisa több 10 órányi hanganyagot tartalmaz. Ennek a beszédadatbázisnak érzelmi töltetű felvétele olvasott beszéd esetén nem lehetséges, a bemondók nem képesek ilyen hosszan egyenletesen – például undorral beszélni. Megoldás lehet, hogy a beszélőktől folyamatosan gyűjtünk felvételeket (Iida et al. 2003). Ez alapja lehet a jó minőségű érzelmi beszédszintézisnek. Az ilyen eljárás lassú és erőforrásigényes, továbbá nagyméretű beszédkorpusz tárolása szükséges a szintetizáláshoz is. Másik megoldás a semleges érzelmű kész szintetizált beszéd utólagos módosítása. Mivel a semleges beszéd jó minőségű, ezért az érzelemmódosítás okozta jelfeldolgozás nem okoz számottevő romlást. Egy megoldásban (Zainkó et al. 2010) a PSOLA algoritmust kombinálja Přibilová–Přibil (2009) módszerével. A műveletsor a következő. Az időtartománybeli jelet aszimmetrikus Hann-ablak segítségével pitch-szinkron ablakozza, majd az ablakozott jelet DFT transzformációval átalakítja frekvenciatartományba. A frekvenciatartományban elvégzi az érzelmi tartalom kifejezéséhez szükséges spektrális transzformációt. A transzformáció az ablakozott jelben torzítást okoz, mivel az ablakok szélén a jel nem tart a nullához, így ismételt ablakozással kell korrigálni a torzítást. Az így módosított spektrumot inverz DFT-vel visszaalakítja időtartománybeli jellé, majd a PSOLA algoritmushoz hasonlóan átlapolva összeadja. Az érzelmi módosításhoz szükséges továbbá az alapfrekvenciamenet módosítása is. Ez az átlapolások eltolásával – ahogy a PSOLA esetén – megvalósítható. A zöngétlen szakaszok esetében két megoldás alkalmazható. Az egyik, hogy zöngétlen hangok esetén a transzformációt nem végezzük el, azaz a feldolgozás nem teljeskörű a beszédjelen. Ennek hiánya a percepcióban nem érzékelhető, mivel a zöngétlen hangok energiája kicsi. Másik módszer, hogy a zöngétlen részekre egyenlő távolságra virtuális zöngehatár-jelöléseket helyezünk el, és azokat a zöngés részekkel azonos módon kezeljük.

HMM alapú szövegfelolvasó. A HMM elvű beszédszintetizátor működése gépi tanuláson alapszik, a különböző paraméterek közvetlen manipulálására nincs lehetőség. A HMM szintetizátor kimeneti jelét módosító eljárás működőképes, de az ismételt jelfeldolgozás és a feldolgozási idő növekedése miatt a módszer nem ad jó eredményt. A HMM szintetizátor tanító-adatbázisai több tíz óra méretűek, és a tanítási fázis is időigényes (több hét nagyságrendű). Ezért csak a meglévő tanítás adaptációját érdemes elvégezni. A megoldás, hogy egy kisméretű – kb. 10 perc összes időtartamú – prozódiailag változatos mondatokból álló eredeti beszédadatbázist módosítunk a kívánt érzelem irányába, majd ezt alkalmazzuk az adaptáció során (Zainkó et al. 2010).

10.4. Beszédszintetizátorok minősítése, szabványosítási javaslatok

Olaszy Gábor

A szövegfelolvasók beszédének minősítésére ma még nincs egységesített, szabványos eljárás. Ugyanakkor a kutatóhelyeken számos példáját látjuk annak, hogy percepciós tesztekkel értékelik a rendszereket, mind magyar, mind más nyelvek esetében. Tehát a kutatási szféra fordít energiát a minősítésre. Az első ilyen témájú publikációk a 20. század utolsó évtizedében jelentek meg (van Bezooijen–Pols 1990, van Santen 1993), hiszen ekkorra már számos nyelvre létezett szövegfelolvasó technológia. A munkában a szerzők áttekintették a lehetséges módszereket és értékelték azok előnyeit és hátrányait. Magyarra az első ilyen publikált értékelést az MTA Nyelvtudományi Intézetében végezték (Gósy–Olaszy 1991). A beszédszintetizátorok értékelésénél általában két paramétert vesznek figyelembe, a hangminőséget és az érthetőséget. Mindkettőt meghallgatásos teszttel mérik az úgynevezett Mean Opinion Score (MOS) értékkel. Ennek lényege, hogy egy előre megadott minőségi skálán kell a besorolást elvégeznie a hallgatónak. Az érthetőségre vonatkozó ilyen skála lehet: nem érthető (1), rosszul érthető (2), nehezen érthető (3), érthető (4), jól érthető (5).

Kutatási-alkalmazási paradoxon. A kutatás és alkalmazás minőségének paradoxona azt próbálja kifejezni, hogy nagy űr tátong a kutatási eredmények és az alkalmazások színvonala között. A kutatási eredmények publicitása (bemutatók, riportok, előadások, újságcikkek) nem elégséges ahhoz, hogy azokat alkalmazzák is az egyébként arra befogadási feltételekkel rendelkező cégek, vállalatok, szervezetek. Két megoldás lehetséges. Vagy szabványokat kell alkotni, vagy speciális szemléletváltásra van szükség, mégpedig a piaci viszonyok között élő és dolgozó cégek vezetésében. A beszédtechnológiát alkalmazó cégek (bankok, mobilszolgáltatók, hangosinformáció-szolgáltatók, telefontársaságok stb.) jelenleg még nem fordítanak energiát arra, hogy tudatosan kialakítsák az „akusztikai arculatukat" is (11.4. fejezet), ami azt jelentené, hogy az általuk ajánlott beszédszolgáltatás beszédminőségét ugyanolyan magas szinten valósítják meg, mint az egyéb szolgáltatásaikat. Mivel nincs szabvány, ezt meg is tehetik. Az úgynevezett minőségbiztosítási rendszerek előírásai nem tartalmaznak meghatározást arra, hogy a beszédszolgáltatás hangminőségének milyen szintűnek kell lenni például egy banki számlainformációs rendszerben vagy egy mobilszolgáltató automatikus beszéddel válaszoló rendszerében. Konkrét példával élve, a számok kifogástalan hangminőségű felolvasására kidolgozott, tudományos igényességgel megvalósított rendszert már 1996-ban elkészült, ismertették, bemutatták. Mindezek ellenére egyetlen banki, tőzsdei, telefonszám-bemondó stb. rendszer üzemeltetőjénél sem merült fel az igény, hogy ezt alkalmazzák a rosszabb minőséggel beszélő rendszerek helyett. Mindez

annak is a következménye, hogy sok esetben olyan nemzetközi világcégek által fejlesztett, főleg angolra optimalizált rendszert alkalmaznak, amelyek szerkezete alkalmatlan a tudományos igényességgel megvalósított, jó hangminőségű beszélő rendszerek beépítésére. A rossz hangminőséggel beszélő beszédszintetizátorok közvetve hatással lehetnek a köznyelvi magyar beszéd romlására is, mivel pár év múlva, már tömegesen fogunk ilyen rendszereket hallgatni akaratunk ellenére. Ezért fontos a szabványosítás. A következőkben javaslatot teszünk egy általános minősítési eljárásra, amelyben különböző szempontok figyelembevételével lehet a szövegfelolvasók teljesítményét értékelni. Ez alapja lehet egy későbbi szabvány kidolgozásának is. A teszt egyes elemeinek részletes kidolgozásához célszerű nyelvészeti, beszédtechnológiai szakértőt bevonni. A minősítéshez szövegeket kell a szintetizátor bemenetére adni, és meghallgatással kell az elhangzottakat – egy megadott skálán belül – leosztályozni. A vizsgálat lépcsőfokai a következők:
Alapteszt. Ezt az akusztikai alapvizsgálatot célszerű minden új beszédszintetizátorra – függetlenül az alkalmazás területétől – elvégezni.

A beszédhangok akusztikai tartalmát vizsgáljuk, hogy a szintetizátor által előállított beszédhangok, hangkapcsolatok megfelelnek-e az adott nyelv hangjainak, hangkapcsolatainak (nem más nyelv hangjaiból fabrikálták-e, amelyre van példa a mobiltelefon-piacon).

A beszédhangok hangzóssági szintjeinél azt vizsgáljuk, hogy a szintetizátor által előállított beszédhangok, hangsorok intenzitása megfelel-e a nyelv hangjaira vonatkozó specifikus szinteknek (kiegyenlített-e a beszéd, nincsenek-e benne túl erősen, túl gyengén hangzó hangok, hangkapcsolatok, szavak).

A beszédhangok időtartamainál és a szüneteknél azt vizsgáljuk, hogy a szintetizátor által előállított beszédhangok időtartamai kiegyenlítettek-e, megfelelnek-e a nyelv hangjaira vonatkozó specifikus időtartam arányoknak (nincsenek-e a beszédben kirívóan hosszan ejtett, illetve túl rövid hangok, amikor a rendszer folyamatosan beszél). A szünet a beszédfolyamat szerves alkotórésze, fontos funkciója van a beszéd értelmezése, megértése szempontjából. A szünetek helyes megvalósítása a gépi szövegfelolvasás fontos eleme. A szünetek nagy részét az írásban is jelöljük a mondat végi írásjelekkel. Mondatok belsejében a vessző, kettőspont, pontosvessző, gondolatjel is jelenthet szünetet. Kötött szótáras szintetizátoroknál is lényeges a helyes szünettartás vizsgálata (nincsenek-e felesleges szünetek a beszédben).

Az akusztikai alapvizsgálati tesztnek az elvégzéséhez össze kell állítani olyan szövegeket, amelyekkel a fenti tesztek elvégezhetők. A szövegekből generált beszédet célszerű szakértővel meghallgattatni és értékeltetni (általános vélemény, hibalista, javaslat a hibák kijavítására, minőségbiztosítási jegyzőkönyv). Naiv hallgatósággal is el lehet végezni a tesztet, ekkor csak egy általános véleményt kapunk a szövegfelolvasó rendszer akusztikai minőségéről.
Nyelvi teszt. Ezt csak igényes, változatos nyelvi alkalmazásra szánt szövegfelolvasóknál kell elvégezni (e-levél-, regény-, fax-, név- és címfelolvasók). A teszt célja,

hogy megállapítsuk a szintetizátor kifejezési képességét (a tervezők milyen mélységig építettek be nyelvi szabályokat a rendszerbe). Itt a nyelv prozódiai sajátosságainak megvalósítását vizsgáljuk. A vizsgálati területek a következők: mondatfajták (kijelentés, kérdés, felszólítás, óhajtás, figyelmeztetés), mondathossz (meg van-e oldva a mondatfajták és a mondathossz összefüggése). Folyamatos dallamvonulat megvalósítása mondatról mondatra, szövegszinten (dialógusok regényrészletekben).

Ennek a tesztnek az elvégzéséhez össze kell állítani egy olyan szöveget, amelyik tartalmazza a magyar prozódiára vonatkozó legfontosabb mintákat. A felolvasott szöveget célszerű szakértővel meghallgattatni, aki értékelni tudja az egyes minták megvalósításának színvonalát (általános vélemény, hibalista, javaslat a hibák kijavítására). A naiv hallgatósággal végzett percepciós tesztet csak ezek után javasoljuk.

Funkcionális teszt. Ennek a tesztnek a célja az, hogy az alkalmazandó felhasználási területhez kapcsolódó speciális működéseket vizsgálja. Speciális feldolgozásra van szükség például elektronikus levelek felolvasásánál, ahol a bejövő karaktersorozatot erősen át kell alakítani, mielőtt a szintetizátor bemenetére küldjük. Ki kell szűrni a nem szöveg karaktereket (csillagok, kötőjelsorozatok, perjelek stb.), szét kell választani a szövegrészeket (cím, dátum, az üzenet szövege stb.), meg kell állapítani a szöveg nyelvét (az egész szöveg magyar-e, egyes szavak nem magyarok stb.).

Hasonló feldolgozásra van szükség az SMS-felolvasók tesztelésénél is, ahol például a felhasználók által használt (az SMS-kommunikációban kialakult) betűsorozatokat kell kimondásra alkalmas szöveggé konvertálni.

Hangoskönyv-szolgáltatásnál például azt kell vizsgálni, hogy a hosszú mondatokon belül a mondat tagolását milyen sikerrel végzi el a rendszer, alkalmaz-e hangsúlyozást a felolvasásnál (ha igen, akkor milyen eredménnyel, hányszor téveszt), továbbá, hogy a mondatok között milyen szüneteket tart. A szövegben elhelyezett XML jelzések feldolgozását is ellenőrizni kell (lehet, hogy a szintetizátor nincs felkészítve ilyen értelmezésre)

Képernyőolvasóknál (vakok és gyengén látók számára kialakított rendszerekben) azt kell vizsgálni, hogy például a betűzési funkcióban a szintetizátor hogyan mondja ki a betűket, továbbá, hogy a beszéd gyorsítása megoldott-e stb.

BESZÉDTECHNOLÓGIAI ALKALMAZÁSOK

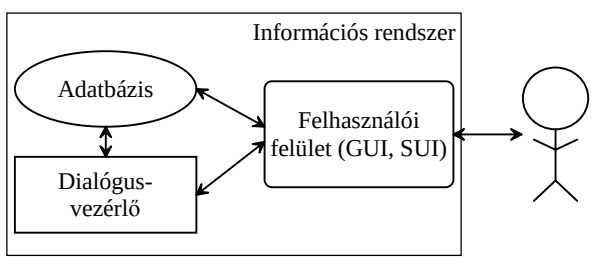
11. fejezet
Beszédinformációs rendszerek

Németh Géza

A beszédinformációs rendszerek a beszédtechnológia eredményeinek felhasználásával teszik lehetővé információ begyűjtését, tárolását, elérését és feldolgozását. Tervezésük, megvalósításuk újabb szintet ad az összetett információs rendszerek egyébként is sok tudást igénylő megvalósításához (Whitworth–Zaic 2003).

11.1. A beszédinformációs rendszerek fő építőelemei

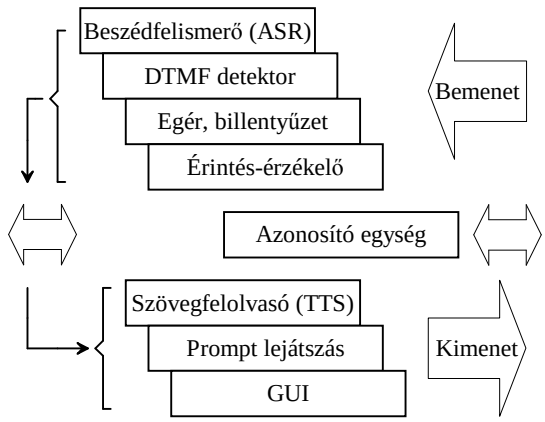
Ezekben a rendszerekben a beszédkommunikáció elemeit építik be, amivel a természetes emberi kommunikáció felé tolják el az információközvetítést, illetve a felhasználó igényének a megadását. Két nagy terület tartozik ide, a beszéd generálásával kapcsolatos megoldások (a felhasználó szóban kapja meg az üzenetet), illetve a beszédet fogadni képes beszédfelismerő motorok megoldási formái. A 11.1. ábrán látható egy általános beszédinformációs rendszer vázlatos felépítése. A felhasználó a



11.1. ábra. A beszédinformációs rendszerek általános felépítése

számára kialakított felületen keresztül ad és fogad információt. Ezek a felületek a következők: grafikus interfész (Graphical User Interface, GUI), illetve beszédinterfész (Speech User Interface, SUI, néha Voice User Interface, VUI elnevezést is használ-

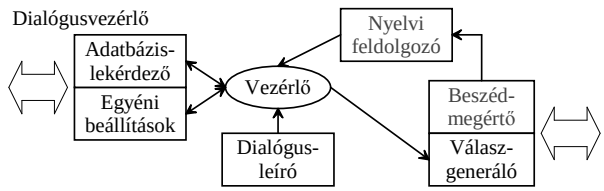
nak). Az információ tárolási helyét és forrását az egyszerűség kedvéért az adatbázis-blokkal jelöltük. Az információ leghatékonyabb kezelését és feldolgozását az úgynevezett dialógusvezérlő biztosítja. A dialógusvezérlő megvalósításának kulcsszerepe miatt a beszédinformációs rendszereket szokták beszéd-dialógusrendszereknek (spoken dialogue systems) is nevezni. A dialógusvezérlés tervezésének tudományát dialógusmérnökségnek (dialogue engineering) nevezik. A felhasználói felület leggyakrabban alkalmazott elemeit a 11.2. ábra mutatja. Amennyiben a felhasználói



11.2. ábra. A beszédinformációs rendszerek felhasználói felületének jellemző összetevői

felület csak az ember egyetlen kommunikációs csatornáját használja (unimodális), akkor a beszédinformációs rendszerben bemeneti csatornaként csak a beszédfelismerőt, kimeneti csatornaként pedig a szövegfelolvasót vagy hullámformavisszajátszást (prompt lejátszás) alkalmazzuk. Erre a megoldásra jellemzőek a telefonos információs rendszerek, például a tudakozók. Ha viszont több érzékszervünket is felhasználjuk az információs rendszerben, akkor mind a bemeneti, mind a kimeneti oldalon több eszközt is alkalmazhatunk a felhasználói felületben (multimodális). Ilyen alkalmazásokat futtathatnak például az információs kioszkok, melyek érzékelik a felhasználó közeledését, vagy azok a játékok, melyek gyorsulásérzékelő vagy vizuális gesztusfelismerés segítségével nyújtanak jelen mű írásakor még újszerűnek számító élményt (lásd a `http://www.xbox.com/en-US/community/events/e3/kinect.htm` bemutatót). Számos alkalmazásban (például bank) a felhasználói felület kritikus eleme lehet a felhasználó megbízható azonosítása. Ennek hagyományos módszere a nyomógombos telefonokon megadott számsorozat (PIN-kód), ami számos hátránnyal bír (könnyen lehallgatható, elfelejthető stb.). Ezért vonzó alternatívát jelenthet a felhasználó hangjának azonosítása, az általa bemondott kulcsszavak visszakérdezése vagy jól ismert személyes adatok (születési év, anyja neve stb.) sorozatának lekérése. A kizárólag beszédalapú (unimodális) dialógusvezérlő vázlatos felépítését a 11.3.

ábra mutatja be. A beszédfelismerőből érkező bemeneti adatokat egy beszédmegértő alrendszer értelmezi az adott feladatra vonatkozóan. A nyelvi feldolgozó pedig a feladat szemantikájának szintjén dolgozza fel a beérkezett információt. A vezérlő a



11.3. ábra. A beszédalapú dialógusvezérlő egy lehetséges vázlata

nyelvi feldolgozóból származó absztrakt jellemzők, a dialógusleíróból származó paraméterek és az adott felhasználó és feladat egyéni beállításai alapján valósítja meg az adatbázis lekérdezését. A lekérdezett adatokat pedig (egy többnyire természetes nyelvi) válaszgeneráló alrendszer hozza a szövegfelolvasó számára kezelhető formára. A továbbiakban az áttekinthetőség kedvéért kizárólag a beszédalapú unimodális rendszerek dialógusainak típusait tekintjük át.

11.2. Emberi-gépi dialógus

Az ember-gép közti dialógusra építő információs rendszereket többféle módon is osztályozhatjuk. Kézenfekvő a *dialógusvezérlésére használt technológia* szerinti osztályozás. A leggyakrabban alkalmazott megoldás a telefon nyomógombjaihoz rendelt hangjellel (Dual Tone Multi Frequency, DTMF vagy touch-tone) vár választ a felhasználótól a rendszer által feltett kérdésekre (ITU-T Q.23 1988). Mivel mindössze 12 szimbólumból választhatunk (0–9 számjegyek és *, #), a legtöbb gyakorlati esetben az információcserét nem lehet egyetlen szinten megoldani, hanem hierarchikus menürendszeren keresztül jutunk célunkhoz. Ez a folyamat önmagában is meglehetősen kényelmetlen lehet. Ha viszont hasonló funkciójú, de mások által üzemeltetett különböző rendszerekben (például különböző mobilszolgáltatók hangpostája, bankok telebankingrendszere stb.) más-más logikát alkalmaznak (például az egyikben a * az egy szinttel visszalépés jele, a másikban pedig a #), akkor a felhasználók az egész technológiát elvethetik. Erre a problémára válaszul már a 90-es években is kidolgoztak és szabványosítottak általános tervezési szempontokat (ITU-T F.902 1995).

A természetesebb dialógus kézenfekvő megvalósítási lehetősége a beszédfelismerők alkalmazása. Ennek előnye az, hogy a legtöbb ember számára legtermészetesebb módon, beszéddel vezérelhetjük a rendszert. Ez a megközelítés azonban jó néhány

csapdát is rejt. A gépi beszédfelismerés tulajdonképpen nem beszédfelismerés, csak mintaillesztés, és belátható ideig csak erősen korlátozott körülmények között éri el az emberi szintet. A kontextus megértése, értelmezése és a megfelelő válasz előállítása pedig tág témakörök esetén még hosszas kutatásokat, igényel. Ezért nagyon veszélyes, ha a gépi megoldás „úgy tesz" mint ha ember lenne a vonal végén. A jól tervezett rendszerek egyértelművé teszik, hogy gépi megoldás „veszi fel a kagylót". Kötött témakörök esetén (például a 12.3.5. fejezet) viszont már ma is hatékony megoldások léteznek nemcsak angol, hanem magyar nyelvre is. A beszédfelismerővel vezérelt dialógus egyik kritikus eleme a felismerési hibák kezelése. A DTMF vezérlés gyakorlatilag hibátlannak tekinthető működésével szemben még jó akusztikai és felhasználási körülmények között is jó eredménynek számít a 95%-os helyes felismerési arány (például csendes szoba, legalább három szótagos, egymástól legalább egy magánhangzóban eltérő szótárelemek stb.). Ez más megközelítésben azt jelenti, hogy átlagosan minden huszadik felismert adat hibás lehet. Ezért különösen fontos beszédfelismerési alkalmazásoknál a megfelelő hibakezelés, a felismert adatok megerősítésére jól megkülönböztethető kulcsszavak használata. Például az *Igen* és a *Nem* használatát nem javasoljuk, mert emberi bemondás esetén is könnyen összetéveszthetők. A *Helyes* és a *Hibás* pár sokkal jobban megkülönböztethető.

Elképzelhetők olyan alkalmazások is, amikor a dialógus vezérlésére az adott alkalmazáshoz tartozó, egyéb adatot is alkalmazhatunk. Ilyen megoldás lehet például az, hogy a mobiltelefonok aktív bázisállomásai, vagy GPS pozíciója alapján egy telefonszám felhívásakor automatikusan azonosítják a hívó fél földrajzi helyzetét és ez alapján felolvassa a rendszer az ahhoz tartozó időjárás-jelentést, megadja a legközelebbi pénzkiadó automatát stb.

A dialógusrendszerek másik osztályozási lehetősége a *vezérlés jellege* szerinti felosztás. Rendszervezérelt esetben mindig a rendszer tesz fel kérdést vagy ad utasítást a dialógus másik résztvevője számára, és annak az így felkínált lehetőségek közül kell választani. A felhasználóvezérelt megoldásban (többé kevésbé) szabadon nyilvánulhat meg az ügyfél. A rendszer nyílt kérdést tesz fel beszéddel (Miben segíthetek?) és a válasz értelmezése alapján lép tovább. Ez a megoldás sokkal rugalmasabb, viszont hiba esetén a felhasználó jobban eltévedhet. Ezért alakult ki a két megoldás ötvözését is lehetővé tevő vegyes kezdeményezésű megoldás. A következő szakaszban részletesebben és példákkal illusztrálva tekintjük át a fenti alternatívákat.

11.3. A dialógus tervezése

A DTMF-alapú, rendszervezérelt dialógusok a ma legelterjedtebben használt megoldások az úgynevezett interaktív hangválasz (Interactive Voice Response, IVR) rendszerekben. Hazánkban a legtöbb telefonos ügyfélszolgálat elsődleges elérési pontja

egy ilyen rendszer. Többnyire az egyik (sokszor mélyen eldugott) menüpont kiválasztása után lehet eljutni emberi kezelőhöz.

Az ilyen rendszerekben nem célszerű egy szinten mind a tizenkét (0–10, *, #) lehetséges választási lehetőséget felhasználni. Ideális esetben legfeljebb 4–5 számjegyhez érdemes az adott szintre vonatkozó opciókat kötni. Mivel a menürendszer elemeit a felhasználó egymás után hallja, ezért ezt is éppen elég megjegyezni. Ezenfelül javasolt az adott alkalmazás minden szintjén azonos funkciókat jelentő támogatásokat megvalósítani. Például a felolvasás megállítása (pillanat állj), vagy az egész menürendszer elejére ugrás tartozhat ebbe a kategóriába. A menürendszer mélységét sem érdemes 4–5 szintnél nagyobbra választani, mert ekkor megnő a rendszerben való eltévedés esélye. Célszerű különböző részletességű és bonyolultságú menürendszert kialakítani (például első jelentkezőknek, gyakorlott felhasználóknak, illetve szakértőknek). Ilyen elveket alkalmaztak például az 1999-ben elindított magyar nyelvű e-levél felolvasó szolgáltatás fejlesztése során (12.3.1. fejezet).

A menürendszer áttekinthetőségét nehezíti az a tendencia is, hogy a nagyvállalatok törekednek arra, hogy egyetlen telefonszámot kelljen reklámozniuk, ezért minden információforrást gyakran egyetlen mamutrendszerbe zsúfolnak. Az áttekinthetőséget segítheti az, ha az eltérő alrendszerekben más-más hangot (férfi, nő) használnak.

A dialógusrendszerek következő szintjét a beszéddel irányítható rendszervezérelt dialógusok jelentik. Az első ilyen megoldásokban angol nyelven a *Yes* és a *No* (*igen/nem*) kifejezésekkel lehetett navigálni egy fastruktúrába rendezett menürendszerben. Minden egyes elágazásnál a felhasználó azt az utasítást kapta, hogy válaszoljon az előre megadott szóval (*Ha már van időpont, akkor a sípoló hang után mondja azt, hogy igen. Egyébként mondja azt, hogy nem).* Természetesen az ügyfelek a napi kommunikációjukhoz hasonlóan mindenféle egyéb kifejezést is használtak (*rendben, mehet, lássuk csak, aha, persze*). Az elvileg két kulcsszót tartalmazó rendszer gyakorlati használhatóságához végül mintegy kétszáz kifejezést kellett beépíteni.

A beszédfelismerés valós körülmények közötti alkalmazásához ezért különösen kritikus a rendszerben alkalmazott bemondások (úgynevezett promptok) pontos és szabatos megfogalmazása. Úgy, hogy a rendszer minden potenciális felhasználója számára egyértelmű legyen, hogy mikor mit kell mondani. A technológia mai (2010) állása szerint az amerikai angol és számos európai nyelv mellett magyarul is elérhető olyan technológia, amivel kényelmesebben, bonyolult tartalmú információhalmazokat is kulcsszavak bemondásával elérhetünk (9.1. fejezet).

Ezen a pontos fontos megjegyezni, hogy a beszédfelismerők egyik lényeges paramétere, hogy képesek-e úgynevezett dinamikus nyelvtan kezelésére, vagy csak statikus nyelvtant alkalmazhatunk. Ez azt jelenti, hogy dinamikus nyelvtannal a felismerő rendszer szótára minden egyes menüpontban változtatható, tehát ha például 16 menüpont van a rendszerben, átlagosan 20-as választási lehetőséggel, akkor nem kell $16 \cdot 20 = 320$ szavas szótárból választani minden esetben, hanem minden egyes

ponton akár más-más 20-as szótár alkalmazható. Ez nagyban növeli a rendszer rugalmasságát és pontosságát. Statikus esetben az előbbi példánál maradva minden egyes menüponton ugyanabban a 320 elemű szótárban keresünk, noha sok esetben nincs is értelme az adott pontban egy-egy szótárelemnek.

A szótárak meghatározása a beszédfelismerőn alapuló rendszerek kritikus eleme. Magyar nyelven célszerű lehetőleg három magánhangzót (de legalább kettőt) tartalmazó parancsszavak, illetve kifejezések alkalmazása úgy, hogy lehetőleg a magánhangzók térjenek el egymástól. Ha ez nem lehetséges, akkor a magánhangzók környezetében levő mássalhangzók legyenek eltérő kategóriába sorolhatók. A dialógusalkalmazások fontos szempontja az is, hogy a beszédfelismerő csak egyetlen felismert szótári elemet ad-e ki vagy képes-e a legvalószínűbb néhány (tipikusan 2–5) szótárelemet megadni. Ez a dialógusvezérlő számára adhat segítséget az adott kontextusba illeszkedő elem kiválasztásában. Amennyiben a felismert elemekhez az azok valószínűségét jelző mérőszám (úgynevezett konfidenciaszint) is kinyerhető a felismerőből, akkor a rendszer rugalmassága növelhető. Például bizonyos konfidenciaszint alatt automatikusan emberi kezelőhöz kapcsol a rendszer. Ismereteink szerint magyar nyelven az egy menüpontban alkalmazott legnagyobb szótárméret elérte a 220 000-es értéket egy név szerinti tudakozó mintaalkalmazásban (például a *Nagy Péter, Cegléd* bemondásra a rendszer válasza gépi hangon a következő: *Nagy Péter, Cegléd, Kossuth utca 55. A telefonszám: 06 25 312 443*).

Amennyiben a dialógus közben több esetben is kis megbízhatóságú eredményt kapunk, vagy a felhasználók a megerősítés során ismételten többször jelzik, hogy hibás a felismerés, akkor a rendszer automatikusan visszaállhat DTMF vezérlésre (úgynevezett fall-back megoldás). Ilyenkor például a billentyűzettel kell az információt beadni (ami nyilván lassabb, de nem vész kárba a hívás).

A dialógusrendszerek esetén szükség lehet a pontos információátvitelre a gép és az ember között. Ilyen eset például az amikor egy személy- vagy cégnevet kell úgy kimondani, hogy az a másik oldalon betűtévesztés nélkül azonosítani lehessen. A 11.1. táblázat megadja az ajánlott betűzést. Amennyiben szükség van a kis- és nagybetűk megkülönböztetésére, akkor a betű elé a *kis* és *nagy* szavakat is mondani kell.

Az ügyfelek számára a legbarátságosabb megoldás a felhasználóvezérelt dialógus. Ekkor adott témakörben a rendszernek bármilyen, az adott nyelven és kontextusban értelmes üzenet (bemondás) megadható. Tipikusan ebbe a kategóriába tartoznak a menetrendi információs rendszerek, a jegyvásárlást támogató megoldások stb. A gyakorlatban használható megoldáshoz ilyenkor mély nyelvi elemzést és hibakezelést is alkalmazó dialógusvezérlés szükséges.

Mivel általában több adat megadása is szükséges a kívánt információ meghatározásához, különösen fontos a felismert lekérdezési paraméterek megerősítése (verifikációja). Ez történhet közvetlen (explicit) vagy közvetett (implicit) módon. A tervezési szempontból legegyszerűbb megoldás esetén minden egyes adatra visszakérdez

11.1. táblázat. Javaslat a magyar környezetben előforduló karakterek betűzésére

Jel	Betűzés	Jel	Betűzés	Jel	Betűzés	Jel	Betűzés
A	a mint Aladár	Á	á mint Ágnes		szóköz	ˆ	kalap jel
B	b mint Béla	Ä	umlautos nagy a	!	felkiáltójel	_	alulvonás jel
C	C mint Cecil	É	É mint Éva	"	idézőjel	{	nyitó kapcsos zárójel
D	D mint Dénes	Ë	umlautos nagy e	#	kettőskereszt	|	függőleges vonal
E	E mint Elemér	Í	hosszú í mint írisz	$	dollárjel	}	csukó kapcsos zárójel
F	F mint Ferenc	Ó	hosszú ó mint óra	%	százalékjel	˜	hullámvonal
G	G mint Géza	Ő	hosszú ő mint őz	&	és jel	€	euró jel
H	H mint Hugó	Ö	ö mint Ödön	’	aposztróf	‰	ezrelék jel
I	I mint Ilona	Ú	hosszú ú mint út	(	nyitózárójel	Š	kalapos nagy s
J	J mint János	Ű	hosszú ű mint űrhajó	)	csukózárójel	Ž	kalapos nagy z
K	K mint Károly	Ü	ü mint üveg	*	csillag	–	közepes kötőjel
L	L mint Lajos	0	nulla	+	plusz jel	—	hosszú kötőjel
M	M mint Mária	1	egyes	,	vessző	™	márkanév jel
N	N mint Nóra	2	kettes	:	kettőspont	¢	cent jel
O	O mint Ottó	3	hármas	;	pontosvessző	£	font jel
P	P mint Péter	4	négyes	<	kisebb jel	¥	jen jel
Q	q betű	5	ötös	=	egyenlőség jel	§	paragrafus jel
R	R mint Róbert	6	hatos	>	nagyobb jel	©	copyright jel
S	S mint Sándor	7	hetes	?	kérdőjel	±	plusz-mínusz jel
T	T mint Tamás	8	nyolcas	@	kukac jel	¾	három negyed
U	U mint Ubul	9	kilences	[	nyitó szögletes zárójel	½	egy ketted
V	V mint Viktor	-	kötőjel	\	fordított per jel		
W	dupla vé mint Walter	.	pont	]	csukó szögletes zárójel		
X	X mint Xavér	/	per jel				
Y	ipszilon						
Z	Z mint Zoltán						

a rendszer. Egy menetrend-lekérdezési példával illusztrálva (F: felhasználó, R: rendszer):

R: Üdvözlöm. Mi a célállomás?
H: Szegedre szeretnék utazni.
R: Ebedre szeretne utazni?
H: Nem.
R: Hova szeretne utazni?
H: Szegedre.
R: Szegedre szeretne utazni?
H: Igen.
...

A beszéd használata ellenére ez a fajta dialógus meglehetősen kényelmetlen és lassú a felhasználók számára. Szerencsésebb és gyorsabb, ha az explicit megerősítés mellett javítással a helyes válasz is megadható. A fenti példán bemutatva:

. . . .
R: Üdvözlöm. Mi a célállomás?
H: Szegedre szeretnék utazni.

R: Ebedre szeretne utazni?
H: Nem, Szegedre.
R: Szegedre szeretne utazni?
H: Oda.
...

Egy szinttel fejlettebb megoldás az, ha egyszerre több adatra kérdezünk rá, és egyben több adatelem is megadható:

R: Budapestről Szegedre szeretne utazni?
H: Nem, Velemre.
R: Budapestről Velemre szeretne utazni?
H: Igen, Budapestről Velemre.
...

Ebben a formában értelmezni kell azt, hogy például a nem és a javítás a kiindulási vagy a célállomásra vonatkozik, ezért a természetesebb megoldás mellett a dialógusvezérlő bonyolultsága nő. Az implicit megerősítés esetén az adott adatelem ellenőrzését a következő adat lekérdezésébe ágyazzuk be:

H: Szegedre szeretnék utazni.
R: Honnan szeretne Velemre utazni?
H: Nem, Szegedre.
R: Honnan szeretne Szegedre utazni?
H: Budapestről.
...

Célszerű a kritikus adatok többszörös megerősítését elvégezni. A példánál maradva a jegy foglalása előtt az időpont, a kiindulási és a célállomás, az ülés osztálya, a jegy típusa stb. ellenőrzésére akár egyetlen bemondás alapján is kérhetünk megerősítést. Fontos, hogy az elfogadási és az elutasítási parancsok jól megjegyezhetőek és megkülönbözhetőek legyenek.

A beszédinformációs rendszerek tervezésével kapcsolatban mélyebben érdeklődő olvasók figyelmébe Gardner-Bonneau–Blanchard (2008) szakkönyvét ajánljuk.

11.4. Az akusztikai arculat

Ebben a fejezetben a jól ismert „vállalati arculat” kifejezésnek az akusztikai jelenségekkel kapcsolatos kiterjesztését ismertetjük Németh (2006) alapján. Az „*akusztikai arculat*” (acoustic company image) fogalmát a szerző egy, a beszédtechnológiák bevezetésére vonatkozó 1998-as vizsgálata kapcsán (Németh 1998) alkotta meg. Az akkori (és sajnos nagyrészt a mai) helyzetet jól jellemzi egy bank informatikai vezetőjének válasza az egyik kérdőív kérdésére:

„Tudjuk, hogy a bemondásaink és a generált válaszaink minősége rossz, de ez nem igazán probléma. A fontos az, hogy rövid időn belül ki tudtunk fejleszteni egy rendszert, ami információt tud adni az adatbázisainkból."

Annak ellenére, hogy a vállalatok jelentős erőfeszítéseket fordítanak a minőség szempontjainak érvényesítésére (lásd total quality management, TQM), az akusztikai minőség fogalma a legtöbb ember fejében koncerttermekre, színházakra, minőségi elektroakusztikai berendezésekre és hasonlókra korlátozódik. Az angol nyelvű szakirodalomban újabban jelentős figyelmet kap a hangminőség (*sound quality*) témaköre (Lyon 2000)) olyan jelenségek tervezése kapcsán, mint a gépkocsik kipufogóhangja, az ajtó becsapódásának zöreje, a háztartási gépek (például porszívó, mosógép) működésével járó hangjelenségek stb.

A *„vállalati arculat"* már hosszabb ideje ismert és használt fogalom. Érdemes megkülönböztetni a vállalati azonosság(tudat) (company identity) és a vállalati arculat (company image) fogalmát (MaGee 2005). A megnevezések általánosított formája a szervezetekre általánosságban vonatkozik (corporate identity/image). Az előbbi *a vállalat „mindazon intézkedéseinek összesége, amely meghatározó a szervezet egészére nézve"* (Barát 2001). Más szóval azon „tulajdonságok, intézkedések összessége, melyek azonosítják a szervezetet önmagával, illetve megkülönböztetik más szervezetektől" (MPRSZ 2000). Az utóbbi pedig arra vonatkozik, hogy *a külső személyek és szervezetek milyennek látják az adott vállalatot* (gyakran beleértik azt is, hogy a vállalat milyennek akarja láttatni magát). Sajnos, a magyar szóhasználat ezekben a témakörökben sem egységes. Még a magyar PR Szövetség ajánlásával kiadott meghatározáslista is számos angol szót tartalmaz (MPRSZ 2000).

A vállalati arculat szubjektív tényezőktől függ. Ezért a cégek hangsúlyt fektetnek az egyszerű, könnyen értelmezhető jellemzőkre. Érthető, hogy elsősorban a vizuális jegyekre, különösen a logóra összpontosítanak, hiszen gyakran annak alapján azonosítják a vállalatot. Jellemző, hogy egy – a vállalati arculat tervezéséhez szempontokat adó – huszonkét jellemzőt tartalmazó paraméterlista (MaGee 2005) csak egyetlen, akusztikához kapcsolható tanácsot tartalmaz: „*Egységes, professzionális módon válaszoljuk meg a telefonhívásokat*." Mindez megfelelő lehet fizikai formában megjelenő, megnézhető és megtapintható termékek esetén, az infokommunikációs szolgáltatások piacán azonban sokkal szélesebb látókörű megközelítés szükséges.

11.4.1. Az akusztikai arculat áttekintése

Az akusztikai arculat definícióját az általános definícióból származtathatjuk:

Az akusztikai arculat azt fejezi ki, hogy külső személyek és szervezetek milyennek látják az adott vállalatot, valamint a termékeit és szolgáltatásait akusztikai jegyek, paraméterek és események alapján.

Az akusztikai arculat néhány eleme hosszabb ideje reflektorfényben van. Ilyenek elsősorban a televíziós reklámok. Ebben az esetben a vizuális és az akusztikai információ együtt jelenik meg. Az akusztikai paraméterek kevésbé lényegesek, mert a vizuális élmény döntő módon befolyásolja a nézőt (Illényi–Csányi 2001). A rádiós műsor és vállalati szignálok, valamint a reklámok azon akusztikai arculati elemek közé tartoznak, melyeket többnyire tudatosan terveznek meg a vállalatok általános arculatáért felelős szervezetek.

A beszédminőség (speech quality) a távközlő hálózatok alapvető jellemzője, és olyan paraméterekkel írjuk le, mint a jel/zaj viszony (SNR), torzítás, bithibaarány (BER) vagy szubjektív minőségi érték (Mean Opinion Score, MOS).

A hangminőség (voice quality) két lényegesen eltérő megközelítést takar. A legtöbb esetben hagyományos hangtechnikai-elektroakusztikai értékelést fejez ki, például hangszórók, fejhallgatók, erősítők, CD-játszók, termek stb. minősítésére szolgál. Az elmúlt években azonban előtérbe került a használata más, hangjelenségeket produkáló eszközök, például gépkocsik, mosó- ill. fúrógépek esetében is. Kiderült, hogy az olyan objektív paraméterek, mint például a hangosságszint, nem feltétlenül korrelálnak a felhasználók értékelésével. Előfordul, hogy egy hangosabb mosógép hangját kellemesebbnek tartják, mint egy csendesebb, de „furcsa” hangú versenytársét (Lyon 2000). Noha az ilyen jellemzők is befolyásolják a vállalati arculatot, de mégis erősen kötődnek egy bizonyos termékhez. Nehéz olyan fúrógépet elképzelni, ami működés közben az egységes vállalati dallamot játssza. Bizonyos termékcsoportok esetében például sportkocsik motor- és kipufogóhangja (az erő hangja) vagy asztali számítógépek ventilátorzaja (csendes segítőtárs), vagyis az adott termékek hangminősége a terméket gyártó egész vállalat arculatára is jelentős hatást gyakorolhat.

Az akusztikai arculat szerepe egyre növekszik és egyre kifinomultabb megközelítést igényel az alábbi okokból:

- erőteljes az a tendencia, hogy egyre több szolgáltatást és információt beszéddel nyújtsanak (főleg telefonkapcsolaton alapuló hívásközpontokban (call center), de az internetalapú megoldások száma is növekszik),
- hangsúlyt kap a célcsoportokra vagy akár személyekre szabott megközelítés (például léteznek már olyan vállalkozások az USA-ban, melyek arra szakosodtak, hogy egy adott vállalatra jellemző hangposta üdvözlő üzenetet, vagy várakozás alatti zenét terveznek és valósítanak meg).

Az akusztikai arculat tudatos formálásához fel kell mérni és kezelni kell azokat a tényezőket, melyek befolyásolhatják azt. Szerencsés, ha az akusztikai arculat tudatos kialakítását a meglevő szervezetbe és eljárásrendbe (például TQM) építik be. Fontos szempont, hogy a helyes megoldáshoz a marketingszakembereknek és az adott akusztikai esemény szakértőinek (akusztikus, fonetikus, gépészmérnök, beszédtechnológus stb.) szorosan együtt kell működniük.

Jellegzetes gond az, hogy a vállalati beszerzési eljárásrendben gyakran nem jelennek meg az akusztikai arculat szempontjai. Helyette elsősorban az ár és néhány műszaki jellemző dominál. Jó példa erre az, hogy a magyar távközlési szolgáltatók többsége olyan hangpostarendszert használ, ami a telefonszámokat úgy mondja be (az angolból átvett szoftver magyarítása következtében számjegyenként), ahogyan egyetlen magyar ember sem beszél. A rendszerek teljes bekerülési értéke több milliárd forint, működésük több millió ügyfelet érint, és a problémát legfeljebb egy emberhónapnyi programozói és beszédtechnológia szakértői munkával meg lehetne oldani. A jó hangminőségű üzenetet csak egyszer kell gondosan, igényesen előkészíteni – beszédtechnológiai szakember bevonásával –, és utána milliók fogják hallgatni nap mint nap. Érdemes lenne tehát az egyszeri befektetés.

Gyakran felmerül a felelősségi körök átfedése. Például hagyományosan többnyire a marketingosztályok a felelősek azért, hogy kiválasszák azt a színészt vagy színésznőt, akinek a hangját televízió- és rádióreklámokban a vállalat azonosítására használni fogják. Mikor azonban egy részben automatizált hívásközpont kialakítása merül fel, akkor elsősorban az ügyfélszolgálati osztály határozza meg, hogy kinek a hangjával veszik fel (a sokszor gyakran változó) rendszerüzeneteket. A két hang többnyire különbözik. A magasabb vezetési szintek bevonása nélkül a két megközelítés harmonizálására kis esély van.

11.4.2. Infokommunikációs szolgáltatások és az akusztikai arculat

A vállalatok akusztikai arculata talán az infokommunikációs szolgáltatások területén a legfontosabb, mert itt a felhasználói interakciók és élmények többsége akusztikai jelenségekhez kapcsolódik, és a bevételek többsége is ebből származik. A következőkben néhány – kimondottan erre a területre jellemző –, az akusztikai arculathoz kapcsolódó tényezőt veszünk számba, valamint kapcsolódó felhasználói teszteket is bemutatunk.

11.4.2.1. Az akusztikai arculat összetevői infokommunikációs szolgáltatásokban

Beszédszolgáltatások esetén a legkézenfekvőbb paraméter az átviteli minőség. Az ezekre vonatkozó jellemzőket szabványosítási folyamatok során részletesen szabályozzák. Az észlelt minőség azonban állandó műszaki minőség mellett is változhat. Jó példa erre az, hogy a mobil- és az internettelefónia széles körű használata során az ügyfelek gyakran találkoznak erősen torzított beszéddel és hozzászoknak ahhoz.

Emiatt a változatlan műszaki tartalom mellett is a hagyományos vezetékes szolgáltatás sok mobilfelhasználás után jobbnak tűnhet.

Egy másik meghatározó tényező az ügyfélszolgálatok alapvető eszközévé vált hívásközpontok szolgáltatásminősége. Két alaptípusukat különböztetjük meg. A kimenő (outbound) hívásközpontból felhívják az ügyfelet. Ilyen rendszereket elsősorban a marketingvállalkozások, közvéleménykutatók, biztosítók, utazásközvetítők stb. alkalmaznak. A legtöbb esetben jól képzett kezelők veszik át a szót az általában automatizált hívásfelépítési folyamat után. Egy jellemző kivétel az ún. SMS-felolvasás, amikor rövid szöveges üzenetet küldenek egy olyan vezetékes telefonszámra, melyhez nem kapcsolódik annak szöveges fogadására alkalmas telefon- vagy faxkészülék. Ekkor az üzenetet egy gépi szövegfelolvasó (Text-To-Speech, TTS) rendszer gépi beszéddel olvassa fel. Attól függően, hogy a TTS megoldást milyen környezetbe ágyazzuk, az észlelt szolgáltatásminőség és ezzel az akusztikai arculat jelentősen eltérővé válhat. A mai TTS rendszerek általában jól érthető hangot szolgáltatnak, azonban a beszédük kissé robotos. Azt is figyelembe kell venni, hogy a TTS szöveget (karaktersorozatot) olvas fel, tehát ha a szöveg hiányos (például hiányoznak ékezetek), akkor a felhangzó beszéd sem lesz helyes.

Ha először magát az üzenetet a tetszőleges szöveg felolvasására felkészített TTS rendszerrel olvastatjuk fel, majd a feladó telefonszámát és a feladási időpontot is szintén a TTS mondja be, a felhasználónak nem lesz lehetősége a gépi hanghoz történő alkalmazkodásra, a feladó száma alapján a várható tematikára való felkészülésre stb. A lényegi üzenet után korlátozott minőségben felhangzó kiegészítő információk feldolgozása nehezíti a fő mondanivaló megértését. Ha viszont előbb a telefonszámot és az időpontot mondatjuk be egy erre a célra fejlesztett, az emberi bemondó hangminőségét közelítő rendszerrel (Olaszy–Németh 1999), és csak ezután érkezik a fő üzenet a TTS segítségével, akkor az ügyfél jobb eséllyel értheti meg a felhangzó információkat és ily módon a vállalat akusztikai arculata is kedvezőbb lesz. Természetesen nagyon fontosak a további kiegészítő tulajdonságok is. Az éjjel egy órakor csörgő telefon, majd a felhangzó SMS-üzenet a legjobb felolvasási technológia mellett is leronthatja a vállalati arculatot.

A fogadó (inbound) hívásközpontokat az ügyfél hívja fel és előfordulhat, hogy azonnal jól képzett kezelőkhöz kapcsolják. Az ellenérzések zöme az automatizált interaktív hangválasz (Interactive Voice Response, IVR) megoldásokhoz kapcsolódik. Ha az ügyfelek költségmentesen vehetik igénybe jól képzett ügyfélszolgálati munkatársak idejét, kevesen kezdenek el összetett menürendszerekben barangolni és a szükséges kezelők száma gyorsan nőni kezd. Ennek ellensúlyozására a vállalatok gyakran arra kényszerítik ügyfeleiket, hogy olyan, viszonylag egyszerű feladatokat, mint előre fizetett kártya feltöltése, egyenleglekérdezés stb. automatizált rendszerekkel oldjanak meg. Ezeket a rendszereket gyakran olyan nagyvállalatok szállítják, melyek elsődleges piaca(i) angol (vagy legfeljebb 6–8 másik) nyelvet használ(nak). A kisebb piacokra készített nyelvi változatok gyakran rossz minőségűek (az elsőd-

leges nyelv szerkezetét követik) még olyan alapvető témakörök esetében is, mint a telefonszámok, dátumok és pénzösszegek felolvasása. A minőséget tovább rontják a határidős elvárások, mert az ilyen vizsgálatokat, illetve honosításokat többnyire a projektek végére hagyják. Gyakran nehezen megoldható, vagy teljesen elmarad a menürendszerek szerkezeti és hangminőségének, hatékonyságának mérése és optimalizálása.

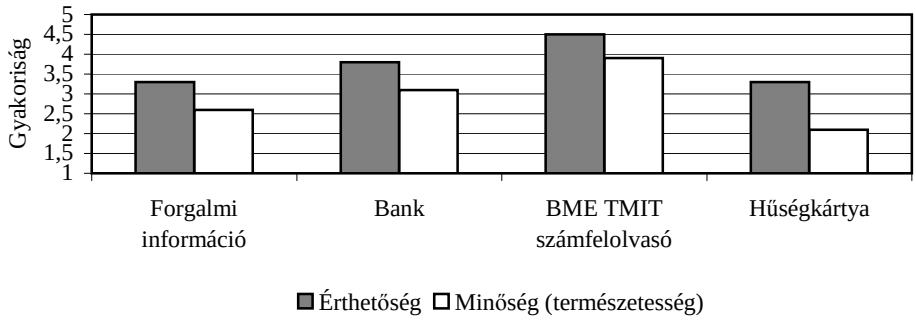
Ha egyszer egy rossz minőségű, de az elemi használhatóságot kielégítő rendszert üzembe állítanak, nehéz javítani rajta. A nagyvállalatok gyakran vizsgálják az ügyfélpanaszok számát és okát, de – még akkor is, ha elégedetlen a rendszerrel – nagyon ritka az, hogy az ügyfél panaszkodással töltse az idejét olyan esetekben, amikor a közvélekedés szerint a változás esélye csekély, nem remélhető, hogy az egyéni megjegyzésekre figyelni fognak. Az illetékes menedzserek esetleg elismerik, hogy a rendszer rossz minőségű, de azzal érvelnek, hogy a javítás profitnövelő hatását nem (vagy nehezen) lehet kimutatni, és a változtatást körülményes és költséges lenne végrehajtani. A helyzetet jól jellemzi egy ügyfélszolgálati vezető véleménye: „*Tudjuk, hogy az ügyfelek utálják az IVR-t. Mi is utáljuk. De nincs pénzünk több kezelőre.*"

Gyakran előforduló hiányosság annak figyelmen kívül hagyása, hogy a stúdiókörülmények között kiválóan hangzó felvétel (vállalati „akusztikus logó", zene, csengőhang, előre felvett üzenetek) nem garantálja azt, hogy a telefonos alkalmazás is sikeres lesz. Egy akusztikus jel 20-ról 3,1 kHz-re történő sávkorlátozása többnyire jelentős torzulásokat okoz (különösen a széles spektrumú – például sziszegő – beszédhangoknál és a tranziens jelenségeknél – például ütős hangszereknél). Ahhoz, hogy reális legyen a megítélés, minden minősítést a végfelhasználói csatornán kell elvégezni.

11.4.3. Az akusztikai arculatot meghatározó néhány szolgáltatás vizsgálata

Annak érdekében, hogy bemutathassuk az akusztikai arculat változását a különböző szolgáltatások tükrében, empirikus vizsgálatokat végeztünk beszédinformációs rendszerekben alkalmazott alapvető bemondástípusokra (Szőke 2003). A hangfelvételeket a BME TMIT-en valódi szolgáltatásokról készítettük. A felvételeket ötven ép hallású egyetemi hallgató értékelte ötfokozatú skálán (1 = legrosszabb, 5 = legjobb). A tesztalanyoknak külön-külön kellett értékelniük a minták funkcionalitását (érthetőségét) és minőségét (természetesség). Minden mintát kétszer játszottunk le. Először valamennyi mintát végighallgatták, majd az újbóli lejátszás során a minták között rövid szünetet tartottunk. Ezalatt kellett az előzőleg meghallgatott minta osztályozását elvégezni.

Az 11.4. ábrán számfelolvasási technológiák értékelésének átlagát láthatjuk. Noha az érthetőségi értékek változása is jelentős és csak egy megoldás haladta meg a 4-es

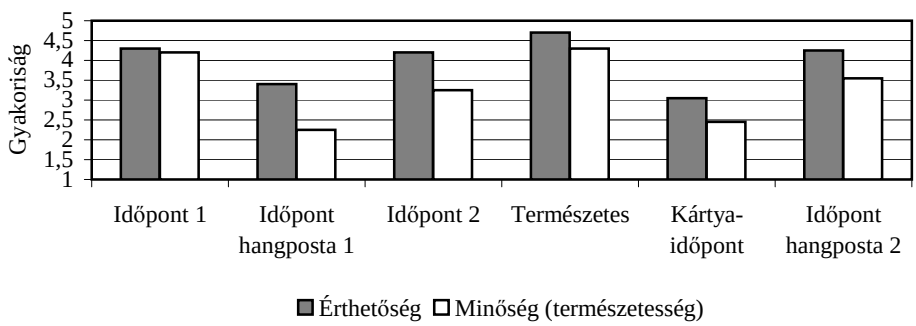


11.4. ábra. Különböző számfelolvasók érthetősége és minősége (Szőke 2003). *Forgalmi információ* = egy távközlési szolgáltató előfizetője beszélgetéseinek havi díja, *Bank* = számlaegyenleg értéke, *BME TMIT számfelolvasó* = a 8.2.1. fejezet szerinti mintarendszer, *Hűségkártya* = egy pontgyűjtő alkalmazás

átlagot, a minőségi (természetességi) paraméterek eltérései a legfontosabbak témánk szempontjából. Még a legjobb rendszer is 0,6 ponttal kisebb értéket kapott a minőségi skálán, mint az alapvető funkcionalitást jelző érthetőségin. A legrosszabb esetben pedig a különbség az 1-et is meghaladja. Egyetlen rendszer közelíti meg a 4-es értéket, ami a hasonló vizsgálatok esetén a széles körű elfogadottság szintjének felel meg. A Hűségkártya jellegű megoldások egyik legfontosabb célja az, hogy kedvező képet alakítsanak ki az adott vállalatról. A jelen esetben az alacsony érték valószínűsíti azt, hogy a számfelolvasó rossz minősége ezt a hatást jelentősen csökkenti, vagy esetleg negatív irányba fordítja.

A legjobb és a legrosszabb természetességű rendszer között a különbség csaknem két jegynyi, ami megalapozhat egy „hagyományos mosópor <-> korszerű mosópor" jellegű marketingkampányt.

11.5. ábra. Különböző dátum- és időpontfelolvasók érthetősége és minősége (Szőke 2003) *Időpont 1* ill. *2* = távközlési vállalatok pontosidő-bemondása, *Időpont hangposta 1* ill. *2* = távközlési vállalatok hangpostájának az üzenet időpontját tartalmazó bemondása, *Kártyaidőpont* = előre fizetendő (pre-paid) előfizetéshez kapcsolódó időpontbemondás, *Természetes* = professzionális bemondó



A 11.5. ábrán különböző dátum- és időpontbemondó technológiák értékelését láthatjuk. Noha az *időpont*bemondások és az egyik *hangposta* érthetőség szempontjából egyaránt megközelítik a természetest, a minőségi oszlopok jelentős eltéréseket mutatnak. Itt már csak az *Időpont 1* rendszer áll közel a természeteshez. Az egyik hangposta és a pre-paid rendszer (*Kártyaidőpont*) minőségi értékelése olyan alacsony (2,5 alatt), hogy az valószínűleg az általános vállalati arculat leromlását is eredményezheti.

Az ebben a szakaszban bemutatott bemondások alapvető és az ügyfelek által gyakran használt rendszerelemek nagy forgalmú infokommunikációs szolgáltatásokban. Sajnálatos, hogy noha viszonylag kis befektetéssel minőségük jelentősen javítható lenne (ehhez hazánkban is rendelkezésre állnak a szükséges műszaki-tudományos ismeretek), egyenlőre a legtöbb vállalat nem fordít erre figyelmet.

A vállalatok akusztikai arculata tehát kialakul, függetlenül attól, hogy azt tudatosan formálják vagy sem. Első lépésként felhívtuk a figyelmet ennek a témakörnek a fontosságára. Rendszerezetten áttekintettük az akusztikai arculat területeit és működési hatásköreit. Kísérletekkel bizonyítottuk, hogy az infokommunikációs szolgáltatások értékelésében a minőségi/természetességi szempontok bevezetése finomabb megkülönböztetést tesz lehetővé. A rendszerek alaposabb tervezése és értékelése lényegesen hozzájárulhat a társadalom életminőségének növeléséhez. Ennek elhanyagolása az ügyfelek számára a természetestől távol álló interakciókat eredményez, ami ellensúlyozhatja a vállalati arculat javítására jelentős költségekkel más médiumokban kifejtett kampányok hatását.

12. fejezet
Példák a beszédtechnológia felhasználásának területeiről

A beszédtechnológia eredményei már kezdik átszőni a mindennapi életünket. Ebben a fejezetben olyan alkalmazásokat gyűjtöttünk össze, amelyek valamilyen formában beszédinformációs rendszereknek tekinthetők és a könyvben tárgyalt modellekre, eljárásokra támaszkodnak. Mindezt a teljesség igénye nélkül tesszük. A válogatásban az vezetett bennünket, hogy minden tárgyalt témakörből egy-egy konkrét, gyakorlati példát mutassunk meg, annak részletesebb ismertetésével. Külön fejezetet szenteltünk az egészségkárosodott emberek segítésére kidolgozott megoldásoknak, és adunk példákat arról is, hogy a beszédtechnológia eredményeit hogyan lehet használni az egészségügyben. Az alkalmazási paletta ebben a fejezetben kibővül még egy dimentózióval, ez pedig az alkalmazott hardvereszközök köre, ami hatalmas fejlődésen ment keresztül az utóbbi évtizedben. A tárgyalt modelleket, eljárásokat tehát folyamatosan más platformokra kell megvalósítani, alkalmazkodva az informatikai környezethez, vagyis az operációs rendszerekhez, a sebességi és memórialehetőségekhez és a hardver/szoftvereszközökhöz.

12.1. Beszédtömörítési megoldások a gyakorlatban

Tatai Péter

A következőkben röviden összefoglaljuk az Európában használatos fontosabb kódolási módszereket, amelyeket az ETSI (European Telecommunications Standards Institute) és az ITU (International Telecommunications Union) szabványosított. Nem térünk ki az amerikai és japán érdekkörben használatos és az ottani szervezetek által szabványosított kódolókra, valamint egyéb, például a tengerhajózásban használatos Inmarsat szabványra, azonban a módszerek és elvek ott is jórészt hasonlóak. Ennek kapcsán megjegyezzük, hogy a digitális telefonátvitel kezdetétől, a 30 és a 24 csatornás pulzus kód modulációs (Pulse Code Modulation (PCM)) rendszerektől kezdve

megtalálható az európai és az amerikai–japán szabványok kettősségének a vonulata a szabványosításban. Az ok nem műszaki, hanem gazdasági, elsősorban az európai cégek piacának védelme. Ugyanis az eltérő szabványok megnehezítik, hogy a fejlettebb amerikai és japán ipar szállítsa az európai távközlési eszközöket. Az interkontinentális távközlés által megkövetelt kompatibilitás szerencsére az együttműködő, szabványos berendezések irányában hat, mind a nagy sebességű optikai hálózatoknál, mind a mobilhírközlésben, ahol a 3G generációnál már szoros az együttműködés például a 3rd Generation Partnership Project (3GPP) szabványosítási törekvés keretében. A 12.1. táblázat a fontosabb szabványos ITU és ETSI kódolókat sorolja fel. Az első kiadási év csak közelítően mutatja a történelmi fejlődést, mert egy-egy szabvány kidolgozása tipikusan 5 év, és évekkel később is kiadnak kisebb-nagyobb módosításokkal újabb változatokat. Lényeges szempont minden kódoló algoritmusnál a

12.1. táblázat. Szabványos beszédkódolási eljárások és műszaki jellemzőik

Szabvány	Kiadási év	Bitsebesség (kbit/s)	Kódolási módszer	Alkalmazás	Keret/look ahead (ms)	Bonyolultság (MIPS)
ITU kodek						
G.711	1972	64	PCM	Általános, telefonos beszédátvitel	0.125	$<<1$
G.721	1985	32	ADPCM	Csatornatöbbszörözés, DECT drót nélküli telefon	0.125	1.25
G.722	1988	48, 56, 64	Részsávos ADPCM	TV, rádió kommentátor csat., ISDN telekonf.	0.125/1.5	5-10
G.722.1	1999-2005	24, 32	Transzform.	High definition VoIP (széles sávú beszéd)	20/20	<15
G.722.2	2002	6,6...23,85	AMR-WBACELP	3G, VoIP, telekonf. (széles sávú beszéd)	20/5	<40
G.723.1	1995	5.3/6.3	MP-MLQACELP	VoIP hálózat és multimédia, videofon	30/7,5	13
G.726, 727	1990	16, 24, 32, 40, 64	ADPCM	Csomagkapcs. hálózat, beágyazott kvantáló	0.125	2
G.728	1992	16	LD-CELP	Kis késleltetésű kommentárcsatorna	0.625	30
G.729, 729A	1995	8	CS-ACELP	Rádiós és multimédia hálózat, telekonferencia	10/5	11-20
ETSI kodek						
GSM 06.10	1987	13	RPE-LTP	GSM mobilkódoló (FR: Full Rate)	20	5-6
GSM-06.20	1995	6,3	VSELP	GSM mobilkódoló (HR: Half Rate)	20	14
06.60	1996	12.2	ACELP	GSM mobilkódoló (EFR:Enhanced FR)	20	14
NB-AMR	1999	4.75-12.2	AMR-ACELP	GSM, 3G, VoIP, WiFi, MMS, video stream stb.	20/5	14-20

bonyolultság, vagyis a megvalósításhoz szükséges műveletek száma, a MIPS (million instructions per second), amely általában egy DSP-n (Digital Signal Processor) futó programra utal, mert többnyire ilyen processzorokon valósítják meg a gyakorlat-

ban használatos tömörítő kódolókat. Azonban a bonyolultság csak tájékoztató érték, annál is inkább, mert a szabvány soha nem tartalmaz tényleges megvalósítási leírást, így az algoritmusfejlesztők és programozók ügyességén, valamint az aktuálisan alkalmazott DSP-k képességein is múlik a tényleges utasításszám.

A hasonló célú, sebességű és képességű kódolók összehasonlításának alapja tehát legfőképpen a bitsebesség (kbit/s) és bonyolultság (MIPS), amellyel egy adott minőséget el lehet érni. Utóbbit leginkább MOS (Mean Opinion Score) értékkel jellemzik (lásd a beszédminősítés fejezetet), amelynél a skála 1-től 5-ig terjed, és általában a 4 feletti, igen jónak számító minőség elérése a cél, de már a 3–3,5 tartomány is alkalmas telefonálásra. A táblázatban feltüntettük a keretidőt és az algoritmus által figyelembe vett következő jelszakasz hosszát, az ún. look ahead időtartamot is. A késleltetés általában kissé hosszabb, mint a két érték összege.

Az alkalmazások szempontjából fontos még az egyes kódolóknak a vonali bithibákra, illetve az internetes átvitelnél elkerülhetetlen csomaghibákra való érzékenysége. Általános szabály, hogy minél több redundanciát távolítunk el egy jelből, vagyis minél inkább tömörítjük, annál érzékenyebb lesz az ilyen jellegű hibákra. Emiatt hibavédő kódolást kell alkalmazni, ami természetesen megnöveli az átviteli bitsebességet. Ha pedig a hiba már nem kerülhető el, akkor annak hatását különféle algoritmusokkal (packet loss concealment) kell csökkenteni. A táblázatból kitűnik, hogy az ITU főként a fix telefonhálózatban, míg az ETSI a GSM mobilhálózatban használatos kódolókat szabványosította. A következőkben egész röviden ismertetjük az ITU, majd az ETSI szabvány szerinti fontosabb kódolókat.

12.1.1. Kódoló ajánlások

ITU-T ajánlások. Az ITU-T (az ITU távközlési ága, szemben az ITU-R rádiós ággal) szabványok elsősorban a közcélú kapcsolt telefonhálózattal foglalkoznak (PSTN: Public Switched Telephone Network). Amint korábban is említettük, ebben a hálózatban a szabványos, logaritmikusan kompandált, 64 kbit/s-os PCM kódolás terjedt el, amelyet a G.711 ajánlás ír le. Ennek a minősége megfelel a szokásos telefonhasználatnak, ezért ezzel vetik össze az újabb, tömörítést alkalmazó kódolókat. Azonban a PSTN átvitel hibaaránya igen alacsony, így ott a hibatűrés nem lényeges szempont. Valóban, a PCM igen érzékeny a vonali bithibákra, 10^{-6}-nál jobb hibaarányt várnak el ezért az átviteltől, 10^{-3}-nál pedig a vonal már teljesen használhatatlan. Ezzel szemben a csomagkapcsolt és még inkább a mobilhálózaton lényegesen nagyobb a hibaarány, illetve a csomagvesztésből eredő hiba, ezért ezt is figyelembe kell venni a kódolók összehasonlításánál.

A G.721, majd később a G.726 és 727 ajánlások az ADPCM egyre fejlettebb változatait szabványosítják. Ezeket is használják a PSTN területén, például a 32 kbit/s-

os változat a PCM-hez képest csatornaszám-duplázást tesz lehetővé, továbbá például a drót nélküli (Digital Enhanced Cordless Telecommunications (DECT)) telefonokban, üzenetrögzítőkben stb. A viszonylag kis bonyolultság miatt az ADPCM már az első generációs DSP elemekkel megvalósítható volt.

A G.727 ajánlás változtatható bitsebességű kivitele beágyazott (embedded) kódoló, amelynél a csomagkapcsolt hálózat terhelésének megfelelően változtatható a bitsebesség. Ennek érdekében az 5 bitre kvantált különbségi jelből csak 2 bites felbontású jelet csatolnak vissza a prediktor felé a visszacsatoló (helyi dekódoló) hurokban. Ily módon, mintánként 5, 4, 3 vagy 2 bites átvitel között lehet választani, és ez a bit-eldobás forgalmi torlódások esetén lényegesen kedvezőbb, mint a csomagvesztéssel történő sebességcsökkentés.

A G.722, 722.1 és 722.2 ajánlások teljesen eltérő kódolást alkalmaznak, de mindegyik széles sávú beszéd, azaz 50 Hz...7 kHz-es, úgynevezett televízió és rádió kommentátor csatorna minőséget nyújt. A legkorábbi, a G.722 ajánlás részsávos ADPCM technikát, a G.722.1 ajánlás pedig transzformációs kódolást alkalmaz, ezért jobban működik zenei jelekre, mint beszédre. Legkorszerűbb a G.722.2 szerinti AMR-WB jelű (Adaptiv MultiRate-WideBand) széles sávú kódoló, amelynél széles tartományban változtatható az átviteli bitsebesség a csatorna kapacitásának és a minőségi követelményeknek megfelelően. Az alkalmazott ACELP kódolás algebrai kódtáblát és CELP struktúrát jelent, amelynél a kódtábla felbontásával változtatható az átviteli sebesség. Az LP elv miatt ez a kódoló beszédre lényegesen jobb, mint zenei jelekre. Ugyanezen kódolót az ETSI, illetve a 3GPP is szabványosította a 3G hálózathoz, ennek keskeny sávú (NB: Narrow Band) változatával együtt.

A G.723.1 ajánlás szerinti kódoló 6,3 kbit/s esetén 24 bites keretekkel és MP-MLQ (Multipulse LPC with Maximum Likelihood Quantization) algoritmussal, 5,3 kbit/s-nál pedig 20 bites keretekkel és ACELP algoritmussal működik. Főként VoIP alkalmazásoknál használják, és ez a szabványos audioátviteli mód a H.324 ajánlás szerinti videoterminálóknál.

A G.728 ajánlás szerinti LD-CELP, kis késleltetésű (Low Delay) kódoló, amely a megismert CELP struktúrában működik, de a hurokban hátra irányú adaptációt alkalmaz, ezért nem kell bufferben tárolni a jelet, ami legalább egy keretnyi késleltetést okozna. 50 LPC paraméter biztosítja az összevont rövid és a hosszú idejű predikciót, továbbá visszafelé irányuló erősítésadaptációt is használ. Ez a kódoló is széles sávú, tehát jó minőségű beszédátvitelt nyújt.

A G. 729 ajánlás CS-ACELP (Conjugate-Structure Algebraic-CELP) elvet használ, és 8 kbit/s sebesség mellett viszonylag jó beszédminőséget nyújt. VoIP hálózaton például Skype telefonálásra alkalmazzák. Vannak kisebb sebességű kiterjesztései is. Ez a szabvány, miként a fentebb említettek többsége is, fejlődött az idők során, és számos újabb kiadás jelenik meg évente.

ETSI kódoló ajánlások. Az ETSI szabványok a GSM hálózat kialakulásával jöttek létre a 80-as évek végén, amit a legnagyobb európai távközlési cégek közösen támo-

gattak. A GSM kódolók nagy késleltetése miatt azonban ezeket az ITU nem tartotta megfelelőnek a hálózatban való alkalmazásra, és nem kívánta szabványosítani, de a mobilhálózat iránti igény olyan erős volt, hogy létrehozták az ITU-tól független ETSI szervezetet, amely (az ITU ellenkezése ellenére) szabványosította a GSM kódolókat és a mobilhálózat többi elemét.

12.1.2. A kódolók fejlődése

A kódolók közül az első, a GSM FR (Full Rate) kódoló forradalmi áttörést hozott a mobiltelefóniában. Az alkalmazott RPE-LTP módszer a korábban megismert nyílt hurkú (AaS) struktúra azzal az eltéréssel, hogy a hosszú idejű prediktor ADPCM struktúrában működik. Az RPE (Regular Pulse Excitation) arra utal, hogy a 20 ms keretidővel (160 minta) működő STP maradékjelét az LTP céljára decimálják, harmadára veszik a mintavételi frekvenciát. Ezáltal a fázishelyzettől függően 4 db, 13 mintából álló jelsorozat áll elő, amelyek közül a legkisebb LTP hibát adó sorozatot választják. Ez a módszer a jó érthetőség mellett jelentősen csökkenti a bitsebességet, 13 kbit/s-ra, amelyhez még 3 kbit/s hibavédő kódolás társul, ami a mobilátvitel gyakori bithibái miatt szükséges.

A másik jelentős előrelépés az FR kódolónál az LP együtthatók kvantálási eljárása. Már említettük, hogy az LPC paraméterek közvetlen kvantálása nem hatékony az együtthatók kedvezőtlen eloszlása, valamint az egymást követő keretek közötti alacsony korreláció miatt. Emiatt a paramétereket először úgynevezett Parcor (Partial Correlation) paraméterekké transzformálják. Ezek a rácsszűrő kialakítású analízis és szintézis szűrő paraméterei. Nevezik őket reflexiós együtthatóknak is, mert a beszédkeltés csőmodelljében a változó keresztmetszetű csőszakaszok átmeneténél keletkező hullám reflexiójának az együtthatói. Kellemes tulajdonsága a Parcor együtthatóknak, hogy ha értéktartományuk +1 és -1 között van, akkor stabil a szűrő. Kvantálás előtt az értékek eloszlásának javítására ezeket az együtthatókat tovább transzformálják LAR (Log Area Ratio) formában az alábbiak szerint,

$$L_i = ln\frac{1-k_i}{1+k_i}, \tag{12.1}$$

ahol k_i a Parcor együttható. (Megjegyezzük, hogy az irodalomban k_i ellenkező előjellel is előfordul, továbbá L_i esetén a fenti érték fele, továbbá tizes alapú logaritmus is használatos.) A LAR paraméter onnan kapta a nevét, hogy ez éppen az említett csőszakaszátmeneteknél található területarányoknak felel meg (Rabiner–Schafer 1978).

A mobilhálózat jobb kihasználása érdekében kifejlesztették és bevezették a fele sebességű HR (Half Rate) GSM kódolót, amely 5,6 kbit/s sebességen működik, és így a minőség rovására lehetővé teszi a forgalom megduplázását. A készülékek

többsége ebbe az üzemmódba is kapcsolható, ha lecsökken a telep töltöttsége, így a kisebb fogyasztás miatt még hosszabb ideig működhet a mobiltelefon. Az alkalmazott VSELP (Vector Sum Excited LP) módszer szintén CELP algoritmus, amely azonban kevésbé jól tolerálja a háttérzajt, ezért nem tekinthető perspektívikusnak.

Jelentős előrelépést hozott azonban az EFR (Enhanced Full Rate) GSM kódolás, amely a 7.30. ábra szerinti, fejlett kódolási struktúrát használja. További jelentős javulást hozott itt a forradalminak is nevezhető LSP (Line Spectral Pair) alapú kvantálás alkalmazása a szintézisszűrő paramétereinél, amelynek során az LPC paramétereket LSF (Line Spectral Frequency) komponensekké transzformálják. Ehhez az *A(z)* függvényt két polinomra bontják:

$$R(z) = A(z) + z^{-p+1}A(z^{-1}), \tag{12.2}$$

$$Q(z) = A(z) - z^{-p+1}A(z^{-1}), \tag{12.3}$$

ahol *R(z)* felel meg a gége zárt, *Q(z)* pedig a nyitott állapotának. Amíg *A(z)*-nek komplex gyökei vannak valahol az egységkör belsejében, addig *R(z)* és *Q(z)* gyökei kizárólag az egységkörön találhatóak, vagyis ezeket elegendő a $z = e^{j\omega}$ változó $0...\pi$ tartományában kiértékelni. A spektrumpár elnevezés onnan következik, hogy a gyökök szimmetrikus párokban adódnak. További fontos tulajdonság, hogy ha *R(z)* és *Q(z)* gyökei felváltva következnek, akkor a szűrő stabilitásának szükséges és elégséges feltétele, hogy a gyökök monoton növekedőek legyenek. Továbbá a spektrumpárok „közrefogják" a beszédjel spektrális csúcsait, és minél közelebb vannak egymáshoz, annál élesebb a rezonancia, vagyis a formánsok kiemelése. A LSP alkalmazása igen jelentősen javította az LPC szintézis minőségét, mert kevéssé érzékeny a kvantálásra, nem lépnek fel stabilitási problémák, és legfőképpen, az egymást követő keretek LSP paraméterei folyamatosan változnak, tehát jól interpolálhatók.

Az AMR-NB (Adaptiv MultiRate-NarrowBand) kódoló hasonló a széles sávú változatához (lásd G722.2), amely ma a legszélesebb körben használatos keskeny sávú (300–3400 Hz) technológia. A jó minőség és a jó zajtűrés miatt nemcsak a mobil (GSM, UMTS) és a VoIP hálózatban, valamint multimédia-szolgáltatásoknál, hanem beszédjelek tárolásánál és médialejátszóknál is kiterjedten alkalmazzák, mint perspektívikus módszert. Dinamikusan adaptálható a hálózati viszonyokhoz, mert az átviteli sebessége 4,75-től 12,2 kbit/s-ig változtatható. Forgalmi torlódás esetén kisebb sebességre állítható, de ekkor is viszonylag jó minőséget nyújt, és 7,4 kbit/s-nál már megfelel a telefonhálózatban elvárt átviteli követelményeknek. Magasabb sebességeknél fokozatosan nő a robusztussága (zajtűrése) és a nem beszéd jelek átviteli minősége.

12.2. Gépi beszédminősítés távközlési rendszerekben

Fegyó Tibor

A telefóniában az átvitt beszéd minősége fontos jellemző. Ennek mérése szubjektív módon igen költséges és időigényes (ITU-T P.800 1996), ezért szükség van automatikus eljárásokra, melyek eredménye jól közelíti a szubjektív minősítést. A beszédminősítés feladata nem újkeletű a távközlési szolgáltatók számára. Mind az analóg vezetékes telefonok, mind a digitális mobiltelefonok, illetve az IP telefonok kódolóinak bevezetésekor nagy hangsúlyt fektettek a felhasználók elégedettségére, és nemzetközi ajánlások írják elő a minőség mérésének módszereit (Beerends–Stemerdink 1994), (Beerends 1998), (ITU-R BS. 1998), (ITU-T P.800 1996), (ITU-T P.861 1996), (ITU-T P.830 1996), (Thiede et al. 2000). A beszédminősítő eljárások használatával lehetőség van többek között:

- a kódoló algoritmusok vizsgálatára és optimalizálására,
- a beszédminőség mérésére és rendszeres monitorozására,
- nagyobb átviteltechnikai rekonstrukció után a minőség változásának kimutatására,
- országos szintű beszédminőségtérkép elkészítésére.

Ezek a feladatok objektív módszereket és automatikus feldolgozást igényelnek, hiszen a szubjektív minősítés hatalmas feladat lenne, amit csak időközönként az objektív minősítő eljárások kalibrálására használunk.

12.2.1. Hanganyag gyűjtése

A beszédátviteli rendszer kiértékeléshez rendszeresen hanganyagot kell gyűjteni. A mai hatékonyan működő beszédminősítő rendszerek összehasonlítási alapon működnek, ezért aktív mérésekre van szükségünk, azaz a hálózat egyik pontján ismert beszédjelet adunk be a hálózatba, és a másik pontján pedig rögzítjük az átvitt jelet, és a mérések során ezt a két jelet hasonlítjuk össze. Fontos, hogy változatos beszédet tartalmazó jelket vigyünk át, különböző hangokat, különböző beszélőktől. Színuszos jelekkel a modern kódolók minősége nem mérhető. Egyszerűbb lenne a hálózat központi elemében rögzíteni beszédet, de akkor nem lenne referenciánk, azaz nem tudnánk mihez hasonlítani a rögzített felvétel minőségét. A rögzített hanganyag értékelésére szubjektív és gépi, azaz objektív módszerek is léteznek.

12.2.2. Szubjektív beszédminősítés

A szubjektív beszédminősítés a gyakorlatban csak kis mintán (néhány száz mondat) alkalmazható, egyrészt mert a hallgatók idővel elfáradnak és nem lesz konzisztens az értékelés, másrészt mert időigényes. Az objektív minősítő eljárás kalibrálásához, verifikálásához azonban elengedhetetlen a szubjektív minősítés alkalmazása. A különböző szubjektív minősítési eljárásokat az ITU-T P.800-as ajánlásai definiálják (ITU-T P.800 1996). A következőkben röviden ismertetjük a módszereket.

Abszolút értékelés előre definiált skála alapján (ACR – Absolute Category Rating). Az abszolút értékelés során a minőségskálát definiálják a minősítést végző embereknek, akik a mondatok meghallgatása után minden mondatot besorolnak egy kategóriába. Ez a megoldás személyenként nem tesz lehetővé túl finom felbontást, mivel az ember számára önmagában egyetlen mondatot nem egyszerű besorolni. Pontos eredmény eléréséhez ezért sok tesztelő ember szükséges.

Jelenségészlelési tesztek. A jelenségészlelési tesztek során bizonyos jelenségek meglétét vagy nem létét kell meghatározni. Ilyen lehet például a zaj, a visszhang, az áthallás. Ezen jellemzők segítségével lehet meghatározni, hogy az ember milyen típusú zajokra mennyire érzékeny.

Romlás megfigyelése az eredetihez képest (DCR – Degradation Category Rating). Az eredeti jellel történő összehasonlítás már pontosabb eredményt adhat, mint az abszolút besorolás, mivel itt már van legalább egy fix viszonyítási pont, a minőségskála felső pontja, azonban ez csak jó minőségű felvételeknél alkalmazható, mert zajos esetekben továbbra is problémát jelenthet a besorolás.

Referenciarendszerrel történő összehasonlítás. A legpontosabb eredményt akkor kapjuk, ha referenciarendszert állítunk föl, és az alapján soroljuk be a minősítendő hanganyagot. A hallgatónak ismert minőségű mondatokhoz képest kell meghatározni az adott mondat minőségét. Ezzel a referenciasorozattal egy teljes skála lefedhető, és így páros összehasonlítások sorozataként meghatározható egy mondat minősége. Nagy előnye ennek a megoldásnak, hogy pontosabb az ACR-nél, kevesebb ember kell a teszteléshez, viszont sokkal időigényesebb, mivel egy-egy mondat minősítéséhez több mondatpár meghallgatására is szükség van. További problémát jelent a referenciahalmaz előállítása. A kerek osztályzatokhoz tartozó referenciák meghatározása általában az ACR segítségével történik, de célszerű olyan mondatokat választani, ahol az egyének véleményének szórása alacsony. Ha finomabb felbontásra is szükség van, akkor páros összehasonlítások segítségével iteratív módon lehet az egyes kategóriákat felbontani. Az eljárások eredménye a MOS érték (mean opinion score), sok ember véleményének átlaga. Ahány eljárás, annyiféle MOS definiálható, és ezek nem vethetőek össze egymással, tehát az ACR által kapott MOS értékéből nem állítható elő a DCR eredménye (ITU-T P.800 1996). A szubjektív tesztek elvégzése ugyan drága és időigényes feladat, azonban kis mennyiségű szubjektív mérésre

mindenképpen szükség van, mivel ezek eredményeit használjuk fel a következő eljárásokban:

- az objektív minősítő kalibrálására,
- az objektív minősítő ellenőrzésére.

Szubjektív beszédminősítési kísérletek. A minősítési kísérletek közül a legegyszerűbb és a leggyorsabb az abszolút értékelés, ezért a gyarkolatban is ezt használják a legtöbbet. A mérés első lépése a minőségi kategóriák definiálása, amelyek közül a hallgatónak választania kell. A hat kategória, a nullát kivéve igazodik a szabványhoz, de a gyakorlatban a nullás kategória is fontos információt mond a hálózatról.

- 5 tökéletes telefonminőség
- 4 a háttérben kicsit zajos, de nem zavaróan
- 3 a háttérben zajos, de még tökéletesen érthető
- 2 a háttérben zajos, de még érthető
- 1 a zaj a beszélgetést már zavarja
- 0 a zaj miatt nem érthető

Amennyiben nem pusztán a szubjektív mérés a cél, hanem későbbiekben objektív minősítőt is szeretnénk készíteni egy adott rendszerhez, akkor célszerű három értékelendő halmazt összeállítani. Az első halmaz a kalibrációs halmaz, amely az objektív minősítő eljárás paramétereinek optimalizálásához szükséges, a második halmaz a teszthalmaz, amelyen az objektív eljárást teszteljük. Ezt a két halmazt általában egyszerre értékeltetjük a szubjektív teszt során. Tipikusan 100, maximum 200 mondatból áll ez a teszt, amelynek mintegy 10%-a jelenti a kalibrációs halmazt és a további 90% pedig a teszthalmazt. Az objektív minősítés során az lesz majd a cél, hogy a teszthalmazon a lehető legjobb eredményt érjük el. A kiértékeléshez legalább 10–15 ember szükséges, de a rendszerek méretétől és a teszt jelentőségétől függően akár több ezer fős tesztek is előfordulhatnak. Az előző két teszthalmaztól, illetve teszttől függetlenül készíteni kell rendszeres időközönként egy harmadik, validációs halmazt is, melynek mérete az előző halmazzal azonos. A validációs halmaz célja, hogy az eredeti tesztektől független halmazon is kiértékelhessük a beszédminősítő rendszert. Az idők során egy hangátviteli rendszer bizonyos részei megváltozhatnak, például technikai fejlesztések során új bázisállomások, újfajta telefonok vagy kódolók kerülhetnek bevezetésre, valamint változhat a felhasználók értékelése is az általános technológiai fejlődés következtében. Meg kell vizsgálni, hogy az objektív minősítő rendszer továbbra is illeszkedik-e a szubjektív véleményekhez. A szubjektív tesztek során előfordulhatnak téves értékelések, részben azért, mert a hallgató elfárad a mérés során, és már nem konzisztensen értékel, részben véletlenül. Az ilyen hibák kiküszöbölésére első lépésben az egyes felvételekre adott értékelések közül a legjobbat és a legroszabbat nem vesszük figyelembe. Következő lépésben megvizsgáljuk, hogy az átlagos mondatonkénti eredmények és az egyes értékelők eredményei

között mekkora a korreláció. Ha mindenki egyformán értékelne, akkor a korrelációnak 1-nek kellene lenni, de a gyakorlatban egy konkrét mérés során ez a korreláció 0,87–0,95 között mozog. Egyértelműen magasabb értéket kaptunk gyakorlottabb értékelőktől. A néhány száz mondatos teszteket nem tekinthetjük olyanoknak, amivel egy rendszert kimerítően lehet tesztelni, viszont alkalmasak arra, hogy egy objektív rendszer kidolgozását segítsék. Ennek lépéseit mutatjuk be a következő fejezetben.

12.2.3. Objektív beszédminősítő eljárások áttekintése

A távközlési csatornákon áthaladó beszéd minőségének objektív mérésére számos eljárás került kidolgozásra (Ascom Infrasys AG 1998), (Ascom Infrasys AG 1999a), (Ascom Infrasys AG 1999b), (Ascom Infrasys AG 2000), (Ericsson NetQual Inc. 2000), (Alonso Frech et al. 1997), (ITU-T P.861 1996), (ITU-T P.862 2000), (Kaenel 1998). Ezek célja a szubjektív minősítés szintjének közelítése. Beszédkódoló eljárások hatékonyságának vizsgálatakor elterjedten használják a jel-zaj viszonyt, azonban ez nem hatékony a beszédminőség jellemzésére, különösen kis sebességű kódolók esetén. Mesterséges szinuszos jelek átvitele alapján mérhető például SINAD (Signal to Noise and Distortion Ratio), de a kódolókat nem szinuszos, hanem beszédjelre fejlesztették ki, ezért ez hamis képet mutat az átvitelről. A gépi minősítő rendszerek többsége a kódolókat ezért beszédmintával értékeli, többnyire összehasonlítási alapon, melyre különböző objektív távolság mértékeket dolgoztak ki (ITU-T P.861 1996):

- LPC kepsztrumtávolság,
- információs index,
- koherenciafüggvény,
- mintaillesztés-alapú módszerek, például EPR (Expert Pattern Matching),
- érzetimodell-alapú beszédminőség-mértékek, amelyek mel-, illetve barkskála-alapú spektrális vagy kepsztrális távolságokon alapulnak. Idetartozik a kis sebességű kódolók esetére kidolgozott PSQM (Perceptual Speech Quality Estimation Method) eljárás (Beerends–Stemerdink 1994), illetve az IP alapú beszédátvitel sajátosságai miatt módosított PESQ (Perceptual Evaluation of Speech Quality) eljárás.

Ezeket az eljárásokat megvizsgálta az ITU-T a beszédminősítési eljárásra vonatkozó ajánlás kidolgozásakor, és kis bitsebességű kódolók esetén a legjobb korrelációt elérő, érzeti modell alapon működő PSQM (Perceptual Speech Quality Estimation Method) eljárásra esett a választás (ITU-T P.861 1996), amelyet később felváltott a komplexebb PESQ (ITU-T P.862 2000) megoldás.

12.2.3.1. Az objektív minősítő eljárás lépései

Előfeldolgozó. A minősítési eljárások során a referencia- és a torzított jelet időben illeszteni kell egymáshoz a megfelelő eredmény eléréséhez. Ehhez a következő lépéseket kell elvégezni:

Időillesztés. A hálózaton átvitt jel késleltetést szenved, ezért szükséges a referencia- és a vett jel pontos illesztése egymáshoz. Ez általában korrelációs csúcs keresésével történik. Különösen nehéz az időillesztés megvalósítása csomagkapcsolt beszédátvitel esetén, ahol a csomagok egyedi késleltetés-ingadozásából adódóan beszédrészek kimaradhatnak, illetve szünetek vagy ismétlések kerülhetnek jelbe. Csomagkapcsolt esetben nem a teljes jelet illesztik, hanem a beszédjel kisebb részeit.

Energiaillesztés. A minősítő eljárások általában megkövetelik, hogy a referencia- és a torzított jel azonos energiájúak legyenek, ezért a két jel átlagos energiáját azonos szintűvé kell tenni a DC szint (egyenkomponens) eltávolítása után. Lehetőség van hangosságérzeti kiegyenlítésre is.

Szegmentálás. A felvételeket rövid idejű átlapolt keretekre vágják, ablakozás után (például Hamming-ablak) ezeken történik a további feldolgozás.

Pszichoakusztikai modell. A pszichoakusztikai modellezés az emberi hallásmodellen alapuló eljárás (lásd a 3.4. fejezet), amelynek lényege, hogy az ember által észlelt különbségeket emelje ki a jelből, a nem észlelteket pedig elnyomja. A lineáris spektrumtartomány nem felel meg ezeknek a követelményeknek, a következőkben felsorolt további leképezéseket kell alkalmazni.

Idő-frekvencia leképzés. Az emberi hallásmodelleknek megfelelően az időtartományi jelet frekvenciatartományra kell transzformálni. A rövid idejű spektrumnak csak az amplitúdóját használjuk a továbbiakban, mert a hallás a fázistorzításra nagymértékben érzéketlen.

Kritikus sávú szűrők alkalmazása. A lineáris frekvenciatartományról az érzeti tartományra (bark- vagy mel-skálára) kell transzformálni a jelet, mivel a kísérletileg meghatározott érzeti tartomány az emberi hallórendszerben meglévő leképzésnek felel meg.

Hangosságkiegyenlítés. Az érzeti hangosság a jelenergia nemlineáris függvénye, ezért a jelet transzformálni kell az érzetihangosság-tartományra.

Időmaszkolás. A hallórendszer két egymást gyorsan követő rövid jelet nem képes megkülönböztetni. Kísérletekkel meghatározták, hogy egy rövid idejű (szinuszos) jel maszkolja a megelőző és őt követő halkabb jeleket. Ennek megfelelő időbeni maszkolást kell alkalmazni a minősítő eljárásokban is.

Frekvenciamaszkolás. Két egymáshoz közel álló szinuszos jelből az erősebb elfedi, maszkolja a gyengébbet. A maszkoló hatás függ a jel frekvenciájától és intenzitásától is. Ezt a hatást szintén modellezni kell a pszichoakusztikai modellben.

Minőségbecslő modell. A referencia- és a torzított jelen is alkalmazott pszichoakusztikai modell egy sokdimenziós jellemző vektorsorozatot állít elő, melyből meg kell határozni az objektív minőséget a következő eljárások alkalmazásával.

Súlyozott összehasonlítás. Általános módszer a távolságszámításra a keretenkénti euklideszi távolságok átlagának képezése. Az egyes keretek súlyozhatóak aszerint, hogy mekkora a jel energiája. A nagyobb energiájú részek fontosabbak a beszédértés, és ezzel a beszédminőség mérésének szempontjából. Az alacsony energiájú szünetet, zajt tartalmazó részeket 0-val célszerű súlyozni, mert a beszéd érthetőségét ezek a részek nem befolyásolják, illetve ha már zavaró mértékű a zaj, akkor a beszéd közben is zavaró.

Transzformálás a MOS skálára. Az összehasonlítás eredménye egy távolság, ami a torzítás növekedésével monoton nő. Ez a távolság monoton, de nemlineáris kapcsolatban van a szubjektív minőséggel, ezért megfelelő leképzéssel a MOS skálára kell transzformálni. Első lépésben linearizálni kell a kapcsolatukat, majd egy lineáris leképezéssel azonos skálára kell hozni az objektív és a szubjektív értékeket.

A fenti elvek alapján több beszédminősítő eljárás került kidolgozásra. Az ITU-T ajánlásban szereplő PSQM eljárás általában a kis sebességű kódolókra alkalmazható, továbbfejlesztése a PESQ eljárás Voice over IP rendszerekre optimalizált. Az ASCOM által kidolgozott QVoice (Ascom Infrasys AG 1998), (Ascom Infrasys AG 1999a), (Ascom Infrasys AG 1999b), (Ascom Infrasys AG 2000), az Ericssonban kidolgozott Auryst (Ericsson NetQual Inc. 2000), illetve a spanyol telefontársaságnál alkalmazott Coverage Information System (CIS) (Alonso Frech et al. 1997) a GSM rendszerekre lett optimalizálva. Az ITU ajánlások alapján készült magyar fejlesztésű Qualiphone-A analóg mobiltelefonokra (PSQM szerint), a Qualiphone-D pedig Vo-IP rendszerekre optimalizált (PESQ szerint). A különbségek részletei az egyes cégek belső információi, azokhoz nem férhetünk hozzá.

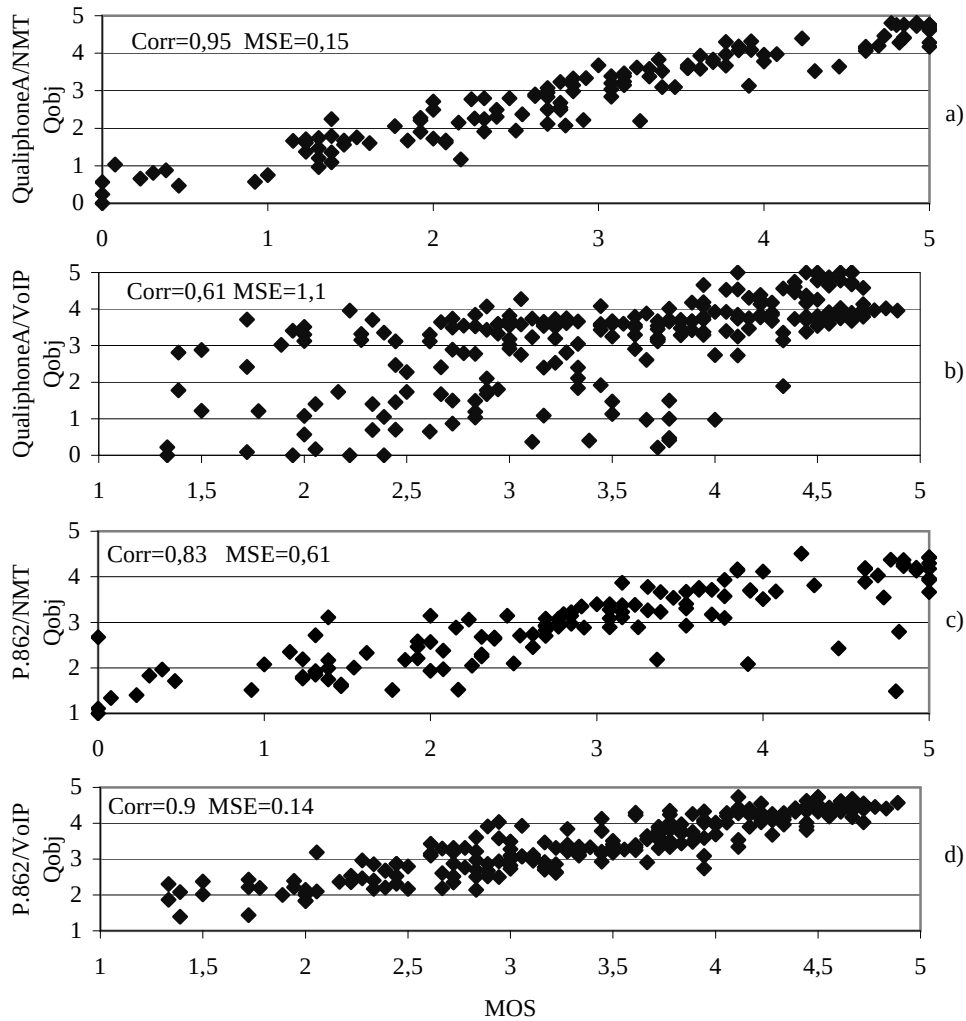
12.2.3.2. Az objektív minősítő eljárások értékelése

Az objektív minősítő eljárásokat minden esetben egy szubjektív minősítési eredményhez képest lehet értékelni. Célszerűen az optimalizálás során használt halmaztól független validáló halmazon. Az értékelés során két mérőszámot vizsgálunk, egyik a korreláció (Corr), másik az átlagos négyzetes eltérés az MSE (mean square error).

$$Corr = \frac{1}{N}\frac{\sum_{i=1}^{N}\{(x_i-\mu_x)(y_i-\mu_y)\}}{\sigma_x\sigma_y}, \tag{12.4}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2, \tag{12.5}$$

ahol N mondat esetén az i-ik mondat szubjektív minősége x_i, objektív minősége y_i, x_i-k átlaga:μ_x, szórása σ_x, és hasonlóan y_i-k átlaga μ_y, szórása σ_y. A korreláció abszolút értéke mindig 0, és 1 közé esik. $Corr = 1$, ha x_i, és y_i lineáris művelettel egymásba átvihető, 0, ha függetlenek egymástól. Nyilvánvaló elvárás, hogy gyakorlati esetekben 1 körüli korrelációt kapjunk, de azt láttuk a 12.2.2. fejezetben, hogy emberek esetén sem értük el az elméletileg maximális korrelációt. Az MSE értékét nulla körüli értékre szeretnénk szorítani, azaz elvárás, hogy a szubjektív és az objektív mérték azonos skálán mozogjon. Az előző rész végén bemutatott néhány eljárásból a Qualiphone-A és Qualiphone-D rendszereket vizsgáltuk a gyakorlatban is. Analóg mobil- és IP telefonrendszerből is rögzítettünk 1-1 teszthalmazt, és azokon teszteltük az egyes eljárásokat. Mivel a Qualiphone-D alapját képező PESQ eljárás a Qualiphone-A alapját képező PSQM-nek kiterjesztése, így első gondolatra mind analóg, mind IP telefonrendszerek esetén a Qualiphone-D-től várjuk a pontosabb eredményeket. Azonban az eredeti ITU-T ajánlások kiegészítésre, optimalizálásra kerültek, a Qualiphone-A analóg, a Qualiphone-D pedig csomagkapcsolt rendszerre optimalizált, így a gyakorlatban az analóg felvételeken a Qualiphone-A, csomagkapcsolt felvételeken pedig a Qualiphone-D teljesített jobban. A 12.1. a), és 12.1. d) ábrákon láthatóak a telefonrendszerek sajátosságaihoz jól illesztett beszédminősítő eljárások tesztelési eredményei. Mindkettő jól teljesített, csomagkapcsolt esetben 0,9, analóg esetben 0,95 az objektív és a szubjektív értékek korrelációja. A Qualiphone-D ugyan újabb, hatékonyabb módszer, de a csomagkapcsolt átvitel sajátosságai miatt a feladat is nehezebb, ezért alacsonyabb a korreláció a 12.1. d) ábrán, mint a 12.1. a) ábrán. Megvizsgáltuk a minősítő eljárásokat illesztetlen esetben is, azaz csomagkapcsolt telefonhálózatban az analóg, míg analóg hálózatban a csomagkapcsolt hálózathoz illeszkedő mérési eljárást alkalmaztuk. A 12.1. c) ábrán láthatjuk, hogy analóg rendszer esetén az újabb, de a feladathoz nem illesztett Qualiphone-D módszer lényegesen gyengébb eredményt ad, mint a régebbi, de az analóg mobilrendszerek sajátosságaihoz jól illesztett Qualiphone-A módszer. A 12.1. b) ábrán egyértelműen láthatjuk, hogy a PSQM-re épülő Qualiphone-A eljárás komoly hiányosságokkal bír csomagkapcsolt beszédátviteli rendszerek esetén, az értékelés gyakorlatilag használhatatlan, hatékonysága messze elmarad a Qualiphone-D mögött. A gyakorlati kísérletek megmutatták, hogy készíthető hatékony automatikus beszédminősítő eljárás, amely mind a korreláció, mind az átlagos négyzetes eltérés szempontjából gyakorlatban használható eredményt ad, de azt is láttuk, hogy nincs egy univerzális beszédminősítő, az eljárásokat optimalizálni kell az adott átviteli rendszerhez. A Qualiphone-A eljárás a Westel-0660 hálózatában 1999-től működött, a Qualiphone-D eljárást a Matáv és az Ericsson kísérleti laboratóriumaiban használták.



12.1. ábra. A Qualiphone-A eljárás tesztelése analóg mobil-beszédátvitelnél a) és csomagkapcsolt esetben b), valamint a Qualiphone-D eljárás tesztelése mobil-beszédátvitelnél c) és csomagkapcsolt eljárásnál d)

12.3. Telefonos és mobilos alkalmazások

A beszédinformációs rendszerek alkalmazásának egyik legkézenfekvőbb területe a távközlés. A telefon lehetőséget ad széles lakossági rétegek számára elérhető automatikus szolgáltatások fejlesztésére. Ezek a rendszerek lehetnek egyoldalú beszédinformációs megoldások, illetve beszédalapú dialógusrendszerek. Az egyoldalú rendszerekben csak a gép beszél az ügyfélhez, a természetes dialógusos megoldásokban már az ember is beszélhet (korlátozott közléseket mondhat), amit a gép azonosíta-

ni tud, majd a választ szintetizált beszéddel adja meg. Mindkét megoldásra találunk példákat ebben a fejezetben, konkrét, működő alkalmazások formájában.

12.3.1. Telefonról elérhető e-levél felolvasó

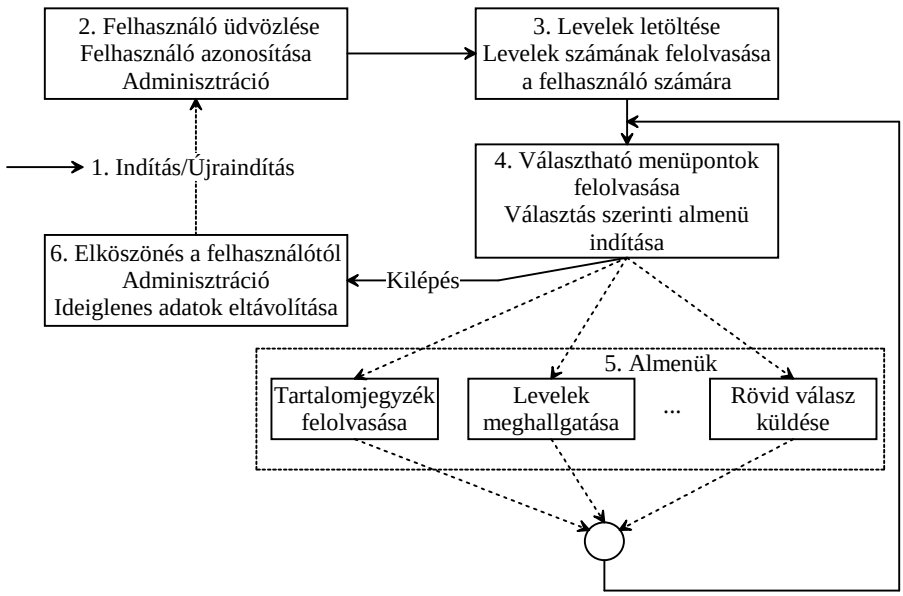
Németh Géza–Zainkó Csaba

Az elektronikuslevél-felolvasó (Németh et al. 2000) arra szolgál, hogy telefonon meghallgathassuk elektronikus leveleink szövegét (számítógép használata nélkül). A szolgáltatás bármilyen nyomógombos – DTMF jelzés küldésére alkalmas – mobil- vagy vezetékes készülékről elérhető, tehát jelentősen kiszélesíti az elektronikus levelekhez való hozzáférés módozatait. Az elektronikus levelek felolvastatása csak kiegészíti az egyéb, megszokott hozzáféréseket, ugyanakkor korlátokkal is rendelkezik:
Hosszú szövegállományok. Az ember sokkal nagyobb mértékben hagyatkozik a látására, mint a hallására, látás útján sokkal gyorsabban és több információt képes befogadni, mint hallás útján. Többoldalas szöveget hosszú ideig tart felolvasni. Egy 2 kbyte-os magyar szöveg (mintegy 30 sor) felolvasása 3,5 percig tart.
A rendszer által nem támogatott formátumú szöveg. Az e-levelekben használt szövegformátumok és karakterkészletek száma nagy és növekszik, ezért nem ismert formátum esetén a rendszer nem képes átalakítani a szöveget beszéddé.
Képek, egyéb állományok. A felolvasó nem alkalmas képi jellegű állományok tartalmának elmondására. Hasonló a helyzet nem nyelvi tartalommal rendelkező karakteres állományok feldolgozásánál is (például, programkódot nem lehet szövegként felolvasni).
A rendszer által nem támogatott idegen nyelvű szöveg. Az elektronikus levél felolvasása nyelvfüggő, tehát ha a szöveget nem a saját nyelve szerint olvassák fel, akkor nem érthető a beszéd.

A megvalósított rendszer működésének folyamatábráját a 12.2. ábra mutatja.

1. A felhasználó tárcsázza a szolgáltatást. Más-más telefonszámon különböző nyelvű rendszer érhető el, tehát már a kezdeti bejelentkező szöveg is a kívánt nyelven szólal meg.
2. Az üdvözlés és a felhasználó azonosítása után a rendszert lehet használni.
3. A gép letölti a megfelelő leveleket a szerverről, és feldolgozza azokat a tartalomjegyzék elkészítéséhez. Ezalatt a felhasználóval közli leveleinek számát, és megkéri, hogy a felsorolt menüből válassza ki a kívánt funkciót (lásd később).
4. A felhasználó bármikor kérhet szóbeli segítséget a rendszer kezelésével kapcsolatosan.
5. A levelek tartalomjegyzékét, tartalmát meg lehet hallgatni, illetve egyéb szolgáltatásokat is igénybe lehet venni (például válaszolni hangban, faxon továbbküldeni stb.).

6. Kilépéskor a rendszer helytakarékossági és adatvédelmi szempontból, minden olyan felesleges információt, adatot, hangfájlt töröl, amire már nincs szükség. A rendszer eltárolhatja a felhasználó cselekvéseit, hogy az esetleges visszaélések felderítését segítse.



12.2. ábra. Az elektronikuslevél-felolvasó működésének folyamatábrája

A rendszer szolgáltatásai

Tartalomjegyzék meghallgatása. A levelekről, feladójukról és tartalmukról a legtömörebb információt a tartalomjegyzék tartalmazza. A bejött leveleket és tárgyukat sorra elmondja a gép. Ha a tartalomjegyzék meghallgatása közben egy olyan levélhez érkezünk, amely érdekes számunkra, tehát a levél tartalmára is kíváncsiak vagyunk, akkor egy gomb megnyomásával lehetőségünk van a levél teljes szövegének a meghallgatására.

Levelek meghallgatása. Ha kevés levelünk érkezett, vagy válogatás nélkül meg akarjuk hallgatni az összes levelet, akkor ezzel a funkcióval a rendszer sorban felolvassa az összes levelet. Lehetőség van arra is, hogy egy levél felolvasását megszakítsuk, a következő levélre ugorjunk, vagy az adott levélben mondatonként előre-hátra mozogjunk.

Levelek törlése. A felesleges leveleket törölhetjük, így például a következő meghallgatásához gyorsabban eljutunk.

Kiegészítő szolgáltatások. A felhasználó rövid választ küldhet (írásost vagy hangüzenetet). Az írásos formák sablonok (1. gomb: Rendben. ; 2. gomb: Válaszolok később

stb.). Lehetőség van a meghallgatott levél faxon történő továbbküldésére is. Erre akkor is szükségünk lehet, ha a levél olyan információt tartalmaz, amelyhez a gépi felolvasás korlátai miatt nem tudunk hozzájutni. Ilyenek például a képek, a hosszú szövegek és az ábrák. A hangban történő választ egy hangfájl formájában rögzíti a rendszer és azt küldi vissza válaszként.

Egyéni paraméterek beállítása. Az ember-gép kapcsolat minőségét lehet javítani, ha az alkalmazás lehetőséget nyújt személyes jellegű beállítások használatára. Ebben a rendszerben férfi és női hang közül lehet választani a felolvasáshoz, és a beszéd paraméterei is hangolhatók (például, beszédsebesség, hangmagasság, tagolás stb.). Lényeges szolgáltatás, hogy a felhasználó választhat a kezdő, átlagos és tapasztalt szinthez illesztett menüstruktúra között is.

Ez a magyar nyelvű e-mail felolvasórendszer működött 1999-től a Westel Telefontársaságnál (Németh et al. 2000). A szöveg felolvasását a ProfiVox első változata látta el.

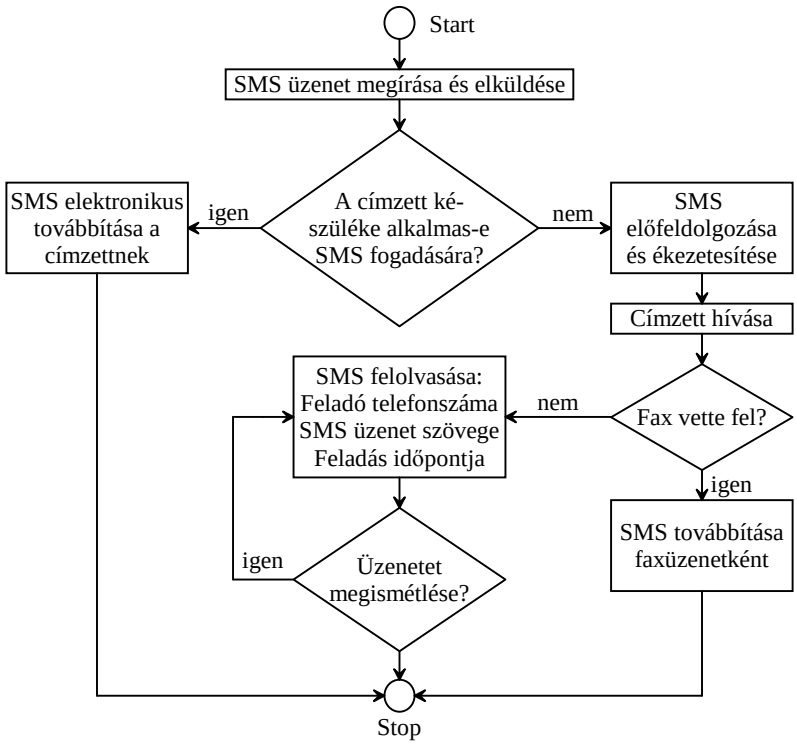
12.3.2. SMS-felolvasó vezetékes telefonra

Zainkó Csaba–Németh Géza

Egy tetszőleges vezetékes telefonra mindaddig nem lehetett SMS-t küldeni, amíg meg nem született egy olyan központi szövegfelolvasó alkalmazás, amelyik ezt lehetővé tette (Zainkó–Németh 2002). Ugyan elérhetőek speciális telefonok, amelyek képesek fogadni és küldeni SMS-t vezetékes vonalakon keresztül, de sok esetben a feladó nem tudja, hogy a címzettnek milyen készüléke van. Fel kellett tehát készíteni a rendszert arra, hogy nem speciális telefonokra is kézbesíteni lehessen az SMS-t. Ez az alkalmazás a Magyar Telekom és más szolgáltatók által üzemeltetett vonalas készülékekre tud SMS szöveget hang formájában eljuttatni. Az SMS-felolvasó egy szerveralapú rendszer, ami telefonon hívja fel a címzettet és az SMS-t szintetizált beszéddel felolvassa vagy rámondja az üzenetrögzítőjére. Ez a szolgáltatás kiszélesíti az SMS célközönségét. A szöveget a ProfiVox beszédszintetizátor alakítja át beszéddé. Mobilkészülékre is kiterjesztették, erről a megoldásról a 12.3.3. fejezetben szólunk. A rendszer működését a 12.3. ábra mutatja.

Az SMS továbbításának menete.

1. Az üzenet küldője megírja az SMS-t, és elküldi a címzett hagyományos, vonalas készülékének telefonszámára (a körzetszámot is megadva).
2. Az SMS-továbbító rendszer ellenőrzi, hogy a címzett előfizetői számán milyen készülék található. Ha a készülék nem képes az SMS elektronikus fogadására (a kijelzőjén megjeleníteni, mint például a hagyományos vezetékes telefon), akkor a rendszer a kézbesítendő üzenetet átadja a SMS-felolvasórendszer részére.



12.3. ábra. Az SMS-felolvasó működésének folyamatábrája

3. Az SMS-felolvasó elvégzi az előfeldolgozási lépéseket, helyreállítja az ékezeteket, amennyiben szükséges.
4. Felhívja a címzett telefonszámát és ha felvették a kagylót, akkor szóban közli, hogy hangos SMS-üzenet érkezett. Felolvassa az üzenet feladójának telefonszámát, ezt követően pedig az SMS szövegét mondja el, majd beolvassa az SMS feladási idejét is.
5. A felhasználó gombnyomásra meg tudja ismételtetni az üzenet és a hozzá kapcsolódó információk felolvasását.
6. Abban az esetben, ha faxkészülék van a vonal végén, akkor nincs szükség beszédszintézisre, a felolvasó fax formájában továbbítja az üzenetet és utána bontja a vonalat.

A szolgáltatás mobiltelefonra is kiterjeszthető (mobilról mobilra) Egy lehetséges megoldás, hogy a feladó az üzenet elejére helyezett speciális kóddal jelzi, hogy az SMS-t mindenképpen fel kell olvasni a címzett számára. Az SMS-felolvasóban háromféle felolvasási eljárást alkalmazunk:

Tárolt beszéd. A gép a rendszerüzeneteket természetes beszéddel mondja el, amelyet egy bemondó felolvasásból rögzítünk.
Szövegfelolvasó szoftver. Az SMS előfeldolgozott szövegét alakítja át szintetizált beszéddé.
Speciális, számfelolvasó beszédszintetizátor. Telefonszámot, dátumot és időpontot a természetes ejtéshez közel álló hangminőséggel olvas fel. Mindhárom technológia ugyanannak a személynek a hangjából készült hangelembázist használja.
Telefonszámok és időpontok felolvasása. A számok, dátumok és időpontok felolvasásáról a 10.2.4. fejezetben ismertetett rendszer gondoskodik. A telefonszámoknál fontos a megfelelő tagolás. Sok, Magyarországon használt telefonos szolgáltatásnál az angol-amerikai típusú számjegyenkénti felolvasást alkalmazzák, amely teljesen idegen a hazai felhasználók számára, és nagyban nehezíti a telefonszám megértését, lejegyzését. Mivel a mobiltelefonok a telefonszámok kijelzésekor a nemzetközi formátumot használják, ahol a telefonszám az adott ország nemzetközi hívószámával kezdődik, az SMS-felolvasóban is ehhez hasonlóan olvassuk fel a telefonszámokat. Valamennyi magyarországi telefonszám tehát 36-tal kezdődik, és a körzetszámmal folytatódik. A körzetszám elhangzása után szünetet tart a rendszer, jelezvén, hogy a körzetszám hangzott el (a beszéddallamot is lebegőre állítja ilyenkor). Innen kezdve következik az előfizető helyi hívószáma, aminek a tagolását a magyar szokásoknak megfelelően végzi el a rendszer. Ez attól függ, hogy hány számjegyű a telefonszám. 7 számjegy esetén a 3-2-2 -es, 6 számjegy esetén a 3-3 -as tagolást alkalmazzuk.
Az SMS-ek előfeldolgozásának problémái. Az SMS-ekben gyakran használnak ötletes rövidítéseket. Lényeges kérdés, hogy magát a küldő által leírt karaktersorozatot kell-e felolvasni vagy a hozzárendelhető szöveget (például: elsajátít6, 5letes, +lehetősen, 6ás, 2séges). A probléma nyilvánvaló: igazodni kellene az SMS szövegek rövidítési stílusához. Olyan megoldást dolgoztunk ki, hogy nagy valószínűséggel megállapítható legyen a szöveg helyes értelmezése. Például rövidítésnek tekintjük azokat a karaktersorozatokat amelyek pontra végződnek, és a rövidítésszótárban szerepelnek. Aminek nem találjuk meg a feloldását, azt karakterenként, betűzve olvassuk fel. Olyan esetekben, ha például a szövegben egy kettőspont és befejező zárójel szerepel :), ami az elektronikus levelezésben és társalgásban bevett hangulatkifejező elem, akkor ezt nem „kettőspont zárójel”-ként olvassuk fel, hanem a szöveg az lesz: szmájli. Az üzenetek feldolgozásakor azt is meg kell állapítani, hogy az üzenet írója milyen karaktert vagy eljárást használt a szóközök jelölésére. A számok feloldásakor meghatározzuk, hogy rövidítésre használta-e az SMS írója, vagyis hogy az előtte, illetve mögötte álló betűkkel egybe van-e írva. Ha pont szerepel a szám mögött, akkor sorszámnévnek tekintjük, kivéve, ha a pont egyben a mondat végét is jelöli, vagy az üzenet írója ezt használta szóközkarakter helyett. Az ékezetesítés az előfeldolgozás utolsó lépése. Az SMS-ek szövegének előfeldolgozása után olyan betűsorozat áll rendelkezésünkre, ami már felolvastatható a géppel és viszonylag kevés kiejtési hibát fog tartalmazni. Ennél a pontnál jegyezzük meg, hogy ha hangos

felolvastatásra írunk SMS-t, akkor a fenti stílust mellőzzük, célszerű a helyesírási szabályokat betartani.

12.3.3. Mobiltelefonba épített SMS-felolvasó

Tóth Bálint–Németh Géza–Kiss Géza

Az SMS-alapú kommunikáció igen kedvelt forma. Az SMS-felolvasó szolgáltatás akkor hasznos, ha az üzenetek olvasása adott szituációkban nehézséget okoz. Ilyen lehet például a gépkocsivezetés, erős napsütés stb. Ekkor a mobiltelefonba beépített beszédszintetizátor felolvassa a szöveget. Vak és gyengén látó felhasználók számára is nélkülözhetelen egy ilyen szolgáltatás. Tudomásunk szerint a világon elsőként a BME-TMIT és a M.I.T. Sytems közös munkájaként, 2003-ban készült el egy SMS-felolvasó műszaki megoldás Symbian alapú okostelefonra, melynek a neve SMS-Mondó (SMSrapper) lett (Zainkó–Németh 2002). Az SMSMondót hazánkban a T-Mobile forgalmazta (Németh et al. 2008). A műszaki teljesítményt a beszédszintetizátor mobiltelefonba való beépítése és illesztése jelentette abban az időben.

Az alkalmazás testreszabása. Az alkalmazás célfunkciója, hogy a telefon beépített SMS kliense helyett az SMSMondó alkalmazás fogadja az üzeneteket, és a felhasználói beállításoknak megfelelően felolvassa azokat. Az alkalmazásban a felhasználók a telefon profiljaihoz (általános, tárgyalás, utcai, csendes üzemmód) külön-külön rendelhetnek egyedi beállításokat. A következő paraméterek beállítására van lehetőség: megadhatja a beszéd hangerejét, a beszéd sebességét, ezentúl, hogy hányszor olvassa fel a bejövő SMS-eket, illetve hogy az üzenet mely paramétereit olvassa fel az SMSMondó (például az üzenet feladója, elküldés dátuma, ideje stb.). Az alkalmazásban a felhasználóknak lehetősége nyílik nyelvdetektálás bekapcsolására is, mely segítségével a program automatikusan érzékeli, ha idegen nyelvű üzenet érkezik. Ekkor megkérdezi, hogy felolvassa-e. További testreszabási lehetőség, hogy a felhasználó megadhat olyan időintervallumot, mely során nem szeretné, hogy az alkalmazás felolvassa a bejövő üzeneteket (például éjszaka alvás közben). Ezentúl meg lehet adni az alkalmazásban, hogy bejövő üzenet esetén az alkalmazás előtérbe kerüljön-e. Ez abban az esetben lehet fontos, ha a meghallgatáson kívül el is szeretnénk olvasni a szöveges üzenetet. Az alkalmazás magyar és német nyelvre készült el (a ProfiVox technológia felhasználásával). Az alkalmazás a telefon hangszóróján túl képes az SMS-eket Bluetooth kapcsolat segítségével Bluetooth-os fejhallgatón, illetve a kocsik Bluetooth-os audiorendszerén keresztül is felolvasni.

Az alkalmazás működése. Bejövő üzenet esetén az SMSMondó azonosítja a bejövő telefonszámot. Amennyiben a telefonszám szerepel a névjegyzékben, akkor a telefonszámhoz tartozó nevet, amennyiben nem szerepel, akkor a telefonszámot olvassa fel. Ezek után bekapcsolt nyelvdetekció esetén az alkalmazás azonosítja az üzenet

nyelvét és a korábban leírtaknak megfelelően jár el. Az alkalmazásnak egy továbbfejlesztett változatában megadhatunk telefonszámlistákat. A bizalmas lista esetében az SMS-eket nem olvassa fel automatikusan a program és nem is értesíti róla hangosan a felhasználót. Ez a funkció hasznos például bizalmas üzleti üzenetek esetén, amikor mindenképp el szeretnénk kerülni annak a lehetőségét, hogy esetleg a környezetünk is meghallgassa az üzenet tartalmát. Az úgynevezett üzleti lista esetében a gép elmondja a küldő nevét / telefonszámát és a küldés dátumát, azonban magát az üzenet szövegét nem olvassa fel.

12.3.4. Automatikus szám szerinti tudakozó

Németh Géza–Zainkó Csaba

2006-ban merült fel az a kérdés a T-Mobile Magyarország mobilszolgáltatónál, hogy lehetne-e teljesen gépesíteni a telefonszám szerinti tudakozót beszédtechnológiai fejlesztéssel. A választ a BME TMIT beszédtechnológiai laboratóriuma adta meg: igen (Németh et al. 2007b). A bemenet egyértelműen lekezelhető, ha egy nyomógombos telefonról adják meg a telefonszámot, a válasz generálására kell a gépi beszédelőállítás, amely speciális megoldásban megfelelő hangminőséget tud biztosítani ahhoz, hogy az előfizető megérthesse a keresett adatot, és elégedett legyen a szolgáltatással. A téma szűkebben arra vonatkozik, hogy neveket (személy- és cég), valamint címeket, esetleg telefonszámokat kell a gépnek felolvasni. A szokásos szövegfelolvasásnál ez nagyságrenddel nehezebb és komplexebb feladat (Németh et al. 2003). A munkához el kellett végezni mintegy 3 millió magyar előfizető személy és cégnév analízisét. A nevek elemeit adatbázisba kellett szervezni, olyan formában, hogy bővíthető is legyen, hiszen az előfizetők száma nőni fog. A kategorizálásnál már figyelemmel kellett lenni a későbbi felolvasásra is, gondolunk itt a megfelelő tagolásra, hangsúlyozásra és dallammenetre. Néhány adatot adunk közre a nevekkel kapcsolatosan. Mitegy 104 000 különböző családnév szerepelt a végső adatbázisban. A nevekhez kapcsolódó egyéb jelek száma 82 volt (betűk, titulusok és azok elhelyezkedése a névhez viszonyítva (*dr. Tóth; Tóth dr.; Dr. M. Tóth, ; Dr. Tóth M.; dr. Tóth nyug. Tóth dr. tans.* stb.). Az utónevek száma elérte az 1800-at. Az utónevek kezelésénél figyelemmel kellett lenni a dupla névadásra is. A cégnevek esetében sokkal szélesebb a paletta, mint a személyneveknél. Több tízezer cégnév kiejtési formáját kellett megadni szabályok és kiejtésikivétel-listák alkalmazásával.

*Beszédformák.*A gépi felolvasás biztos megértésére is gondolni kellett. Nem azért, mert a gépi felolvasó érthetetlenül beszél, hanem azért, mert a nevek esetében rövid, egy szótagos, egyetlen hangot tartalmazó elemek is szerepelhetnek (például *Hó Alíz; Mász Vajk*), amiknek a megértése egyébként sem könnyű, de ezt még nehezítheti a telefon torzítása is. Ezért terveztek a folyamatos kiejtésen felül szótagolási funkciót

és betűzéses formát is. Ha a hívó fél a folyamatos felolvasásból nem tudja megérteni egyértelműen az előfizető nevét, akkor folyamodhat a részletező felolvasási módok egyikéhez. Ehhez a munkához kapcsolódik a szerzők azon törekvése is, hogy ajánlatot tettek a magyar betűzési forma szabványos megoldására is (11.1. táblázat). E munka keretében született meg az első magyar szótagoló hangautomata is.
A beszéd-előállítás megoldása. A célnak megfelelően olyan megoldást kellett találni, amelyik a legérthetőbb hangminőséget biztosítja. Ebből következik, hogy nem lehetett a feladatot egy szövegfelolvasóra bízni, mivel az túl gépies hangzást adna, a rövid, egy-két szótagos elemek felolvasásánál a hangminősége sem lenne optimális. A fejlesztők hibrid megoldást gondoltak ki. Ebben alkalmaztak egyrészről kötött szótáras szintetizátort (10.2. fejezet), amely előre felevett elemekből dolgozik, szövegfelolvasót (10.3.6.1. fejezet) és számfelolvasót (10.2.4. fejezet). Mindhárom modult ugyanazon személy hangjára tervezték meg, és készítették el.
A felolvasási stratégia. Az előre felvett elemek a következők voltak: fix üzentek, tájékoztatások a dialógusban, valamint az utónevek, a cégformák, a városok, a leggyakoribb utcanevek, a közterületek elnevezései és a betűzés szavai. Ezekben a felvételekben minden tételt úgynevezett vivőmondatban olvasott fel a bemondó, hogy a későbbi összefűzésnél a beszéd prozódiája optimális maradjon. A többi felolvasandó üzenetet, közlést szövegfelolvasóval oldották meg. Számfelolvasót használtak az irányítószámok, valamint a telefonszámok és a házszámok felolvasására. A három beszédszintetizátor hangját gondosan összeillesztették (már a hangfelvételeknél is figyeltek erre), tehát a hallgató nem érzi a kapcsolódási pontokon, hogy váltják egymást a technológiák. A megoldás eredménye az, hogy úgy tűnik a felhasználó számára, mint ha természetes személy mondaná be az előfizető adatait akkurátusan, magyarázó formában. A rendszer üzembeállítása után a megrendelő meghallgatásos tesztet is végzett kis létszámú, azonban szigorú belső munkatársakkal. Az eredmény egy 5 fokozatú skálán a következő volt: a beszéd érthetősége és természetessége 4,58; beszéd tempója 4,75; a közölt információ korrektsége 5,0; a dialógus menete és korrektsége 4,58. Ezt a rendszert 2008-ban a Vodafone mobilszolgáltatónál is bevezették.

12.3.5. Gyógyszervonal, automatikus telefonos információs rendszer

Olaszy Gábor–Bartalis Mátyás

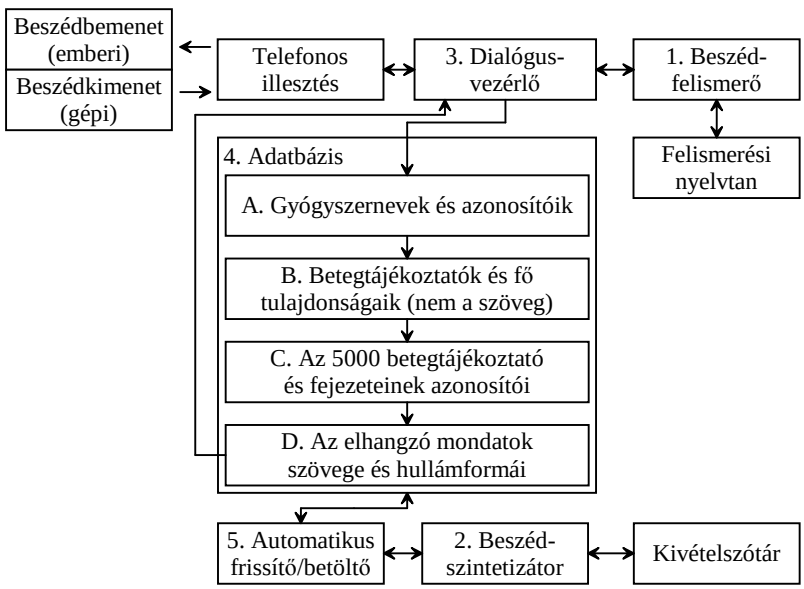
A széles körű lakossági tájékozódás korszerű támogatására ad példát az alább ismertetett gyógyszerinformációs rendszer (Németh et al. 2007a). A szolgáltatás kicsit a jövőbe mutató, ötvözi a gépi beszédfelismerés és a beszédszintézis módszerét egy korlátozott humán-gép (beszéddialógussal is megvalósított) gyógyszer-információs rendszerben. A célkitűzés az volt, hogy olyan információs szolgáltatást fejlesszenek,

amelyik magyar nyelven képes a hivatalosan forgalomban lévő gyógyszerekről a legalapvetőbb tájékoztatást megadni, vagyis a gyógyszer dobozában elhelyezett nyomtatott betegtájékoztató szövegét hozzáférhetővé tenni a társadalom bármely tagja számára helytől és időtől függetlenül. A rendszer interneten, WAP-on és telefonon is ad információt. Ezzel a három hozzáférési lehetőséggel a lakosság teljes körének módja van hozzájutni az általa keresett gyógyszer adataihoz, jellemzőihez. Ilyen széles körű adatelérési lehetőség 2006-ig nem állt rendelkezésre (tudomásunk szerint külföldön sem). A rendszert az Országos Gyógyszerészeti Intézet (OGYI) működteti, rendszeresen frissítve az adatokat. A *Gyógyszervonal* lakossági információs rendszerrel sok esetben megelőzhető a téves gyógyszerhasználat, lehetőséget teremt a társadalmilag hátrányos helyzetű csoportok (például idősek, sérült emberek, kórházban fekvők) számára is a betegtájékoztatókban leírt, fontos információk megszerzésére. A közegészségügyi információs ellátottság egyik fontos eleme lehet az ilyen típusú szolgáltatás. A megoldás interdiszciplináris együttműködés eredménye, és alapvetően beszédtechnológiai alapelemeket használ kiindulásként. A fejlesztés a BME Távközlési és Médiainformatikai Tanszéke (TMIT) és az OGYI közös munkája, amelyet 2005-ben elnyert GVOP-pályázat támogatott (Olaszy–Haraszti 2007). Magyarországon átlagosan ötezer, különböző kiszerelésű OGYI által törzskönyvezett gyógyszer van, mindegyikhez kiadnak egy nyomtatott, szöveges betegtájékoztatót. Évente mintegy 400 új gyógyszer jelenik meg, és hozzávetőlegesen ugyanennyit vonnak ki a forgalomból. A *Gyógyszervonal* információs rendszer mind a szakembereknek, mind a lakosságnak segítség a betegtájékoztatók szövegéhez történő szabad és gyors hozzáférésben.

A rendszer felépítése. A *Gyógyszervonal* telefonos információs rendszer telefonos formája, valamint az internetes és WAP-os megoldása ugyanazt a közös, egységesített szerkezetű adatbázist használja. A telefonos rendszer előnye, hogy nem igényel semmiféle speciális tudást és infrastruktúrát, az információt hangban kapja meg az érdeklődő, akár többször is meghallgathatja, mondatonként is ismételheti a meghallgatást, lassabb-gyorsabb beszédtempót is kérhet a géptől stb. Az internetes rendszer előnye pedig az, hogy hatásosan lehet keresni a gyógyszeradatbázis írott anyagaiban. Az internetes megjelenítés vakbarát, a fejlesztők hangsúlyozott célja volt a látássérültek igényeinek kielégítése is. Mindkét hozzáférés más-más lakossági rétegeket érint. A rendszer elérhetőségei: `http://www.gyogyszervonal.hu`, telefonon: 06-1-886-94-90.

A telefonos működési mód. A rendszer tervezői célul tűzték ki a természetes beszédkommunikáció géppel támogatott formájának a megvalósítását, *ez a fejlesztés fő újdonsága*. Felhívjuk a telefonszámot, és emberi hangon „beszélgetést folytatunk” a géppel (erősen korlátozva a dialógust a gyógyszerekkel kapcsolatos kérdésekre). A gép kérdez, tájékoztat, az ember pedig válaszol és megítéli a gép által elmondott információt. A telefonos kapcsolattal 30 másodpercen belül hozzá tudunk férni a

keresett gyógyszerhez kiadott tájékoztatóhoz (hangban). A „géppel” történő beszélgetés megvalósításához számos nyelvi, fonetikai, akusztikai, beszédtechnológiai és műszaki problémát kellett megoldani. Beszédszintetizátor szükséges ahhoz, hogy a hívó fél részére az információs rendszer a betegtájékoztató szövegét felolvassa, a beszélő fél hangjának feldolgozásához pedig gépi beszédfelismerőt kell alkalmazni. Mindkettő egy-egy beszédtechnológiai modulként működik az információs rendszerben (12.4. ábra). A hívó fél az adatmegadáshoz két forma közül választhat. Beszéd-



12.4. ábra. A „Gyógyszervonal” telefonos információs rendszer felépítése

del megadja a gyógyszer nevét, ilyenkor a dialógusvezérlő az 1-es beszédfelismerő bemenetére irányítja a jelet, avagy kéri a gyógyszer törzsszámát (4-jegyű szám, ez minden gyógyszerdobozon megtalálható), amit a telefon nyomógombjaival lehet megadni. A párbeszédet a hívó fél és a gép között a 3-as dialógusvezérlő irányítja. A kétoldalú beszédkommunikációt barátságos és egyszerű dialógusrendszer biztosítja. A bejelentkezéskor a gép üdvözli a hívót. Ha az ügyfél először hívja a rendszert, akkor részletes információt is kérhet. Ha a hívó gyakrabban használja a rendszert, akkor a bejelentkező szöveg átugorható. Ebben az esetben azonnal a gyógyszer nevét kéri a gép. A párbeszéd során a hívó fél a kiválasztott gyógyszerhez tartozó minden fontos adatot meghallgathat. A dialógusvezérlő mondatait kellemes hangszínezetű női hang olvassa fel, nyugodt tempóban. A 4-es modul az adattár, amelyik a gyógyszerek törzskönyvi számát és a hozzájuk rendelt gyógyszerneveket tartalmazza. Itt található a gyógyszerek nevéhez tartozó betegtájékoztatók szövege, az OGYI által jóváhagyott formában. Hatféle információt kérhet az ügyfél: általános tájékoztató;

szedés előtti tudnivalók; alkalmazási javaslatok; mellékhatások; a készítmény tárolása; egyéb tudnivalók. 5. Ez az adattár a beszédszintetizátornak előkészített kódolt formátumú szövegeket tartalmazza. 6. A beszédszintetizátor olvassa fel a 4. adattár szövegeit. A szintetizátor a BME TMIT ProfiVox beszédszintetizátorából (Olaszy et al. 2000b) kialakított, speciálisan gyógyszerinformációs szövegek meghangosítására fejlesztett szoftver (Olaszy et al. 2007), amely férfi hangon szól. Több ezer, főként latin szó kiejtését kellett meghatározni és beépíteni a kivételszótárba, hogy a felolvasások elfogadhatóan hangozzanak mind a szakemberek, mind a gyógyszert felhasználók számára. Ez a szótár nyitott, hiszen az előforduló újabb szavak kiejtését is meg kell határozni. A betegtájékoztatók szövegének fogalmazási stílusa, néha a központozása is, eltér az általános szövegekben használtaktól. Ezért az itt megkívánt felolvasási stratégiához sajátos szabályokat kellett kialakítani. A beszédfelismerő modul egy beszélőtől független, mintaillesztésen alapuló megoldás (Fegyó et al. 2003). Az ügyfél beszédének akusztikai jeleiből kiszámított paramétermezőt hasonlítja össze az adattárában eltárolt sok ezer minta paramétereivel, és kiválasztja a leghasonlóbbat. A felismerő modul tehát csak akkor működik jól, ha az ügyfél olyan gyógyszernevet mond be, amelyik szerepel a rendszer 5000-es adattárában. A rendszer a bemondotthoz leghasonlóbb ejtésű gyógyszernévre dönt, az eddigi tapasztalatok szerint megközelítőleg 92%-os pontossággal. A beszédfelismerő különlegessége, hogy új felismerendő elem hozzáadása esetén (új gyógyszer jelenik meg új névvel) nem kell tanítást végezni. Speciális adatábrázolási forma teszi lehetővé, hogy fonetikai szimbólumok segítségével megadható az új felismerendő elem gép számára érthető paraméterrendszere. A beszédfelismerő szótára éppúgy nyitott, mint a beszédszintetizátoré. Erre szükség is van, hiszen a változásokat folyamatosan be kell vezetni a rendszerbe.

Internetes és WAP-os elérés. A Gyógyszervonal internetes honlapjának használatakor a felhasználó kereshet az adatbázisban a készítmény neve, hatóanyagtartalma, illetve törzskönyvi száma alapján. A honlap vakbarát szerkezetű. A webes felület fejlesztésénél is figyelembe vettük a Magyar Vakok és Gyengénlátók Szövetsége (MVGYOSZ) javaslatait, hogy a mai képernyőolvasókkal gond nélkül tudjon együttműködni a rendszer. Ezen felül figyelembe vettük a W3CD ajánlásait az akadálymentesített weboldalak készítéséhez. A legfontosabb ilyen kérésekről alább szólunk. Fontos, hogy a rendszerről szóló általános információs rész, mint telefonszám vagy internetes elérés címe, ne az oldal elején legyen, mert ekkor minden alkalommal, amikor meglátogatják az oldalt a képernyőolvasó felolvassa ezt, ami hosszadalmassá teszi az oldal használatát, valamint a sokadik látogatás alkalmával már nem nyújt érdemleges információt. Akadálymentesített weboldalak esetén kötelező a stíluslapok használata, tehát maga a HTML kód nem tartalmazhat semmilyen formázásra (betűtípus, betűméret...) vonatkozó jelölést, mert ez megzavarhatja a képernyőolvasók működését. Helyette a stíluslapokban definiálni kell a megfelelő formázásokat és arra kell hivatkozni a kódban. Képek esetében mindenképp meg kell adni az ALT

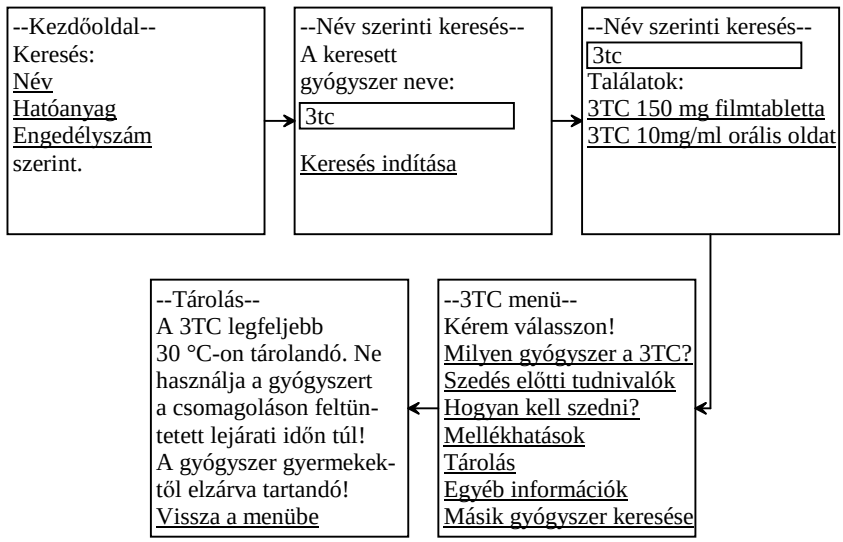
címkét, mely egy rövid leírást tartalmaz a kép tartalmáról, mivel a képernyőolvasók nem képesek a képek tartalmát másképpen értelmezni. Akadálymentesített weboldalak esetében táblázatokat csak indokolt esetben lehet használni, mivel ezeket a képernyőolvasó mindig úgy olvassa fel, hogy az egyes mezőkhöz bemondja a hozzá tartozó oszlop fejlécében található meghatározást is. Valamint táblázatok esetében is ki kell mindig tölteni az ALT címkét, mely egy rövid összefoglalása a táblázatnak Mi történjen azokkal a gyógyszerekkel, amelyekhez nincs betegtájékoztató? A program első verziójában először egy hibaüzenetet tartalmazó oldal jelent meg ebben az esetben. Az MVGYOSZ kérésére ezt megváltoztattuk úgy, hogy azok a készítmények, melyekhez nem tartozik betegtájékoztató, ne legyenek linkek, így nem keltik azt az illúziót, hogy további információhoz juthat a link követésével a hasznló személy. WAP-os felületnél figyelembe kellett venni, hogy a mobiltelefonok kijelzőjén nem jeleníthető meg bármekkora mennyiségű szöveg. Korlátozott memória és egyéb okok miatt a keresési találati listák hosszát és adott esetben a betegtájékoztatók tartalmát is ezek figyelembevételével kellett darabolni, valamint meg kellett oldani, hogy egyszerűen lehessen navigálni a töredékek között. WAP-os hozzáférés esetén a készítmény neve, fő hatóanyaga és a törzskönyvi száma szerint is lehet keresni. Mivel mobiltelefonokon általában nincs teljes értékű billentyűzet, ezért itt is megengedett szótöredékek megadása a keresési feltételeknél, így az is könnyedén használhatja a mobiljáról a rendszert, aki esetleg nem annyira gyakorlott a készülék használatában. Alább megtekinthető, hogy a „3TC” nevű készítmény tárolására vonatkozó információk keresése közben a mobiltelefonon milyen megjelenő tartalmakat lehet látni (12.5. ábra).

12.3.6. Automatikus, mobiltelefonos, helyfüggő kereső szolgáltatás

Fegyó Tibor

A helyfüggő szolgáltatás olyan speciális információt ad (emberi kezelő nélkül), ami a földrajzi helyzethez igazodik. Megtudhatjuk például a helyi időjárást, megismerhetjük a környék látnivalóit, programajánlatokat kérhetünk, vagy megkereshetjük a legközelebbi gyógyszertárat. Egy ilyen helyfüggő szolgáltatás például a T-Mobile úgynevezett Célravezető szolgáltatása (Location Based Service, LBS), mely természetes beszéddialógussal működik. Segítségével a mobilfelhasználók lekérdezhetik a közelükben lévő benzinkutak vagy mozik listáját (beszéddel), de akár helyi időjárásjelentést is kérhetnek. Minden információt gép olvas fel. A fejlesztés egy megvalósíthatósági tanulmány keretein belül történt, ahol megvizsgálatuk, hogy beszédalapú, telefonos környezetben hogyan működne egy ilyen szolgáltatás.

A rendszer felépítése. LBS szervernek nevezzük a helyfüggő szolgáltatásokat nyújtó szerveralkalmazást. Egy LBS szerver tipikusan többféle funkciót is megvalósít, töb-



12.5. ábra. A „Gyógyszervonal” WAP-os változatának képernyő képei

bek között lehetővé teszi egy mobilfelhasználó földrajzi helyzetének lekérdezését, valamint adott földrajzi koordináta környezetében lévő objektumok (például benzinkút, mozi, étterem, időjárás, bank) kategorizált megkeresését. Előfordul, hogy az LBS szerver nem konkrét információt tartalmaz az adott keresési kategóriáról, hanem csak egy arra mutató linket, például az időjárás esetében.

Az LBS szerverrel egy LBS kliens alkalmazás tartja a kapcsolatot. A szerverrel való kapcsolattartás jelen esetben egy HTTPS/XML API-n keresztül történik, aszinkron módon, ezalatt azt értjük, hogy nem szükséges megvárni például a pozíció meghatározásának a végét, az érdeklődési kör kérelmet közben is el lehet küldeni. A végeredményt természetesen csak a két kérelem feldolgozása után lehet lekérdezni, de az aszinkron megoldással gyorsabban jutunk az információhoz. A kliens modul kezeli azt az esetet, amikor a keresés sikertelen volt, például tágabb környezetben keres, újra kérdez stb. Sikeres keresés esetén, amennyiben a visszaadott objektumok rendelkeznek pontos koordinátainformációkkal, automatikusan sorba rendezi a találatokat a felhasználótól való távolság szerint. A felhasználóval történő beszélgetés menetének vezérlését egy dialógusvezérlő irányítja. A teljes dialógusfolyamatot úgynevezett dialógusfákból lehet felépíteni, ezek határozzák meg, hogy az automata mikor mit mondjon a felhasználónak, mit kérdezzen tőle, illetve milyen egyéb tevékenységeket hajtson végre. A dialógusfákat néhány előre definiált elemi dialógus segítségével építhetjük fel. Minden ilyen elemi dialógus valamilyen egyszerű művelet végrehajtását teszi lehetővé, mint például számítások végzése, felhasználó informálása, információ bekérése, dialógus befejezése vagy másik dialógusfa meg-

hívása. Egy dialógusfa tehát egy olyan fa, melynek csomópontjaiban elemi dialógusok állnak, az ágakat pedig úgynevezett címkék alkotják. A címkék segítségével dönti el a rendszer, hogy milyen esetben a fa melyik ágán kell folytatni a vezérlést. A dialógusrendszer természetesen egyszerre több felhasználót is ki tud szolgálni, a dialógusok végrehajtása párhuzamosan történik. Akár arra is van lehetőség, hogy bizonyos feltételektől függően különböző felhasználók esetén különböző típusú dialógusok kerüljenek végrehajtásra, így akár személyre is szabható a hangos célravezető szolgáltatás. A kereshető objektumok nevét kérésre a rendszer felsorolja. A felhasználó által bemondott kifejezéseket a beszédfelismerő motor dolgozza fel. A dialógus egyes csomópontjaiban külön-külön meghatároztuk, hogy mely kifejezésket kell felismerni, és a gyakorlatban előforduló alternatív kifejezéseket is rögzítettük (például *bank, pénzkiadó automata, ATM, bankautomata, pénzfelvevő*). Az ismertetett LBS szolgáltatásban használt kategóriák száma meglehetősen alacsony, így a csomópontonként felismerendő kifejezések száma tipikusan kevesebb, mint 100 elem. Száz, vagy néhány száz kifejezést tartalmazó egyszerű párhuzamos ágakból álló nyelvi modell esetén a felismerő motor a valós idő töredéke alatt képes meghatározni az optimális illesztési útvonalat, így egy mai PC 10–20 csatorna egyidejű felismerését képes kiszolgálni. Fontos a kiejtésikivétel-szótár megfelelő feltöltése (8.3. fejezet). A telefoninterfész-modul feladata a telefonvonalon az adatok és a jelzésrendszer kezelése. Jelen esetben PRI ISDN típusú vonalon fogadjuk a bejövő hívásokat, ehhez speciális vonali illesztő PC kártyára, valamint ehhez illeszkedő programra volt szükség. Ez a program teremt kapcsolatot a dialógusrendszer és a telefonvonal között, felveszi és bontja a hívást, lejátssza, illetve rögzíti a beszédet. A modul további feladata, hogy a magasabb szintű alkalmazások felé visszaadja a hívó MSISDN számát is, ugyanis a felhasználó helyzetét ez alapján lehet meghatározni, letiltott hívószám-kijelzés esetén a szolgáltatás nem elérhető.

A rendszer működése. A hívás elején a felhasználó hívószámának azonosítása után a rendszer pozíciómeghatározást végez a dialógus futásával párhuzamosan. A kommunikáció a felhasználó üdvözlésével és tájékoztatásával kezdődik. A főmenüben lehetőség van beszédfelismerés és nyomógombok használatára is. Ha az automata másodjára sem érti meg, amit mondtunk, vagy a felhasználó másodjára sem mond be semmit, akkor a DTMF menüre történő átváltás automatikusan megtörténik. Amennyiben beszédfelismeréssel vagy gyorsbillentyű megnyomásával sikerült kiválasztani, hogy mit akarunk lekérdezni, a rendszer elindítja a keresést az érdeklődésnek megfelelően (például az étterem bemondásra éttermeket keres a környéken). Az automata ellenőrzésként visszamondja az érdeklődési kört, és ekkor még lehetőség nyílik a visszalépésre, módosításra. Ha megvan a keresés eredménye, akkor a rendszer beszédszintetizátora elkezdi felsorolni a találatokat (speciális felolvasó kell, nevekre, címekre, telefonszámokra, vö. a 12.3.4. fejezettel). Az eredményt a gép kérésre többször is elismétli. Lehetőség van egy kiválasztott, konkrét eredmény adatainak SMS-ben történő elküldésére is.

12.3.7. Automatikus áru- és árlista-felolvasó

Zainkó Csaba–Bartalis Mátyás–Németh Géza

Nagy kihívás, ha egy meghatározott személy hangjára kell beszédszintetizátort tervezni (például vállalati megkeresésre), olyant, aminek beszéde összetéveszthető az emberi bemondással, tehát azzal egyenértékű. E könyv tartalmának ismeretében azonnal adódik a válasz, hogy csak igen szűk tématerületre lehet ezt elvállalni. Az ilyen fejlesztések elindításához kétféle adatállománnyal lehet számolni. A) Vagy korábban rögzített hangfelvételek állnak rendelkezésre és a hozzájuk tartozó szövegek írott formában, nagy mennyiségben. B) Vagy új hangfelvételeket kell tervezni és ehhez a felolvasandó szövegeket is ki kell alakítani. Ebben a fejezetben egy olyan szolgáltatást mutatunk be, melyik nagy tömegű, korábbi hangfelvételek alapján elkészített beszédszintetizátorral segíti a T-Mobile Magyarország mobilszolgáltató munkáját (Németh et al. 2009). A konkrét megoldásban a rendszer beszédet generál előre megadott szövegformátumokból (mobiltelefonok és hozzájuk kapcsolódó termékek árlistája és a fizetési opciók megadása). A megoldás támogatja az illetékes operátort, hogy képes legyen a hetente közreadott mobiltelefon-árlisták beszédfájljait automatikusan előállítani kismértékű emberi beavatkozással, felügyelettel.

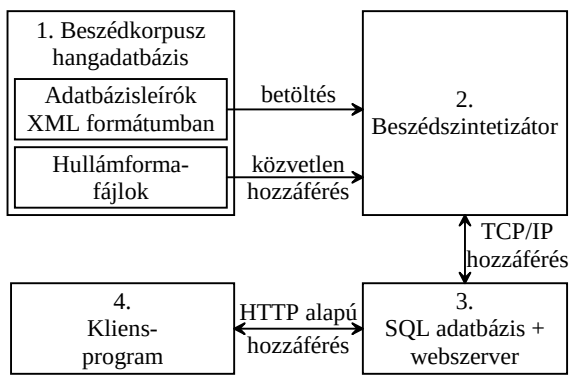
Motiváció. A mobilszolgáltató telefonos ügyfélszolgálati rendszerét üzemeltető és folyamatosan fejlesztő részlegnek sok munkát okozott, hogy minden héten a cég bemondójának (a teljes telefonos ügyfélszolgálatban az ő hangja szól) fel kellett olvasni az összes általuk forgalmazott készülék típusát és árát, valamint a hozzájuk tartozó különböző előfizetési opciókat, hogy azok telefonon is meghallgathatók legyenek. Ez heti 3–4 órányi hangfelvételt jelentett. Felmerült az igény, hogy ezt a monoton munkát gép végezze el, de legalábbis nagy részben segítse. A bemondó hangját kellett géppel létrehozni, ezért csak az elemkiválasztásos technológia jöhetett szóba (10.3.7. fejezet). A BME TMIT által javasolt és azóta rendszeresen használt szoftverrel ennek a feladatnak az elvégzése heti fél órára csökkent. Nagy előnye, hogy a bemondónak nem kell felolvasni a teljes szöveget. Ez azért is előnyös, mert abban az esetben, ha a bemondó megbetegszik, vagy szabadságon van, akkor is lehetőség van az aktuális árlista hangos formájának elkészítésére, melyre egy rövid betanítás után egy operátor képes. Csupán szerkesztési munkát kell végezni az esetek nagy részében. Bemondásra csak akkor van szükség, ha új gyártó új készüléke jelenik meg a cég kínálatában. Ez viszont előre tudható.

A fejlesztés. Első lépésben meg kellett vizsgálni az információs anyag tartalmi változásait. A kérdés az volt, hogy milyen változó részek vannak ezekben az információkban. Az elemzés kimutatta, hogy évente 2–3 új gyártó termékei jelennek meg a kínálatban, valamint 60–80 új készüléktípus (új név). Ez azt jelenti, hogy csupán ennyi új szövegelem kerül bele a szöveglistába éves szinten. A heti változó elemek tehát döntően az árak és a fizetési feltételek voltak. A 10.2.4. fejezetben

ismertetett megoldással ez algoritmizálható. A beszédszintetizátor fejlesztéséhez a kiindulási korpuszokat (beszéd és szöveg) a szolgáltató biztosította. Ez a bemondója által a fejlesztést megelőző 1 évben felolvasott hasonló tartalmú bemondásokat tartalmazta (5 és fél órányi beszéd, ugyanabban a stúdióban felvéve), valamint annak szöveges formáját (413 szövegblokk, ami 3431 mondat, mintegy 30 000 szó). Egy szövegblokk alatt egy adott gyártó készülékeit értjük, azonos előfizetési konstrukciók mellett, például előre fizetéses, illetve 1 vagy 2 éves hűségnyilatkozat aláírásával történő előfizetés vásárlása esetén. Egy-egy blokk átlagosan 10–12 mondatot tartalmazott. Példa egy blokkból:
Samsung GT- ES7350 1990 Ft, 24x2000 Ft-os részlettel
Samsung GT-ES8000 19 990 Ft, 12x5000 Ft-os részlettel
Samsung Ultra Touch S8300 19 950 Ft, 12x 4170 Ft-os részlettel
Samsung SGH-U900 1990 Ft, 24x2000 Ft-os részlettel, Ezt a korpuszt több szinten fel kellett dolgozni, hogy a szintetizátort ki tudja szolgálni. Mind a beszéd szintjén, mind a szövegén végeztünk kiegészítő munkákat (vö. 8.2.3. fejezet). A beszédet blokkonként kezeltük, egy-egy blokkot automatikus eljárással mondatokra bontottunk, majd a mondatokat elláttuk hang- és szóhatárcímkékkel. Ilyen méretű beszédanyag címkézését emberi munkával nem lehet elvégezni. A BME TMIT beszédfelismerő motorját használtuk úgynevezett kényszerített felismerési módban (9.5.3.2. fejezet). A hang-szöveg összerendelésre példa a következő sor (minden esetben a hang határozta meg a szöveget)
Samsung GT-ES7350 1990 Ft, 24x2000 Ft-os részlettel=számszung gété, ees-hetvenhárom, ötven, ezerkilencászkilencvenforint, huszonnétyszerkettőezerforin-tosrészlettel A sikeres felolvasáshoz kombinálni kellett a számfelolvasási és az elemkiválasztásos technológiát. Ezzel biztosítottuk, hogy bármilyen szám kimondására is alkalmas lett a rendszer ugyanazon a hangminőségen (a cég bemondójával felolvastattuk a a 8.2.1. fejezetben ismertetett számlistát). A legérzékenyebb része egy ilyen felolvasórendszernek a készüléknevek, cégnevek jó kiejtése (vö. 8.3. fejezet). Ezek helyes kiejtésére külön modult terveztünk, amelyben kivételszótárt és ezzel párhuzamosan szabályalapú megoldásokat helyeztünk el. A szótáras megoldást a gyártó és készülékek neveire például: *LG* - /el dzsí/, *Samsung* - /számszung/, *Shine* - /sájn/; a szabályalapút pedig általában a számsorozattal megadott konkrét típusnevek felolvasásakor. Ilyen például a Nokia XXXX alakban leírható készülékei, ahol X egy-egy számot jelöl. Ilyen esetekben, a Magyarországon megszokott 2-2-es tagolást használtuk. Példa: *Nokia 5110* - /nokia ötvenegy tíz/ Ehhez a résztémához kell kapcsolni a részleges emberierőforrás-igényt is, amire akkor van szükség, ha új nevű készülék, illetve új nevű gyártó kerül be a felolvasandó listába. A rendszer csak olyan szavakat képes felolvasni, amik a beszédadatbázisában és szövegadatbázisában már szerepelnek és szinkronizálva is vannak egymással. Mivel folyamatosan jelennek meg új készülékek, illetve újabb gyártók a szolgáltató kínálatában, ezért szükséges volt biztosítani a korpusz bővítését eseti hang- és

szövegelemekkel. Ilyen esetekben a bemondó felolvassa azt a mondatot, amelyben az új készülék neve szerepel (a cég stúdiójában). Ez egy új bemondásnak számít, ezt hozzá kell csatolni a nagy korpuszhoz, amiből a beszédszintetizátor dolgozik. Ennek lépései ugyanolyanok, mint amit a korpusz készítésénél elmondtunk. Lehetséges, hogy bővíteni kell a kivételszótárt is, esetleg új szabályt kell implementálni a szövegfeldolgozó modulban. Mivel a korpusz elkészítésének nem minden lépését lehet automatizálni, ezért a működés során is folyamatos és szoros együttműködésre van szükség a fejlesztők és az operátor között. A szolgáltató részéről a kiindulási korpusz adatai nem minden esetben voltak egységes formában, például az, hogy *részlet* vagy *részlettel* a bemeneti szövegben néhol keverve volt viszont az adott mondat helyes értelmezése szerint a *részlettel* kifejezésnek kellett elhangzani az esetek döntő többségében (gépelési hiba). Ezért ezeket általában a program helyettesítette a megfelelő változattal, ügyelve rá, hogy a szöveg értelmes maradjon. Az ilyen jellegű hibák, amikor egy bemondó olvassa fel, nem jönnek elő, mert ő automatikusan kiegészíti a mondatot a megfelelő ragokkal, szótagokkal esetleg szavakkal (értelmessé teszi a felolvasást). Viszont, ha egy gép végzi a feldolgozást, akkor azt fel kell készíteni az ilyen esetek kezelésére.

A rendszer elemei. A rendszer elemei, valamint az azok közti kommunikáció a 12.6. ábrán láthatók:



12.6. ábra. Az árfelolvasórendszer blokkjai

1. A Beszédkorpusz, mely tartalmazza a szintézis során használt hangfájlokat, valamint a hangfájlokhoz tartozó egyéb leírókat, melyeket XML struktúrában tárol a rendszer. Ilyen leírók a hangfájlban elhangzó szöveg, a hang, illetve szóhatárok, a zöngés hangok alapfrekvenciái, az egyes elemek intenzitásértéke.

2. Az elemkiválasztásos technológiára épített beszédszintetizátor, mely a megadott szöveg alapján előállítja a hangzó formát. Ennek része a szövegfeldolgozó mo-

dul, valamint a kivételszótár is. A beszédszintetizátor a különböző leírókat a beszédkorpuszról a memóriában tárolja, míg a hullámformákat tartalmazó fájlokat csak abban az esetben nyitja, meg, ha szükség van rá (ezek együttes mérete több Gbyte).

3. A vezérlő információkat az SQL adatbázison keresztül kapja a beszédszintetizátor. Ebbe az adatbázisba kerül bele az új szintetizálandó mondat, valamint – amikor végzett – a szintetizátor szintén ebbe tölti be az elkészült hullámformát is és egyéb jellemző értékeket.
4. A kliensprogram a felhasználó operátor számára teszi lehetővé a rendszer használatát. Ennek segítségével lehet bevinni a szintetizálandó szöveget és meghallgatni az elkészült hullámformát. A kliensprogram és a többi komponens HTTP protokollon keresztül kommunikálnak egymással.

A gyakorlati használat pontokba szedve:

1. Az operátor megkapja a felolvasandó szövegeket elektronikus formában
2. A szövegeket blokkokra bontja a szoftverrel
3. Ellenőrzi a blokk szövegét (ránézéssel)
4. Szintetizálja a blokkot
5. Meghallgatja a szintézis eredményét
6. Ha hangzásilag rossz mondatot talál, akkor korrigálhat a szövegen is, és újból szintetizálhatja a blokkot (3 blokk)
7. Elvégzi az összes blokk szintetizálását, ezzel elkészült a hanganyag.

Ez az előállítási forma zajlik le hétről hétre. Ha új készüléknév, cégnév van a kiindulási szövegben, a szintézis után ki fog derülni, hogy a beszédszintetizátor képes-e megfelelő minőséggel azt előállítani. Amennyiben nem képes, akkor stúdióban felveszik a készülék vagy cég nevét egy, az árlistában szereplővel azonos struktúrájú mondatba illesztve, majd elküldik a fejlesztőknek, hogy ezzel egészítsék ki a szintetizátor beszédkorpuszait. Erre a műveletsorra csak néhányszor van szükség évente.
Tapasztalatok. A rendszer 2009 decembere óta működik a szolgáltatónál, és azóta a telefonos ügyfélszolgálati rendszerben elhangzó árlista jellegű promptok 80%-a ezzel készül. A rendszer hetente átlagosan 20 percnyi beszédet állít elő. A felhasználó operátorok elmondása szerint a régebbi heti 3–4 óra helyett jelenleg átlagosan 30 percet foglalkoznak ezzel a feladattal. A rendszernek még az is előnye, hogy előre is lehet dolgozni vele, például, ha tudják, hogy a következő héten megjelenik egy új készülék, melynek nevét nem olvassa fel megfelelő minőségben a gép, akkor már előtte, amikor még nem szerepel az árlistában, fel tudják venni a stúdióban, elküldhetik a fejlesztőknek, és mikor már szükség van rá, akkor már elérhető a rendszerben is.

Összefoglalásként elmondható, hogy sikerült egy olyan hibrid beszédtechnológiai megoldást készíteni, melynek segítségével egy telefonos ügyfélszolgálati rendszerben együtt lehet használni a természetes bemondásokat, illetve a mesterségesen

előállított, szintetizált mondatokat anélkül, hogy a felhasználók ezt észrevennék. Lényeges eleme a működtetésnek a folyamatos emberi felügyelet és a bővítés. Ez az ára a magas minőségű beszédszolgáltatásnak. Ez az első ilyen megoldás Magyarországon.

12.3.8. Beszéddel vezérelt automatikus telefonközpontok

Fegyó Tibor

A telefonos beszédvezérelt alkalmazások egyik tipikus felhasználási köre a névalapú tárcsázás (name-dialing), ennek egy konkrét megvalósítása a VOXenter rendszer (Fegyó et al. 2003). A szolgáltatás előnye, hogy nem kell kezelőre várni, nem kell a melléket fejből megtanulni, hanem elegendő ismerni és bemondani a keresett személy nevét, a rendszer már kapcsolja is.

A rendszer felépítése. A 12.3.6. fejezetben bemutattuk a helyfüggő szolgáltatásokat, az ott leírt hangportálrendszer magja megegyezik a VOXenter megoldásban használttal. Néhány eltérést érdemes kiemelni. A VOXenterben az LBS kliens szerver modulok nem szükségesek, viszont egy adatbázisillesztő modulra szükség lehet, amely egy külső telefonkönyv-adatbázishoz csatlakozik. A telefoninterfész ebben az esetben ritkán PRI-ISDN, gyakrabban alkalmazunk analóg, BRI-ISDN vagy VoIP (SIP) interfészt, de a rendszer működése szempontjából ezek nem jelentenek különbséget. A beszédfelismerő nyelvi modellje opcionálisan kiegészült, lehetőség van kulcsszókeresés alkalmazására, így nem csak a név bemondását, hanem a mondatba foglalt neveket is elfogadja a rendszer (például *Kérem Horváth Gyulát kapcsolni; Horváth Gyulát keresem; Horváth úrral szeretnék beszélni*). A kulcsszókeresés nem csak annyiban különbözik, hogy más nyelvi modellt töltünk a rendszerbe, hanem a betöltés módja is módosult, ugyanis a nyelvi modell mérete ez esetben sokszorosa az egyszerű parancsszófelismerő modellének. A kis modelleket a dialógus futása közben is be lehet tölteni, azonban a nagyobbak betöltése akár egy percig is eltarthat, ami egy telefonos dialógus esetén nem megengedhető késleltetést eredményez. Ezért a nagy méretű nyelvi modelleket előre a rendszer indulásakor be kell tölteni. Amennyiben módosul a modell, akkor azt csatornánként le kell cserélni, amikor a csatornák inaktívak.

A rendszer működése. A VOXenter dialógusával az alapvető híváskezelői információk kerültek megvalósításra:

- a bemondott személy, esetleg részleg/osztály neve alapján, vagy a beütött DTMF kód (mellék) alapján kapcsolja a keresett személyt
- termékekkel kapcsolatban általános információkat lehet közölni, mielőtt kapcsolná a rendszer az illetékest

- amennyiben nem kapcsolható a mellék, akkor több lehetőség közül is lehet választani: hangposta, továbbkapcsolás mobilszámra, továbbkapcsolás híváscsoportban, új keresés stb.

A VOXenter száz alkalmazottat foglalkoztató kisvállalkozást, és több ezer fős biztosító társaságot is képes kiszolgálni, sőt kísérleti jelleggel közel kétszázezer nevet tartalmazó tudakozó jellegű szolgáltatással is sikerrel teszteltük. A felismerendő nevek számának növekedése tipikusan a hívásforgalommal is összefügg, de mivel a rendszer elosztott modulokból épül fel, így több párhuzamos csatorna esetén az egyes komponenseket kell csak megsokszorozni. 4–8 csatornáig egyetlen mai PC is képes kiszolgálni közepes szótár esetén a rendszert, nagyobb szótárnál a gépigény is nőni fog. A VOXenter kiterjesztésének másik lehetősége a nevek számának növelése mellett a funkciók kiterjesztése. Sok esetben szükség van arra, hogy sokféle „termékről" nyújtsunk információt, ahol a termék tág értelemben értendő. Például egy okmányiroda esetén a számos különböző ügyintézéshez szükséges tudnivalók, bemutatandó iratok felsorolása rutin jellegű feladat, így hatékonyan automatizálható. Az ügyfélnek elegendő bemondania a keresett ügytípust, és a rendszer felolvassa az összes vonatkozó tudnivalót. Az okmányirodába érkező hívások túlnyomó többsége ilyen rutin jellegű munkát igényel, az ügyintéző ha ezek alól felszabadul, akkor hatékonyabban tudja végezni a nagyobb odafigyelést és szaktudást igénylő feladatokat. A számos ügytípus miatt egy nyomógombos menürendszerben kikeresgetni a megfelelő almenüt még akkor is több időt vesz igénybe, ha a felismerő esetleges tévesztése miatt újra be kell mondani a kívánt ügytípust. Egy ilyen egyszerűnek tűnő névalapú tárcsázórendszer is számos beállítási lehetőséggel rendelkezik, fel kell venni a neveket, azok összes kiejtési variációjával együtt, mellékeket, csoportokat, üzeneteket, időzítéseket stb. Így a paraméterek beállításának elősegítésére egy grafikus szerkesztői felület készült, ahol az összes paraméter könnyen beállítható. A VOXenter beszédvezérelt telefonközpont számos magyar vállalatnál és intézménynél működik.

12.4. Internetes alkalmazások

Az internet átszövi világunkat és minden új műszaki megoldás előbb-utóbb ide is beépül. Így van ez a beszédtechnológiával is. Az automatikus szövegfordítók, a szöveget hanggá átalakító szoftverek, a tartalomelemzők, az adatbázisokban hang alapján kereső megoldások szerves részeivé válnak a világhálónak. A gép és ember közötti kommunikáció hatékony megoldásában a nyelv- és beszédtechnológia egyre komolyabb szerepet kap, és a jövőben ez csak növekedni fog. A következőkben olyan magyar beszédinformációs alkalmazásokat mutatunk be, amik működnek, elérhetők és használhatók az interneten.

12.4.1. Időjárás-előrejelzés írott szöveges és hangos modalitással

Zainkó Csaba–Németh Géza

Az időjárással kapcsolatos információk fontosak és állandó igény van rájuk. Egy automatizálható hangos szolgáltatás lehetőségét vetíti előre az alább ismertetett kísérleti alkalmazás. A cél, hogy az időjárással kapcsolatosan megkapott szöveget a gép szép hangon, jól érthetően elmondja. A megoldást a 10.3.7. fejezetben ismertetett elemkiválasztásos beszédszintézis-technológiára alapoztuk. A BME TMIT által fejlesztett rendszert csatoltuk a `http://www.metnet.hu/` portál szöveges időjárásjelentéseihez. A szolgáltatás teljesen automatikus, ahányszor cserélik a szöveget a portálon, annyiszor kerül automatikusan előállításra annak hangos változata. Az alábbiakban ismertetjük a szintézis fő modulját képező szöveg- és hangkorpuszok kialakítási folyamatát erre a konkrét feladatra.

Mondatkorpusz. Első lépésben meg kell határozni azt a mondatkorpuszt, ami jó hatásfokkal lefeldi a témát szóhasználatban, mondatszerkezetben. Ezt egy bemondó a második fázisban fel fogja olvasni. Belátható, hogy ehhez olyan nagy méretű nyers szöveganyagot kell összeállítani, amelyiknek a szóállománya lefedi a majdan szintetizálandó időjárás-jelentéses mondatok szóállományát és megfelel a 10.3.1.5. fejezetben leírt komplex prozódiai modell követelményeinek is. A szóanyag lefedéséhez egy éven keresztül gyűjtöttünk (saját, automatikus szoftverrel) magyar időjárásjelentés-szövegeket 20 különböző weboldalról. Az eredmény 56 000 mondat, bennük 493 000 szó és 43 000 szám. A teljes szöveg 5200 különböző szóalakot tartalmazott (a szóalak akkor különbözik, ha a szó betűkarakteres formájában legalább egy karakter különbözik). A statisztikai analízis azt mutatta, hogy a leggyakoribb 500 szóalak lefedte a mondatállomány 92%-át, 2300 szóalak pedig a 99%-át (prozódiai szempontok nélkül). Ez a kis szám abból ered, hogy a témakört limitáltuk az időjárásos mondatokra. A második lépésben alakítottuk ki az 56 000-es mondatállományból a későbbiekben felolvasásra kerülő mondatkorpuszt (5260 mondat), amely tartalmazta az 5200 szóalakot és azok prozódiai variánsait (összesen 82 000 szó). Ez a mondatkorpusz képezi a beszédszintetizátor elsődleges (szóalapú) keresési terét.

Beszédkorpusz. A beszédkorpusz a mondatkorpusz felolvasásából jött létre. Ez a hangalapú keresés tere a szintézis során. A hivatásos női bemondó 4 héten át, heti 2–3 alkalommal, naponta 4–5 órát olvasva mondta fel a mondatkorpusz 5260 mondatát. Az eredmény 11 órás folyamatos beszédanyag. Ez képezi a beszédkorpuszt a szintézishez. A keresés biztosításához feldolgoztuk a beszédkorpuszt, a folyamatos beszédanyagot mondatokra daraboltuk, minden mondat kapott egy azonosítót. Ezután a mondatkorpusz szöveges formáját (minden mondatát) manuálisan össze kellett vetni a beszédkorpusz tartalmával (vö. a 8.2.3. fejezettel), kijavítottuk az esetleges felolvasási hibákat (a bemondó néha tudat alatt átformálta az írott szöveget, ilyenkor a szöveget a felolvasott formához igazítottuk, továbbá a felolvasott számokat és rövi-

dítéseket is szövegesen ki kellett fejteni (például *4–6 C fok* szövegből a kimondáshoz igazított szöveg (fonemikus átirat) így alakul: *négy, hat celziusz fok*, ahol a vessző azt jelenti, hogy a bemondó szünetet tartott. Ezzel a módszerrel teljesen szinkronba hoztuk a hangot és annak szöveges formáját (a szünettartást is jelöltük vesszővel). A következő lépésben minden mondat hullámformáját elláttuk szóhatárokkal, banghatárokkal és a hangok szimbólumaival (fonetikus átírat). Ezt a BME TMIT automatikus beszédfelismerő szoftverének (Mihajlik et al. 2002) támogatásával végeztük (vö. a 9.5.3.2. fejezettel). Az automatikusan átírt és címkézett anyagot – a szó- és hanghatárok pontosságát – ezután félautomata módszerrel ellenőriztük. Itt felhasználtuk például a magyar beszédre kidolgozott időtartammodellt (10.3.1.4. fejezet), minden hangra jósoltunk egy időtartamot és összehasonlítottuk a bejelölt értékkel. Nagy eltérés esetén manuálisan megkerestük a hiba helyét és korrigáltuk a rossz jelzést. A szóhatár jelölése sok esetben nem végezhető el egyértelműen, alkalmaztuk a 4.5. fejezetben leírt módszert.

A fentiekből látható, hogy egy tényleges korpuszalapú alkalmazás szövegszintű és beszédkorpuszának elkészítése komoly munkát igényel. Az időjárásfelolvasó meghallgatható a `http://www.metnet.hu` honlapon.

12.4.2. Híradókereső – internetes hang-videókeresés kulcsszavak alapján

Fegyó Tibor

A technológia fejlődésével egyre nagyobb számban és méretben születnek köz-, illetve magángyűjteményi videó- és audioarchívumok, mint például a Magyar Országgyűlés archívuma, a Nemzeti Audiovizuális Archívum, a rádió- és televíziótársaságok archívumai, de idesorolható a Youtube is. Az archívumok méretének növekedése a keresés problémáját hozza előtérbe. A szerzők néhány címkét, metaadatot fűznek az egyes felvételekhez, mint például a készítés helye és ideje, témakör stb. Ezen címkék alapján azonban nem lehet az archívumokban tartalom alapján keresni, ami jelentősen szűkíti az archívumok értékét, használhatóságát és a feltöltött anyagok életciklusát. Az archívumok mérete nő, így elkerülhetetlen igényként lép fel a tartalom szerinti keresés gépesítése. Elvileg lehetséges megoldás, ha valaki legépeli ezeket a tartalmakat, mint például az Országgyűlés esetében, de ennek az erőforrásigénye nagyon nagy, ugyanis 1 órányi hanganyag lejegyzése nagyságrendileg 10 órányi munkát vesz igénybe. A beszédtechnológiai feladat tehát adott, a kézi lejegyzést ha egyszerűbb formában is, de automatizálni kell. A korábbi fejezetekben foglalkoztunk a nagyszótáras folyamatos gépi beszédfelismerés problémakörével. A technológia adott, csak alkalmazni kell. Modelleket kell tanítani az adott feladathoz. A betanítás során, mint mindig, itt is két modellt kell készítenünk, egy akusztikai és egy nyelvi modellt. A modellek elkészítése előtt azonban fel kell mérnünk, hogy milyen téma-

körhöz készítjük el azokat. A gépi beszédfelismerő (és az ember is) pontosabban működik, ha szűkebb a témakör és ezzel együtt kisebb a keresési tér. Gyakorlati korlátok is adódnak, az számítógép operatív memóriájában el kell férni az alkalmazott modelleknek. Egyszóval keresnünk kell egy viszonylag szűk témakört, ilyen például a politikai jellegű hírek témaköre, amelyre hatékony modellt tudunk építeni. Erre mutatunk be példát alkalmazási szinten. A felismerő szoftvert tehát fel kell készíteni az elhangzott híranyagok tulajdonképpeni annotálására (felismerésére). Ennek két fő lépése van. Az első lépés az akusztikai modellek betanítása. Ezt a felismerendő híranyagok akusztikai minőségéhez kell igazítani, ennek leghatékonyabb módja, ha különböző hírforrások hanganyagát gyűjtjük össze, és ezek segítségével tanítjuk a modellt, vagy adaptáljuk a korábbi modelleket. Többnyire jó minőségűek a hírműsorok felvételei, így célszerű széles sávú, például 16 kHz-el mintavételezett adatokkal tanítani a modellt. A híranyag közvetlenül még nem alkalmas a tanításra, el kell végezni az annotálást és címkézést (8. fejezet) A megtisztított szöveget a szónál kisebb egységekre kell bontani, mivel a magyar nyelv ragozó jellegéből fakadóan túl sok szó szerepel csupán egy-egy alkalommal a tanítókorpuszban, ami a statisztikai modellezést megnehezíti. A szónál kisebb egység a morféma vagy morfémaszerű elem már kellő számban szerepel (hangkapcsolat, szótöredék). A morfémák bevezetésével a modellben alkalmazott szótárméret is csökken, és a modell általánosítóképessége is nő, hisz képes korábban nem látott szóalakokat generálni, de természetesen ezzel együtt a hibás alakok elfogadásának valószínűsége is megnő. Az akusztikai modell a témakörtől nem függ, így ez a modell újrafelhasználható hasonló körülmények között készített egyéb felvételek esetén is. A nyelvi modell készítéséhez az adott témakörben lejegyzett nagy méretű szöveges korpusz szükséges (8.1.1.2. fejezet). Hírműsorok esetén élhetünk azzal az egyszerűsítéssel, hogy a csak írott és nem az elhangzott tartalmakat alkalmazzuk a tanítás során. Ezzel egyik oldalról megspóroljuk néhány 100 órányi hanganyag kézi lejegyzését, másik oldalról viszont az élő beszédre jellemző elemeket kihagyjuk a modell tanításból. Minél spontánabb a beszéd, amit fel kell ismernünk, annál kevésbé használhatjuk ezt a megoldást. A hírekre optimalizált akusztikai és nyelvi modellek segítségével készített automatikus kereső megoldás (feljesztők: BME TMIT, AITIA, Digital Natives), a híradókereső a `http://www.mindroom.hu` oldalon tekinthető meg. A híradókeresőben 7 televízió-csatorna közel 50 hírműsora kerül feldolgozásra folyamatosan. Amennyiben egy-egy beszélőre adaptáljuk a fenti akusztikai modellt, akkor elérhető a 80%-os szópontosságú felismerés. Ha az alap akusztikai modellt használjuk, akkor is 70% feletti a pontosság. A kidolgozás során a politikai jellegű hírműsorokra fókuszáltunk, természetesen kipróbálható másfajta műsorokra is a rendszer, például orvosi vagy bulvár tartalmú műsorokra, itt azonban, ahogy várható, a pontosság messze elmarad a korábbiaktól, 20–50% körül szórnak az eredmények. A gyakorlati tapasztalatok alapján a 70% feletti felismerési eredmény nagy hatékonyságú keresést tesz lehetővé. A keresésen túl további elemzésre is lehetőséget ad a rendszer. Az adott időszakban leg-

gyakrabban előforduló szavakat szófelhőben lehet ábrázolni, így hamar képet kaphatunk egy-egy időszak vagy műsor fontosabb eseményeiről, pontosabban az ezekhez tartozó kulcsszavakról.

12.4.3. Szövegfelolvasás a webfordítás színesítésére

Olaszy Gábor–Németh Géza

Az internet helyet ad számos hasznos nyelv- és beszédtechnológiai szolgáltatásnak is (lásd `http://hlt-platform.hu`). A szövegek automatikus felolvasása is egyre népszerűbbé válik mint színesítő eszköz. Ilyen módon került bele a ProfiVox magyar beszédszintetizátor a `http://www.webforditas.hu` oldal kínálatába (Prószéky–Tihanyi 2009). Az alapvetően gépi fordítást ajánló honlapon a magyarra fordított szövegeket és a magyarul beírtakat a felhasználó fel is olvastathatja (hangszóró ikon). a hangszóró ikon használatával. Választani lehet férfi és női hang között. A felolvasáskor a szintetizátor kizárólag a szöveget veszi alapul. A modalitásokat érzékeli és megvalósítja, tehát akár egy szavas kérdőmondatot is helyesen alakít át beszéddé. A felolvasás a portál második legnépszerűbb szolgáltatása. A fejlesztés a Morpholgic Kft. és a BME TMIT közös munkája.

12.5. Közlekedési alkalmazások

A közlekedés területén számos példáját láthatjuk a beszédtechnológiai alkalmazásoknak (változó hangminőséggel). Ezzel a technológiával magas fokú automatizálások valósíthatók meg. A jegyrendeléseknél géppel beszélgethetünk, a járműveken a gép bemondja a megállókat és egyéb információkat is, az állomásokon gép adja tudtunkra, hogy mikor érkezik és mikor indul egy-egy járat, szerelvény, a GPS készülékünk beszél hozzánk, a mobiltelefonok is beszélhetnek és adhatnak közlekedési útmutatást. A következő fejezetben egy jövőbe mutató, távvezérelt állomási tájékoztató mintarendszert mutatunk be, amely működése során (2008–2009) bebizonyította, hogy magas minőségű, ugyanakkor rugalmasan távkezelhető gépi felolvasórendszereket hatékonyan lehet alkalmazni a közlekedésben.

12.5.1. Vasútállomási utastájékoztató

Zainkó Csaba–Németh Géza

A vasútállomások hangosbemondóiból hallhatóak az érkező és induló vonatok időpontadatai és egyéb, az utazóközönség számára fontos közlemények. Ezeket az információkat hagyományosan úgynevezett kötött szótáras technológiával állítja elő egy gép. A technológiáról, annak előnyeiről, hátrányairól a 10.2. fejezet szól. A leggondosabb tervezés ellenére a jó hangminőség folyamatos biztosítása ezekben a rendszerekben egyre nehezebb. A közlendő információk függnek az aktuális menetrendtől és egyéb körülményektől (sztrájk, időjárás, műszaki hiba). A 10.3.7. fejezetben ismertetett új technológia (BME TMIT fejlesztés) kiválthatja a régebbit, rugalmassága megoldást adhat az eddigi üzemeltetési problémákra. A hangüzenetek állandó újrafelvétele költséges és időigényes, függ a bemondótól, ha csak a megváltozott részek felvétele történik meg, akkor nehezen biztosítható az egységes minőség, gyors változtatásra nincs lehetőség.

Az új megoldás Magyarország egyik északkeleti vasútállomását érintő közlemények bemondására készült el női hangra (kiterjeszthető az egész országra). A gyakorlatban is tesztelték, és kifogás nem merült fel sem a hangminőséggel, sem a rugalmassággal kapcsolatban. Működése ugyanazon az elven alapszik, mint az időjárásfelolvasó (10.3.7 és 12.4.1. fejezet). A gép egy megfelelő nagyságú beszédadatbázisból dolgozik, nem előre felvett üzeneteket kapcsol össze. A rendszer több szabadságot ad az üzemeltetőnek, mint a korábbiak, ugyanakkor tökéletes hangminőségével teljes szolgáltatást nyújt. A rendszer nagy előnye a rugalmasság és a gyorsaság, hiszen egy-egy információ megváltozása esetén a bemondások generálása kevesebb, mint egy perc, és emberi munkát sem igényel, csak a szöveget kell megadni. A hagyományos rendszerhez képest (emberi hangfelvétel rögzítése, feldolgozása és telepítése) ez lényeges előny és költséghatékony megoldás. A beszédkorpusz 1200 mondatot tartalmaz. Amennyiben az egész országra kiterjesztenék a rendszer beszédkorpuszát, akkor további, közel 3000 mondatot kellene hangfelvétel formájában rögzíteni és megfelelő címkékkel ellátni. Ez nem jelent drasztikus növekedést. A bővítés során ugyanis csak a megállóhelyekkel kapcsolatos kifejezéseket kellene felvenni, az összes többi már benne van az egyetlen állomást érintő rendszerben is (időpontok, tájékoztató üzenetek stb.). A jövő mindenképpen az ilyen megoldásoké.

12.6. Diktálórendszerek

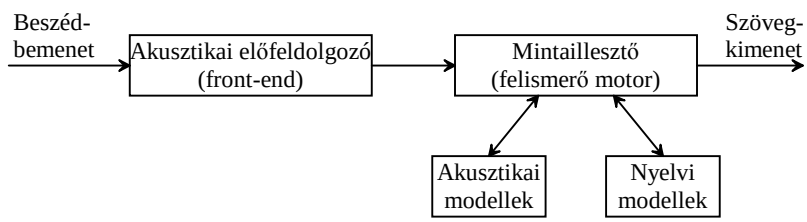
A diktálásalapú szöveglejegyzés automatizálása már régi vágya az emberiségnek. A beszédfelismerő szoftverek fejlődésével egyre közelebb kerülünk ennek a teljesítéséhez. A feladat nagyon bonyolult, a gátló tényezőkről a beszédfelismerés fejezete-

iben olvashattunk. Gyakorlati megoldásként jelenleg még csak olyan rendszer jöhet szóba, amelyik személy- és témakörspecifikus. Ez megkötést és kellemetlenséget jelent a felhasználásban. Mindezek ellenére a gyakorlatban már használnak ilyen rendszereket. Korlátozás, hogy a használat megkezdése előtt rá kell tanítani a felismerőt a személy hangjára, be kell állítani a témakört és a környezeti zaj szintfokozatát, fajtáját (iroda, ügyfélváró, porta, autó stb.). A következő fejezetben ismertetett diktálórendszer az egyik legígéretesebb területe a beszédfelismerési kutatások felhasználásának. Az Amerikai Egyesült Államokban számos helyen már napi gyakorlatban alkalmaznak orvosi leletező diktáló szoftvereket.

12.6.1. Leletező beszédfelismerő

Szaszák György–Vicsi Klára

A leletező beszédfelismerő szakorvosi vizsgálatok leleteinek elkészítéséhez nyújt segítséget beszédinterfész biztosításával. A leletet nem kell begépelni, hanem elegendő bediktálni, az esetleges hibákat azonban utólag javítani kell (Velkei–Vicsi 2004). A rendszer használata akkor kifizetődő, ha a diktálás és a hibajavítás együttes ideje elmarad a begépeléshez szükséges időtől, így a szakképzett orvos idejének nagyobb részét tudja a betegekkel való közvetlen foglalkozásra fordítani. A leletező felismerő voltaképpen egy diktálórendszer. Felépítése és működése nagyjából megfelel a rejtett Markov-modellre épített, folyamatos, beszédhangalapú beszédfelismerőnek. Az akusztikai előfeldolgozó modul (nevezik front-endnek is) végzi a keretképzést, vagyis a kepsztrum, illetve a mel-frekvenciás kepsztrális együtthatókhoz (MFCC) nagyon közeli, bark-szűrősoros elemzéssel nyert együtthatók kiszámítását 10 ms keretidővel. A mintaillesztő egység (a felismerő motor) az akusztikai és nyelvi modelleket használja tudásforrásként (12.7. ábra). Az akusztikai modellek külön-külön



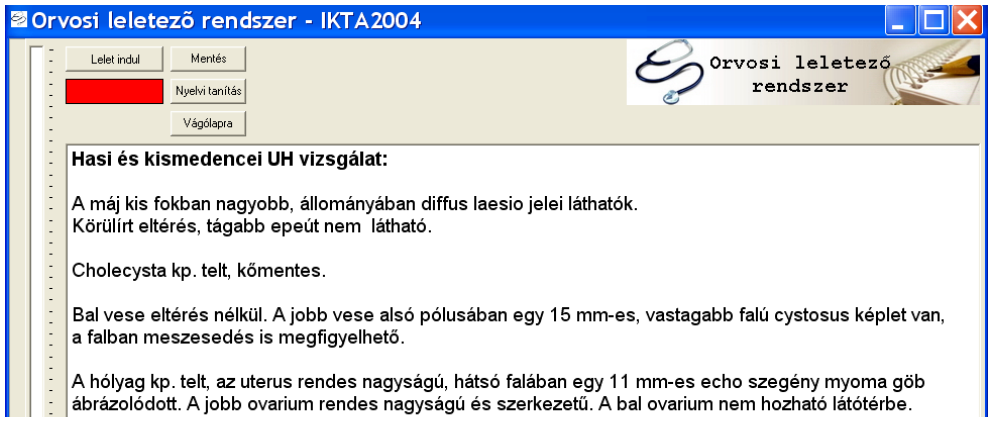
12.7. ábra. Leletező beszédfelismerő blokkszintű felépítése

készülnek női és férfi beszélőkre, széles sávú, tehát 16 kHz-en mintavételezett, viszonylag csendes környezetben rögzített beszédjel alapján. A tanítóminták a Magyar Referencia Beszédadatbázisból származnak (8.1.1.1. fejezet). A nyelvi modell bigram statisztikai modell, amelyet korábban még gépeléssel előállított leletkorpusz-

ból taníthattunk. A nyelvi modell a vizsgálattípustól függően változik, például külön nyelvi modell szükséges a hasi és kismedencei ultrahangos vizsgálatokhoz, külön a röntgenezéshez, az endoszkópos vizsgálatokhoz stb. Ebben a konkrét alkalmazásban a felhasználónak a rendszer indításakor, illetve a bejelentkezéskor ki kell választania a leletezés területét fedő nyelvi modellt, sőt, akár személyes nyelvi modelljét is. A beszélőadaptálás során mindig a felhasználó korábban diktált – és szükség esetén javított – leletei szolgálnak (12.8. ábra) alapul. A nyelvi modellek szétválasztásának oka egyrészt az, hogy ezzel a modell jobban illeszkedik az egyes feladatokhoz. Az optimális működést segíti, ha több kisebb modellt használunk, ugyanis ha egyszerre csak egy nyelvi modell lesz aktív, a futási idő igen rövid (valós idejű a működés). Egy-egy nyelvi modellhez 5–10 ezer szavas szótár tartozik, azaz a szótár közepes méretű, emiatt a nyelvi modell sem túl nagy. A valós idejű működéshez azonban még így is legalább 1 GB RAM és gyors CPU szükséges (a műveletigény döntő részét a mintaillesztés teszi ki). A leletező diktálórendszer felhasználói felülete lényegében

12.8. ábra. Leletező beszédfelismerő felhasználói felülete

egy szövegszerkesztő (12.9. ábra), amely képes fogadni a beszédfelismerő kimenetét, de biztosítja a billentyűzetről történő beírás lehetőségét is. A két üzemmód között a felhasználó nyomógombok segítségével választhat. A rendszerrel 90% fölötti szótalálati pontosság érhető el, különösen a beszélőadaptáció elvégzése esetén. A beszédfelismerő komponens futhat szerveren is, illetve offline is, azaz a diktafonra rögzített leletbemondások utólag is felismertethetők és szöveges formátumúvá alakíthatók. A rendszer struktúrájából adódóan lényegében tetszőleges diktálórendszerré alakítható, amennyiben a nyelvi modellt a kívánt felhasználási területre cseréljük. Ha a használat akusztikai körülményei is változnak (eredetileg csendes környezet, széles sávú beszédbemenet), akkor az akusztikai modelleket is cserélni kell. A leletező felismerőt a Budapesti Műszaki és Gazdaságtudományi Egyetem Távközlési és Médiainformatikai tanszéke és a Szegedi Tudományegyetem fejlesztette ki a Semmelweis



12.9. ábra. Leletező beszédfelismerő felhasználói felülete

Egyetem II. Belgyógyászati Klinikájával és a Szegedi Orvostudományi Egyetemmel együttműködve.

12.7. Beszédtechnológia a vakok és gyengénlátók szolgálatában

Kiss Géza

A beszédtechnológia azok számára is nagyon fontos, akiknek az életminőségét jelentősen javíthatja a használata. Ezen belül is kiemelkedő jelentőségű azoknak, akiknek elvesztett vagy soha nem birtokolt képességét segít pótolni, a fogyaték nélkül élőkkel egyenlő esélyt biztosítva számukra. Az ilyen, fogyatékkal élőket segítő vagy rehabilitációjukat elősegítő (általában elektronikus) technológiákat összefoglaló néven segítő technológiáknak (assistive technology) nevezzük. Ezek egyik fajtája az augmentatív és alternatív kommunikáció (augmentative and alternative communication), melynek célja a verbális kommunikáció segítése vagy helyettesítése.

A segítő technológiák alkalmazhatóságát a kommunikációban két mai tendencia is növeli. Az egyik, hogy a korábban személyes kapcsolatokban vagy toll és papír segítségével zajló kommunikáció egyre nagyobb mértékben digitálisan zajlik, például telefonon és elektronikus levelezéssel, így kézenfekvő a számítógépes feldolgozás bevonása a kommunikációs láncba. A másik, hogy mindennapi életünk segítőiként egyre több elektronikus eszközt hordunk magunknál, amelyeket az eszközök képességeinek és teljesítményének növekedésével egyre több célra tudunk felhasználni. Jó példa erre a mobiltelefonok növekvő használata, amelyek a beszélgetési funkció mellett időmérő, naptár-, játékeszköz-, fényképezőgép-, rádió-, zenelejátszó, internetkliens-, navigációsrendszer-feladatokat is betöltenek, hogy csak a leggyakrabban használtakat említsük. Az ilyen eszközök segíthetik a fogyatékkal élők

személyes kommunikációját is, érzékelőikkel felfogva a külvilág jeleit, és a szükségletnek megfelelően transzformálva azokat az audio, a vizuális és a taktilis modalitások között. Az állam is egyre nagyobb figyelmet szentel állampolgárai eme körére, ami megjelenik többek között esélyegyenlőségi törvényekben és állami projektekben. Az Egyesült Államokban például az 1996-os telekommunikációs törvény kötelezővé tette, hogy (meghatározott kivételektől eltekintve) minden televízióadást feliratozással lássanak el. Az Európai Unióban a 2007-ben módosított audiovizuális szolgáltatásokról szóló irányelv írja elő, hogy a tagállamoknak ösztönözniük kell a joghatóságuk alá tartozó médiaszolgáltatókat annak biztosítására, hogy szolgáltatásaikat fokozatosan tegyék hozzáférhetővé a látási vagy hallási fogyatékkal élők számára. Bár a médiában a feliratozás ma javarészt emberi munkával valósul meg, a gépi beszédfelismerés fejlődése ezt várhatóan egyre inkább ki tudja majd váltani. A segítő technológiákban a beszédtechnológia ágai között a beszédfelismerés a hallássérült, a beszédszintézis elsősorban a látássérült és a beszédükben gátolt (beszédsérült vagy akár bezártságszindrómás) emberek kommunikációját segíti. A beszédszintézis fontos lehet még például a diszlexiásoknak, akik számára érthetőbb a beszélt, mint az írott szó; emellett az idegen nyelven tanulóknak a megértés gyakorlására, és mindenki másnak is olyankor, ha más tevékenységre kell koncentrálnia, de közben tud figyelni a felolvasott beszédre. Az alábbiakban bemutatunk néhány példát a beszédtechnológia olyan alkalmazásaira, amelyeknek célja a fogyatékkal élők segítése, és amelyek meg is valósultak magyar nyelvre. A sort még sokáig lehetne folytatni további alkalmazásokkal, melyek a világ valamely nyelvére már elkészültek.

12.7.1. Képernyőolvasás

A számítógép legfőbb kimeneti eszköze hagyományosan a monitor. Azonban ahhoz, hogy a nem látók is használni tudják a számítógépeket, a vizuális helyett más modalitást használó kimeneti eszközre van szükség. Az egyik lehetőség ma a taktilis (tapintási), a másik az akusztikus (hallási) modalitás használata. A taktilis kijelzőkre példa a Braille-kijelző, amely Braille-karakterek egy sorozatát tudja megjeleníteni. A Braille-karaktereket egyenként hat, téglalap alakba rendezett pont segítségével lehet megjeleníteni, amelyek mindegyike lehet kiemelt vagy besüllyedt állapotban, így összesen 64 kombinációt adva. Ennek használata nagy gyakorlatot igényel, és az idősebb korukban megvakulók vagy látáscsökkenést elszenvedők általában nem sajátítják el. További korlátja a használatának, hogy a Braille-kijelzők nem mehetnek egy bizonyos méret alá (ma a legkisebbek kb. 20 cm hosszúak), ami gátolja az egyre elterjedtebb okostelefonokkal való használatukat. A másik lehetőség a beszédkimenet: a számítógépen fut egy képernyőolvasó program, amely egy beszédszintetizátor segítségével felolvassa a képernyőnek az éppen aktív részletét, valamint beszéddel

ad visszajelzést a felhasználó által bevitt információról is. Ezt a kimenetet minden, a beszédet halló és értő látássérült használni tudja. A gyengén látók számára is rendelkezésre áll hasonló lehetőség az úgynevezett képernyőnagyító programok formájában. Ahogy az elnevezés is mutatja, ezek képesek kinagyítani a képernyő aktív részletét, és több más vizuális funkciót is tartalmaznak, mint például kontrasztosabb megjelenítés, fekete-fehér kijelzés, a kurzor helyét mutató célkereszt. Emellett általában tartalmaznak a képernyőolvasókéval azonos beszédfunkciót is. Mivel most csak a beszédfunkciót tárgyaljuk, az egyszerűség kedvéért a továbbiakban csak a képernyőolvasókról beszélünk, de a leírtak vonatkoznak a képernyőnagyítók beszédfunkciójára is.

A vak és gyengén látó emberek a képernyőolvasókkal nem egyszerűen csak használni tudják a számítógépet, sőt az okostelefonokat is, hanem gyakorlatilag a látókkal mind sebességben, mind hatékonyságban egyenértékűen tudják használni. Ez lehetőségek óriási tárházát nyitja fel a számukra: szabadon kommunikálhatnak (elektronikus levelezés, csetelés), ők is hozzáférnek a manapság elérhető szinte korlátlan mennyiségű digitális információhoz (például hírekhez, irodalmi művekhez és tananyagokhoz), valamint segítségével munkát is végezhetnek (például szoftverfejlesztés, szövegszerkesztés, hívásközpontban információ szolgáltatása). Ezért ők a beszédszintézis-technológia talán egyik legrégebbi és leglelkesebb felhasználói, akik egyben nagyon hibatűrők is. Akár gyengébb hangminőségű beszédet is elviselnek (erre a kezdetekben sok példa volt). A használat során eljutnak oda, hogy olyan gyors felolvasási sebességgel is jól elboldogulnak, ami tapasztalatlan hallgató számára szinte érthetetlenül gyors; ez még hatékonyabbá teszi számukra a számítógép használatát. Egy ilyen program és a futtatásához szükséges személyi számítógép sajnos nem olcsó. Szerencsére a rászorulóknak időről-időre lehet jelentkezni különböző szervezetek pályázati kiírásaira, hogy kedvezményes áron vagy ingyen jussanak hozzájuk. Léteznek tanfolyamok is, ahol elsajátíthatják a számítógép és az alkalmazás használatához szükséges alapvető ismereteket.

A képernyőolvasók működése. A képernyőolvasó programok az operációs rendszer szolgáltatásain keresztül érik el a képernyő tartalmát. Például a jelenleg legelterjedtebb Microsoft Windows alatt az Accessibility Frameworks segítségével (Microsoft Active Accessibility, illetve User Interface Automation). A meghangosítandó tartalom felolvasásához az alkalmazástól független szöveg-beszéd átalakító (Text-to-Speech, TTS) komponenst használnak. Rendszerint a képernyőolvasóval együtt települhet egy vagy több, különböző technológiájú és nyelvű TTS változat megfelelő szoftvertámogatással. Erről részletesebben írunk a 13.2. fejezetben.

A képernyőolvasótól elvárt funkciók.

- Beszélő telepítővel rendelkezik, hogy a nem látók is tudják telepíteni.
- Könnyen elérhető, részletes, kontextusfüggő súgóval rendelkezik, hogy segítsék az alkalmazás megismerését.

- Segít navigálni a futó alkalmazások között, és az egyes alkalmazások ablakán belül. Ehhez új képernyőelemre való lépéskor felolvassa annak néhány tulajdonságát. Például az épp kijelölt ikon feliratát, az aktivált ablak fejlécét, az aktuális menüpont nevét és gyorsító billentyűjét, egy dialógusablakban az épp aktivált szerkesztőmező nevét és tartalmát, egy szövegszerkesztőben az aktuális sort.
- Felolvas szöveges dokumentumokat, weblapokat megfelelő hangsúlyozással, tagolva, a speciális szövegrészeket a szerepüknek megfelelően olvasva (például címek, táblázatok, URL-ek). Lehetővé teszi a dokumentumok struktúrájában a navigálást.
- Lehetővé teszi dokumentumok, levelek írását, szerkesztését.
- Beállítható a beszédhez használt hangkarakter és annak sebessége, hangereje, alaphangmagassága. A különböző fajtájú üzenetekhez ezek általában külön beállíthatók (például a menükhöz, szöveghez, a súgó üzeneteihez).
- A szöveges kurzor és az egér kurzor kezelésének lehetősége.
- Az olvasandó szöveg nyelvének megfelelő felolvasása, a nyelv kézi vagy automatikus kiválasztásával.

A képernyőolvasókban használt beszédszintetizátoroktól elvárt funkciók

- Olvassa fel a szavakat a köznapi kiejtésük szerint, valamint az írásmódot pontosan tükrözően is. Az utóbbira sok nyelven a betűzést használják, de magyar nyelven egy könnyebben használható, jól érthető lehetőség, hogy a betű szerinti olvasat hangzik el. A *Széchenyi* név esetén például a tényleges kiejtés [seːt͡ʃeːɲi], a betű szerinti olvasata pedig [s eː t͡s h ɛ ɲ i].

- A gépelés során a lenyomott billentyűkhöz rendelt betűket érthetően olvassa vissza. Például: *á* →[aː], *p* →[peː], *y* →[ipsilon]. Egy másik lehetőség a magyarban, hogy a betűhöz tartozó fonémát önmagában modja ki a rendszer. Például *á* →[aː], *p* →[p], *y* →[i]. Ilyenkor egy szó gépelésekor szaggatottan ugyan, de sok esetben maga a gépelt szó hangzik el (ehhez hozzá lehet szokni). Sok más nyelvre ez a lehetőség nem használható.

- Legyen több hangkarakter, széles tartományban szabályozható sebesség és hangmagasság.
- Nagy sebességen is jól érthető hangzás.
- Minél kellemesebb, hosszú hallgatáskor sem zavaró hangzás.

Képernyőolvasó programok. A képernyőolvasó programok besorolhatók négy kategóriába: kereskedelmi termékek, ingyenes programok, operációs rendszerbe épített szolgáltatások és a felolvasásra is képes dokumentumkezelők. Alább mindegyikre megadunk példákat, de természetesen az ilyen programok palettája is gyorsan bővül, változik. Jelenleg az Informatika a Látássérültekért Alapítvány honlapján (`http://infoalap.hu`) megtalálható a magyar közönség számára elérhető legtöbb

szoftver- és hardvertermék aktuális listája. Néhány termékként kapható professzionális képernyőolvasó, számos funkcióval és terméktámogatással: Jaws képernyőolvasó és MAGic képernyőnagyító a Freedom Scientifictől; Window Eyes a GW Microtól; ZoomText az Ai Squaredtől. A Jaws, MAGic és ZoomText programok magyar változatával a ProfiVox magyar beszédszintetizátor települ fel alapértelmezett hangként, de bármelyikhez tetszőleges beszédszintetizátort lehet illeszteni, ez a kapcsolódási felület Windows alatt általában az MS SAPI technológia (lásd 13.2. fejezet), amely ismeri a képernyőolvasó kapcsolódási felületét. Egyre több ingyenes képernyőolvasó is elérhető, Windows alatt például a Thunder, Linux alatt pedig az Orca elnevezésű program. A modern operációs rendszerekbe beépített számos kisegítő lehetőség között az egyik egy minimális funkcionalitású képernyőolvasó: például Microsoft Windows alatt a Narrator, Mac OSX alatt a VoiceOver. Az Informatika a Látássérültekért Alapítvány fejlesztése a BeLin (Beszélő Linux), melyet kifejezetten a látássérült emberek számára hoztak létre. Ez az Ubuntu Linux disztribúció egy speciális változata, amely a telepítéstől fogva beszédtámogatással működik a beépített Orca képernyőolvasó segítségével. Egyes programok nem hangosítják meg a képernyő teljes tartalmát, ezért nem számítanak képernyőolvasónak, de a betöltött dokumentumokat képesek felolvasni. Ilyen például az Adobe Acrobat Readernek a Read Out Loud funkciója és a Browsealoud webböngésző.

12.7.2. Dramatizáló

Ebben a fejezetben a szintetizált hangoskönyvkészítőket mutatjuk be. Ezeknek a célja hasonló a hangos online könyvtárakhoz, mégpedig szövegek hallgathatóvá tétele beszédszintetizátor felhasználásával, de a megközelítés több szempontból eltér. Először is a cél itt audioállományok létrehozása, amelyek ezután bárhol használhatók, nem szükséges a meghallgatásukhoz élő internetkapcsolat. Másodszor nem központosított megoldás, hanem egyéni felhasználásra készült: egy számítógép és egy erre telepített hangoskönyvkészítő szoftver segítségével bárki előállíthat hangoskönyvet az otthonában, saját maga vagy közössége részére. Harmadszor, nem előre rögzített a használható szövegek köre, hanem teljesen hétköznapi szövegekből (például a napi levelezéséből) is készíthet ilyen hangállományt, amit akár meghallgathat utazás közben egy hordozható lejátszóval. Erre a célra számos szoftvert találhatunk már ma is az interneten.

A fejlettebb szoftverekben arra is van lehetőség, hogy a felolvasás módját részleteiben meghatározzuk, mintegy „dramatizáljuk” a szöveget; ez a lehetőség főként irodalmi szövegek esetén válik nagyon értékessé. A dramatizálásba beletartozik például szereposztás rendelése a párbeszédekhez, valamint a részletekbe menően pontos és életszerű kiejtés, ritmus és hangsúlyozás meghatározása a szöveg bármely mon-

datához. Ez ugyan energiabefektetést igényel a szöveget előkészítő „dramaturgtól”, viszont elég egyszer elvégezni, és utána számos hallgató profitálhat az eredményből, a dramatizált szövegkönyv vagy szintetizált változatának terjesztésével.

DEX — Dramatizált Elektronikuskönyv-szerkesztő. Példaképpen ismertetünk egy magyar nyelvre készült szintetizált hangoskönyvkészítőt, a DEX 1.0 dramatizált elektronikus könyvszerkesztő és hangoskönyv-konvertáló programot. Ez egy látássérültek számára készült ingyenes Windows alkalmazás, amelyet az Informatika a Látássérültekért Alapítvány készített a BME TMIT együttműködésével. Az első változatának elkészítéséhez anyagi segítséget nyújtott a T-Online, a második verzióhoz pedig a Vodafone. A program kifejezetten rászorulók részére készült, a szövegbeszéd konverzió csak akkor működik benne, ha fut a számítógépen a Jaws képernyőolvasó vagy MAGic képernyőnagyító program is. Viszont a szerkesztői felülete ettől függetlenül működik tetszőleges szövegszerkesztővel, hogy bárki segíthesse a rászorulókat könyvek dramatizálásával, azaz bárki lehessen dramaturg.

A DEX több különböző bemeneti formátumból képes beolvasni szöveget (egyszerű szöveg, html stb.), melyekből mp3 formátumú fájlokat állít elő. Mivel egy szintetizált könyv több száz gigabájt méretű is lehet, a szöveg hosszától és a kimenet bitsebességétől függően, a dramatizálást végző személy beállíthatja, hogy milyen időtartamonként és/vagy milyen pontokon darabolja el a program a szintetizált fájlokat. A szövegbe a felolvasást vezérlő címkéket is beszúrhat, melyeknek két fajtája van: egy adott ponton ható (ilyen a szünet és a következő szó hangsúlyát meghatározó címke), és az adott ponttól kezdődően érvényes beállítások (ezek a hangerő, hangmagasság, beszédsebesség, betűzés, szünetek be/kikapcsolása, intonáció be/kikapcsolása, szereplőváltás). A dramaturg összeállíthat egy „társulatot” az egyes könyvekhez, amelynek a tagjai olvassák majd fel a szereplők szövegrészeit. Az egyes szereplők hangját több beállítással tehetjük egyedivé: négy különböző hangkarakterből (két férfi és két női hang) választhatunk, és a hang többi tulajdonságát is megadhatjuk (hangerő, hangmagasság, sebesség stb.). A fájlhoz tárolható a könyv néhány adata (szerző, cím, kiadás éve) és a dramatizáló személye; ezeket a kimenetként készülő mp3 fájlban is tárolja címkeként.

Jelenleg nem áll rendelkezésünkre pontos adat a program felhasználói táborának számosságáról, de Szuhaj Mihály, az Informatika a Látássérültekért Alapítvány kuratóriumának elnöke elmondta, hogy a visszajelzések alapján számos látássérült számára a hétköznapi szoftverhasználatuk részévé vált a DEX-szel való hangoskönyvkészítés.

12.7.3. Hangoskönyvek

Amióta csak készülnek könyvek, azóta felolvassák őket hangosan. Ennek csak egyik oka, hogy sokáig a magányos olvasást is hangos tevékenységnek tartották; a néma olvasás gyakorlata viszonylag új keletű, csak a 10. és 14. század között kezdett elterjedni (Saenger 2000). Az egymásnak való felolvasás ősi gyakorlat (lásd például Fischer 2004, 56., 192–194. o.). Számos okból olvasunk fel egymásnak. Egyrészt a felolvasás közösségi élményt teremthet: lehetőséget arra, hogy egy csoport tagjai ugyanarra figyeljenek, és azután együtt gondolkodjanak és megbeszéljék a hallottakat (például író-olvasó találkozókon, kiadatlan művek előzetes bemutatásán, vallási könyvek istentiszteleti felolvasásán). Másrészt olyanoknak olvasunk, akik nem tudnának maguknak olvasni: azért, mert épp mással vannak elfoglalva, de a felolvasásra tudnak figyelni, vagy mert nem ismerik az írást (kisgyermekek és írástudatlan felnőttek); és azoknak, akik nem látnak elég jól az olvasáshoz, vagy egyáltalán nem látnak.

A 19. század vége óta a felolvasást sok esetben ki tudja váltani a rögzített beszéd, a hangrögzítés és lejátszás technológiájának használata. Ennek nem csak az az előnye, hogy nem kell minden egyes alkalommal megkérni vagy épp megfizetni valakit a szöveg felolvasására, hanem az is, hogy a felvétel készülhet egy színész professzionális olvasatából, amit sokszorosítanak. Így sokan élvezhetik a felolvasott könyvet ugyanabban a kifinomult előadásmódban, ugyanakkor hozzáférhető áron. Máig ez a legelterjedtebb módja a felolvasás kiváltásának, melynek számos formátumát használják, például CD-lemezek (korábban bakelitlemezek), kazetták, audiokönyv, e-book, mp3 fájl, podcast.

Hangos online könyvtárak. Az online könyvtárak olyan, az interneten elérhető digitális könyvtárak (digital library), ahol sok különféle témájú és stílusú írott művet találunk, a lejárt szerzői jogú könyvek esetén ráadásul ingyen. Ezek között számos olyant is találhatunk a világhálón, amelyek hangos formában elérhetővé teszneknek szövegeket, méghozzá ezt több különböző megközelítést használva teszik. Több weboldalon olyan hangoskönyveket tesznek közzé, amelyek szerzői jogi védelmet nem élvező könyvek felolvasott változatát tartalmazzák. A könyveket önkéntesek olvassák fel, hogy azok széles közönség számára ingyen elérhetők legyenek. Magyar nyelven a Magyar Vakok és Gyengénlátók Országos Szövetsége (MVGYOSZ) tett közzé számos hangoskönyvet ingyen a Magyar Elektronikus Könyvtáron keresztül (mek.oszk.hu) mp3 formátumban. Több idegen nyelvű weblap is létezik, amelyen ilyen könyveket adnak közre, és önkéntesnek is lehet jelentkezni. (Ilyenek például az angol librivox.org, a francia litteratureaudio.com és a spanyol leerescuchando.net; ezek közül a LibriVox oldalán könyvünk írásának idején is találhatók már magyar nyelvű hangoskönyvek is). Egy másik megközelítés, hogy a weblapon található könyveket a böngészőhöz letölthető kiegészítővel olvastathatjuk fel, amely beszédszintetizátort tartalmaz. Például a freebookstoread.com ezzel a megoldással

működik. Természetesen egy ilyen böngészőkiegészítő kiváltható egy képernyolvasó szoftver (lásd a 12.7.1. fejezetet) használatával is. Egy harmadik megközelítés, hogy a szolgáltatást nyújtó weblap maga alakítja a szöveget hanganyaggá beszédszintetizátor segítségével. A fejezet hátralévő részében ezzel a megközelítéssel foglalkozunk, és bemutatunk egy magyar példát, a VilágHallót: vilaghallo.hu.

Könyvek hangosítása beszédszintetizátorral. A beszédszintézis fejlődése mára lehetővé teszi, hogy emberi közreműködés nélkül, automatikusan alakítsuk beszéddé az írott szöveget. Ez minőségben és élvezhetőségben egyelőre jelentősen elmarad attól, ahogy egy színész vagy akár egy átlagos képességű felnőtt felolvasná, viszont olcsóbb, vagy legalábbis gyorsabban megvalósítható módja a felolvasásnak. Vannak olyan szövegfajták is, amelyekre kevés érdeklődő volna, vagy éppen bizalmas információt tartalmaznak; ezek meghangosítására talán az egyetlen mód a beszédszintézis használata.

A beszédszintézist használó hangos online könyvtárak célja az, hogy ennek a viszonylag új technológiának a felhasználásával elérhetővé tegyenek irodalmi műveket, oktatási anyagokat és egyéb szakirodalmat felolvasott formában a rászorulóknak, valamint bárkinek, aki egyéb elfoglaltságai végzése során szeretné ezeket hallgatni, például autószerelés vagy házimunka végzése közben. A felhasználóknak ilyenkor nem kell saját beszédszintetizátorral rendelkezniük. Ennek előnye, hogy az ilyen szoftverekből a jobb minőséget nyújtók költségesek lehetnek; valamint egy kisebb számítási kapacitással rendelkező, lassabb, de megfelelő internet-sávszélességgel rendelkező gépen is igénybe tudják venni a szolgáltatást.

A beszédszintetizátorok, a beépített nyelvi elemzés segítségével egyre inkább képesek helyes kiejtést, életszerű hangsúlyozást, természetes hangzást produkálni. Ahol mégis tévednének, ott a felolvasás javítható a szövegben elhelyezett címkék segítségével, amelyekkel tetszőleges szintig feljavítható a felolvasás minősége. Ezzel a technikával megvalósítható, hogy a rendhagyó kiejtésű nevek megfelelően hangozzanak el, hogy ott és akkora szünet legyen a szövegben, amit az tartalmilag megkíván, és hogy a párbeszédekben minden szereplő saját hangszínt kapjon. Sőt, akár egy részletes, a színészi felolvasásnak megfelelő minőségű címkézés is megadható a mondatokhoz, ha a hangsúlyozás és dallamvonal minden részletét szabályozni akarjuk. Ha ez mégsem így történik, annak fő akadálya nem technikai, hanem az, hogy ehhez rengeteg költséges emberi munka befeketésére volna szükség. Mivel az ember számára kisebb hibák esetén is jól értelmezhetők maradnak az ilyen felolvasások, a gyakorlat rendszerint az, hogy a szöveg ilyen címkézésébe nem fektetnek munkát, hogy cserébe ingyenes vagy nagyon olcsó legyen ez a szolgáltatás.

VilágHalló. A VilágHalló egy magyar nyelvű hangos online elektronikus könyvtár. Ez a fejlesztés 2003-ban készült el a BME TMIT közreműködésével. A Magyar Elektronikus Könyvtár több ezer híres művét teszi elérhetővé hangos formában. Az első verziója az ITEM kutatás-fejlesztési pályázat támogatásával valósult meg, a második az eredeti továbbfejlesztése, amit 2008-ban egy GVOP pályázat keretén belül

végeztek. A szolgáltatást a készítők adatai szerint már több tízezer felhasználó igénybe vette. Az első változatban a BME-TMIT ProfiVox beszédszintetizátora szólt (lásd a 10.3.6.1. fejezetet), a második változatban kiegészítették még egy beszédszintetizátorral, a Speech Technology Kft. férfi és női hangjú eszközével (ez utóbbi nem intonál).

A program a `http://vilaghallo.hu` oldalon elérhető, és egy Java program telepítése után használható. Szerveroldalon alakítja a szöveget beszéddé, a felhasználók internetkapcsolatukon keresztül a beszédet töltik le és szólaltatják meg a gépükön. A felhasználók fejezetenként vagy akár mondatonként navigálhatnak a szövegben, és könyvjelzőket is elhelyezhetnek, hogy az előzőleg olvasott vagy éppen kedvenc részt könnyen kikereshessék.

12.7.4. Beszélő bankautomaták

Németh Géza–Kiss Géza–Bartalis Mátyás

A pénzkiadó automaták ellátása beszédkimenettel kézenfekvő gondolat. Olyan felhasználóknak lehet ez nagy segítség, akik látássérültek. A bankautomata egyrészről beszéddel tájékoztatja a használóját, hogy mi látható a képernyőn, ezzel könnyíti a tájékozódást, mivel minden billentyűművelet hanggal is támogatva van. Másrészről felolvassa a felvenni kívánt pénzösszeget és az esetleges hibaüzeneteket. Az alkalmazás egyszerűnek látszik, a megvalósítás komoly műszaki együttműködést kíván az automata gyártója és a beszédszintetizátor szállítója között. Mind hardvermind szoftveroldalon el kell végezni bizonyos bővítést. Az automatát el kell látni fejhallgató-csatlakozóval, valamint a beszélőszoftvert be kell illeszteni az automata szoftverrendszerébe. Az ilyen fejlesztések viszonylag hosszú időt vesznek igénybe, mivel a gyártási technológia szigorú szabályrendszerében kell változtatásokat végezni, ezeket tesztelni kell és jóvá is kell hagyni.

Az automatákon a hangkimenet igénybevételéhez az ügyfélnek saját fejhallgatóval kell rendelkeznie, azt tudja bedugni az automata hangkimeneti csatlakozójába. Ezzel biztosítva van, hogy az automata beszédét csak az ügyfél hallja.

Magyarországon a ProfiVox beszédszintetizátor hangjával először a Volksbank egyik automatáját látták el 2005-ben, a Vakok Állami Intézetének közelében lévő bankfiókban. A fejlesztő (BME TMIT) ingyenesen bocsátotta a bank rendelkezésére a speciális szoftvert. Ez volt az első beszélő bankautomata Magyarországon. A későbbiekben két nagy gyártó építette be automatáiba ugyanezt a beszédkimenetet.

A Wincor Nixdorf 2008-ban bővítette ki automatáit a fenti beszélő szoftverrel. A cég a beszédkimenet biztosításán túl két lényeges újítást is bevezetett. Az első, hogy a készülék vezérlését immár nem csak a képernyő melletti, nehezebben elérhető funkciógombok segítségével, hanem – a nemzetközi szabványnak megfelelően

– a számbillentyűk, az úgynevezett PIN-pad segítségével is meg lehet oldani, amelyek elérése és kezelése a vak ügyfelek számára lényegesen egyszerűbb. A másik fontos különbség, hogy a fülhallgató-csatlakozó feletti gomb segítségével a hangerő is egyénileg beállítható.

2010-ben az NCR cég, a világ legnagyobb bankautomata-gyártója is követte a példát és beépítette automatáiba a ProfiVox magyarul beszélő szoftvert, hogy vakbaráttá tegye automatáit.

12.7.5. NaviSpeech – beszélő navigátor látássérült gyalogosoknak

Viktóriusz Ákos–Németh Géza–Tóth Bálint

Az okostelefonok a látásszervi fogyatékossággal élők nélkülözhetetlen segítőtársává váltak a számos, nekik készült speciális beszélő programmal kiegészített alkalmazásnak köszönhetően. Így ők a telefont a grafikus kijelző használata nélkül is tudják kezelni. A létező alkalmazások között található többféle képernyő-felolvasó, SMS és e-mail kliens, továbbá egyéb más programok is. Ezek a szoftverek jelentősen megkönnyítették életüket, hiszen az érintett személy az okostelefonjára feltelepítheti a programot, mindenhová magával viheti, és szükség esetén használhatja (telefonálhat, internetezhet).

A közlekedés és a navigáció azonban még mindig jelentős problémát okoz számukra. Különösen ha önállóan, emberi segítség nélkül kell közlekedniük vagy eljutni új, eddig ismeretlen helyekre. Utazásaik során a tömegközlekedési eszközöket is rendszeresen igénybe veszik, itt is számos nehézségbe ütközhetnek. Az általános célú kézi navigációs eszközök nehézkesen használhatók vagy teljesen használhatatlannak bizonyultak a látássérült emberek számára. Ennek oka, hogy a készülékek gyártói nem vették számításba e célcsoportra jellemző követelményeket. Emiatt felmerült a szükségessége egy olyan eszköz megtervezésének, amely megfelelő mértékben pótolhatná az emberi segítséget a közlekedés során.

Egy olyan navigációs céleszközt ismertetünk, amely a látássérült felhasználók körében kényelmesen használható, és megbízható navigációt szolgáltat gyalogos környezetben, mindemellett kihasználja az okostelefonok által nyújtott egyéb lehetőségeket. Az okostelefon-alapú navigációs alkalmazás célja, hogy a látássérült felhasználók közlekedését megkönnyítse. Az alkalmazás grafikus felülete szöveges megjelenítésre van korlátozva, az alkalmazás fő jellemzője a beszédalapú irányítás. A rendszer számos funkcióval rendelkezik (aktuális pozíció lekérdezése, útvonalfelvétel, útvonaltervezés stb.), melyeket a következőkben ismertetünk.

A BME-TMIT Beszédtechnológiai Laboratóriumában kifejlesztett *NaviSpeech* névre hallgató rendszer *Symbian* alapú okostelefonokon futtatható (Tóth–Németh 2007). A navigációt a legelterjedtebb helymeghatározó rendszer, a *GPS* (*Global Po-*

sitioning System) alapján végzi, melyet vezeték nélkül csatlakoztatható vevőegység, vagy a telefonba épített GPS vevő biztosít. A telefon beszéddel kommunikál a felhasználóval. A beszédkeltés a BME-TMIT ProfiVox szövegfelolvasó rendszerre épül (Olaszy et al. 2000b). A navigáció során beszéddel tájékoztat az aktuális pozícióról (utca, sarok), és utasításokat adva vezeti a látássérültet a helyes irányba. Emellett a képernyőn megjelenik a kimondott utcanév szöveges formája is. A betűk mérete nagyítható, tehát a gyengén látó emberek innen is tájékozódhatnak, újra elolvashatják az információt. A program által értelmezhető útvonalak a programba épített automatikus útvonalrögzítő használatával alakíthatók ki. A megtervezett útvonalak ezután fájlok formájában korlátlan számban eltárolhatóak a telefon memóriájában és onnan betölthetőek az alkalmazás futtatása közben, amikor elindulunk a megtervezett útvonalon.

Mi kell a *NaviSpeech* használatához? Rendelkezni kell egy Symbian-alapú okostelefonnal és beépített vagy külső GPS vevőegységgel. Az új generációs és a régebbi operációs rendszert futtató Symbian-alapú okostelefonokkal is kompatibilis, ezért a legújabb technológiák kihasználása mellett az elérhető árkategóriába tartozó készülékeken is lehet használni.

A Navi-Speech felépítése. Az alkalmazás induláskor beszéddel üdvözli a felhasználót, majd próbál kapcsolatot létesíteni a *GPS* rendszerrel (külső egység alkalmazásakor *Bluetooth* adatátviteli csatornát használ). A csatlakozás eredményéről szóban tájékoztatja a felhasználót. Sikeres csatlakozás esetén az alkalmazás másodpercenként lekérdezi a GPS vevőtől megkapott, majd értelmezett adatokat. Eközben várakozik a felhasználói parancsokra, amelyek a telefon billentyűi által vagy a menüből adhatók ki. A felhasználó bármikor kérhet segítséget a „0” gomb megnyomásával, a program ekkor felolvassa a billentyűk aktuális funkcióit. A navigáció közben elérhető főbb funkciók könnyen aktivizálhatók az okostelefon billentyűzetének használatával.

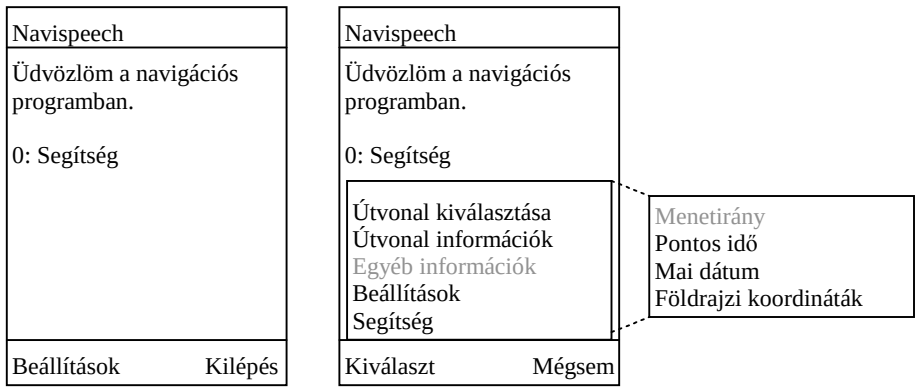
A telefon számbillentyűzetét tekintve három használati mód létezik: „Normál”, „Útvonalpontok kézi léptetése” és „Útvonal felvétele”.

Normál módban az alkalmazás alapfunkciói működnek. A GPS vevő által küldött vagy az útvonallal kapcsolatos információkhoz juthat hozzá a személy.

Útvonalpontok kézi léptetésénél a felhasználó végignézheti (hallgathatja) az útvonalpontok listáját, valamint módosíthat az útvonal pontjain (írásban). Az alkalmazás mindig az aktuálisan kiválasztott útvonalpont felé vezeti a felhasználót, ugyanakkor van lehetőség a legközelebbi pont kiválasztására is (ahol éppen áll, ahhoz melyik a legközelebbi megadott útvonalpont).

Útvonalfelvételhez egyszer végig kell járni az útvonalat. Szerkesztéskor a felhasználó az adott ponthoz érve az útvonalpontot gombnyomással rögzítheti. Ekkor vagy akár később is az adott ponthoz tartozó nevet rendelhet (például: Bojtos utca sarok). Ugyanígy kell rögzíteni a bejárt útvonal minden kívánt pontját. Az útvonal végén el kell menteni az készített útvonaltervet. Egymás után több útvonaltervet is létrehozhatunk ilyen módon. A készülék kezelésére a két funkciógomb és az iránybil-

lentyűk szolgálnak. Az alkalmazás a menü megnyitásakor és a kurzor mozgatásakor hangosan felolvassa az aktuális menüpont nevét, így látássérült emberek is tudják kezelni (12.10. ábra). A menüpontok felolvasása a *Beállítások* menüben opcionálisan ki/bekapcsolható. Az *Útvonal kiválasztása* menüpontot aktiválva a felhasználó



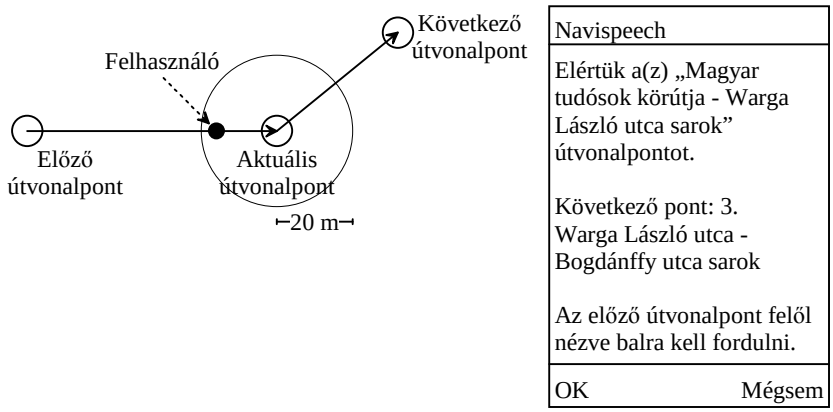
12.10. ábra. Az alkalmazás az indításkor (bal), a menürendszer választható pontjai (jobb)

végigtallózhatja az előre megtervezett útvonalfájlokat, majd ezek közül választhat egyet. Ezután a navigáció elkezdődik, a készülék hanggal vezeti az embert. Folyamatosan figyeli a GPS vevőtől kapott földrajzi koordinátákat, és összehasonlítja azokat a betöltött útvonal hasonló koordinátáival. Ennek alapján képes információkat szolgáltatni a felhasználónak. Navigáció közben a program a következő események bekövetkezésekor szólal meg:

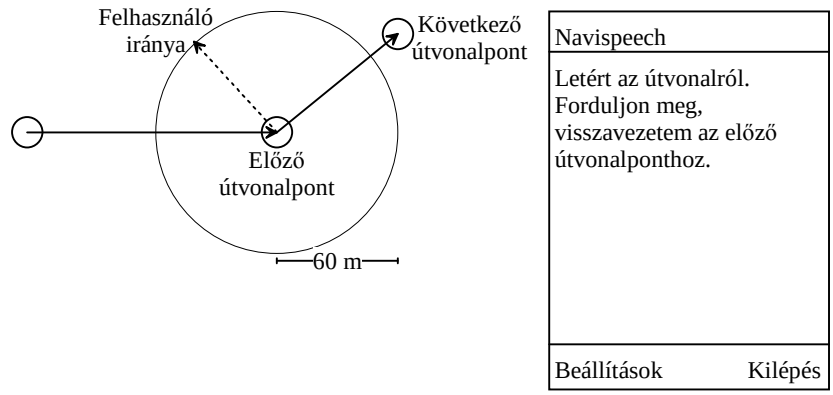
Aktuális útvonalpont elérése. Ha a felhasználó az aktuális útvonalpontot 20 méterre megközelíti, a gép felolvassa azt, valamint a soron következő útvonalpont nevét. Elmondja a következő pont eléréséhez szükséges relatív fordulási irányt az előző pont felől nézve (kivéve az első útvonalpont elérésekor, mert ekkor még nem tud mihez viszonyítani). A cél útvonalpont elérésekor az alkalmazás közli a felhasználóval, hogy megérkezett úti céljához (12.11. ábra).

Letérés az útvonalról. Ha a felhasználó már 60 méterre eltávolodott az előző útvonalponttól és haladási iránya feltehetően rossz a mérési eredmények alapján (nagy szöget zár be a helyes útvonal következő szakaszával), az alkalmazás figyelmezteti, hogy letért a helyes útról (12.12. ábra). Ekkor a felhasználónak meg kell fordulnia, és a program visszavezeti az előző útvonalponthoz, újból elmondja, hogy merre kell fordulnia, innen a helyes irányban folytathatja útját.

Grafikus felhasználói felület. Az alkalmazás grafikus felhasználói felülete kiegészíti a folyamatos beszédalapú visszacsatolást. A rendszer a szóban elhangzó utasításokat szövegesen is megjeleníti az okostelefon nagy méretű képernyőjén (a betűk nagyíthatók).



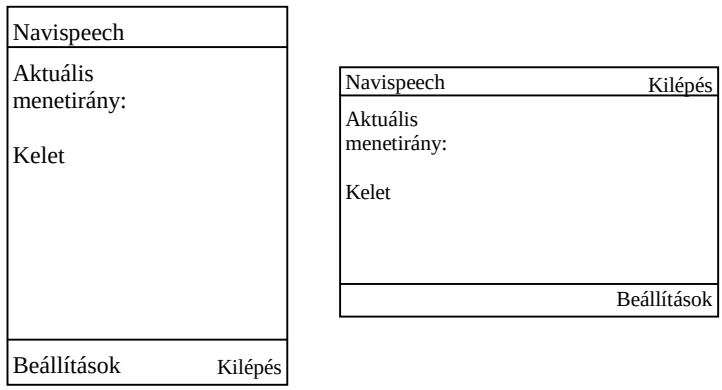
12.11. ábra. Útvonalpont elérése szimulációs ábrán (bal) és a képernyőre kiírt üzenet az útvonalpont elérésekor (jobb)



12.12. ábra. Az útvonalról való letérés szimulációs helyzetképe (bal), figyelmeztetés letéréskor (jobb)

A betűnagyság és a színek beállítására a *Beállítások – Kijelző beállítások* menüpontban van mód. Amennyiben a megjelenített szöveg nem fér el a képernyőn, a „Le” és „Fel” kurzorbillentyűkkel görgethető a megfelelő irányba.

Az új generációs *Symbian* operációs rendszert futtató okostelefonok lehetővé teszik a képernyő 90 fokos elforgatását (ennek neve „*landscape wiew*”, míg az eredeti a „ *portrait view*”). A két nézet között lehet az alkalmazás futtatása közben is váltani. Ekkor a kijelzőn lévő szöveg tördelése és görgethetősége a fordított képarányhoz igazodik (12.13. ábra).



12.13. ábra. Függőleges nézet (bal) és vízszintes (jobb)

12.8. Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére

Takács György

Ebben a fejezetben olyan eljárást mutatunk be, amelyik a beszédjelből mozgó száj képet készít, ezzel segítve a siketek kommunikációját (Feldhoffer et al. 2007). Siket emberekben hosszú gyakorlás után fantasztikus szintre fejlődik ki a beszéd megértése pusztán a szájmozgást nézve. Erre alapozva kommunikációs segédeszköz készíthető siket felhasználók számára, amely pusztán a szájról olvasáson alapul, és egy alkalmas mobiltelefon készülékben vagy egy IPTV „set-top-box" egységében megvalósítható. A Pázmány Péter Katolikus Egyetem Információs Technológiai Karán (PPKE-ITK) a Siketek és Nagyothallók Országos Szövetségének közreműködésével kifejlesztett rendszerben egy beszélő ember szájmozgásának képe jelenik meg grafikus kijelzőn. A rendszer a beszédszervek mozgását utánzó mozgó fej vezérlő paramétereit közvetlenül a beszédjelből származtatja.

Kiinduló megfontolások. Tisztában voltak a fejlesztők azzal, hogy a teljes emberi beszédfolyamatnak ez csak egy részleges megjelenítése, de számoltak azzal, hogy korlátai ellenére a siketek hasznos kommunikációs segédeszközhöz juthatnak ezzel a megoldással. A rendszer nagyban épít a siketek kifinomult szájról olvasási képességeire és a közvetlen kommunikációban kialakult folyamatos kiegészítő és hibajavító mechanizmusaira. Jelfeldolgozási szempontból a rendszer sarkalatos eleme, hogy időkeretenként meghatározott folyamatos jellegű beszédjellemzőkből folyamatos képjellemzőket számol. Az eddig ismert megoldások leképezték a folyamatos beszédfolyamatot diszkrét elemek (vizémák) halmazára, egy második lépésben pedig a diszkrét elemek halmazát alakították át mozgó fej képévé. Nagy előnye az ismertetésre kerülő közvetlen átalakításnak, hogy megőrzi a beszédfolyamat eredeti időbeli

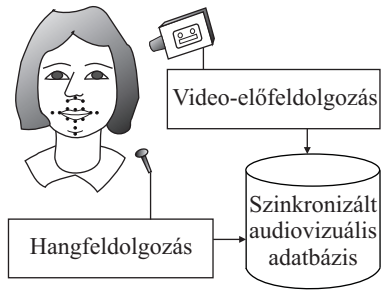
szerkezetét. Ezáltal a természetes beszédritmus eleve megőrződik. Releváns újdonság a rendszerben, hogy a folyamatot nem átlagos beszélők jeleivel tanították, hanem olyan hang- és képadatbázissal, amelyet képzett jeltolmácsok felvételeiből állítottak össze. A jó jeltolmácsok artikulációs stílusa és artikulációs dinamikája kifejezetten alkalmazkodik a siketek szájról olvasási igényeihez.

Tanító és tesztelő adatbázis. Az előzetes mérésekből a legfontosabb végkövetkeztetéseink egyike volt, hogy a szájról olvasott beszéd érthetősége nagyon függ az artikuláció minőségétől. A szájról olvasás sokkal nagyobb figyelmet igényel, mint a beszéd megértése hallás útján. A teljes beszédfolyamatról csak részleges információt ad, ezért a tévesztések eleve gyakoribbak. Az olyan artikuláció, amely eleve kiemeli a megkülönböztető jegyeket, valamint a lassú beszédtempó nagyon sokat segít a helyes megértésben. A hallók között messze legjobban teljesítik ezeket a követelményeket a képzett jeltolmácsok. Ők napi kapcsolatban állnak a siketekkel, és ezért alkalmazkodik artikulációjuk a szájról olvasás igényeihez. Ezért választották a fejlesztők azt, hogy a tanító adatbázis jeltolmácsok kép- és hangfelvételeiből álljon össze, még akkor is, ha a tényleges használatkor bárkinek a hangja szolgálhat jelbemenetként.

Kiderült az előzetes kísérletek során azt is, hogy a siketeknek komoly nehézségeik vannak a természetes nyelv komplikált nyelvtani szabályaival. Ha elemezzük az elektronikus leveleiket és SMS üzeneteiket, akkor látszik, hogy ugyanez megnyilvánul írott kommunikációjukban is. Amikor az érthetőséget teljes mondatok, rövid közlendők formájában adott nyelvi egységekkel próbáltuk mérni, akkor tapasztaltuk, hogy nem képesek a teljes üzenet pontos, szó szerinti visszaadására, hanem csak a legfontosabb üzenetelemek maradnak meg emlékezetükben. Sokszor az a kulcselem, amely az előzetes információk alapján a figyelmük középpontjába kerül. Az rögződik. A toldalékokkal sem nagyon foglalkoznak. Konkrét nevek fontosabbak számukra, mint a személyes névmások. Egy hirtelen témaváltás is igen nehezen követhető számukra. Ennélfogva az érthetőségvizsgálatok szokásos szövegei és módszerei eleve nem használhatók esetükben. Ezért kellett speciális szövegű adatbázist kialakítani a rendszer tanításához és teszteléséhez.

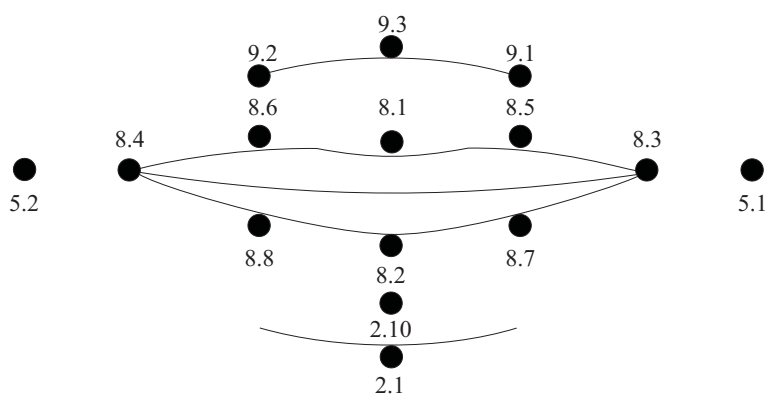
Az előzetes vizsgálatok eredményei alapján kétdimenziós fejmodellt alkalmaztunk, mivel a mélységinformáció elvétele alig csökkentette a szájról olvasás pontosságát. További méréseink igazolták, hogy a száj és környékének képe, akár mobiltelefon kijelzőjének méretében is elegendő a gyakorlatilag teljes megértéshez.

Audiovizuális adatbázis készítése. Az audiovizuális beszédadatbázis nem más, mint különböző bemondók összerendezett hangfelvételeinek és képfelvételeinek rendszere. A felvételek jelét azonos időkeretekben összeszinkronizálva dolgoztuk fel (12.14. ábra). A bemondók fejét puha korlátokkal rögzítettük, hogy a fej ingatását megakadályozzuk. Az egyes pontokat abszolút koordinátáikkal jellemeztük. Az MPEG-4 szabvány az emberi arcot 86 jellemző ponttal (Feature Point, FP) írja le. Előzetes kísérleteink alapján ezekből 15-öt választottunk ki a szájnak és környezetének le-



12.14. ábra. Adatbázisgyűjtő rendszer

írására. A felvételek során ezeket a pontokat könnyen lemosható és egészségre nem ártalmas sárga festékkel jelöltük meg a bemondók arcán. A beszédfolyamat képének leírása az MPEG-4 szabvány szerinti jellemző pontokkal több szempontból is előnyös. Egyrészt a száj és arc mozgásának tömör és elég pontos leírására alkalmasak az FP koordináták, másrészt a bevált szabványos fejmodellek alkalmazhatók ezekkel a pontokkal vezérelve, így az igen összetett modellek alapvető fejlesztésére nem kellett erőinket pazarolni. Csak képzett jeltolmácsokkal készítettünk felvételeket. A felvételekhez egyszerű kamerákat használtunk: 720x576 pontos felbontással, másodpercenként 25 képpel, PAL szabvány szerint. Ez azt jelenti, hogy 40 ms hosszú időablakokban készülhettek az összeszinkronizált kép- és hangelemzések. A felvételek a szájat és környékét rögzítették annak érdekében, hogy a kiválasztott jellemző pontok helyzete minél kisebb hibával meghatározható legyen. A fej többi részét (bár a szem környéke, vagy akár a hozzáfűzött tekintet is hordoz tartalmi információt) nem vontuk bele vizsgálatainkba. A képfelvételeket ezután emberi beavatkozás nélkül dolgoztuk fel. A képjelet a kontraszt, a fényesség és telítettség tekintetében úgy



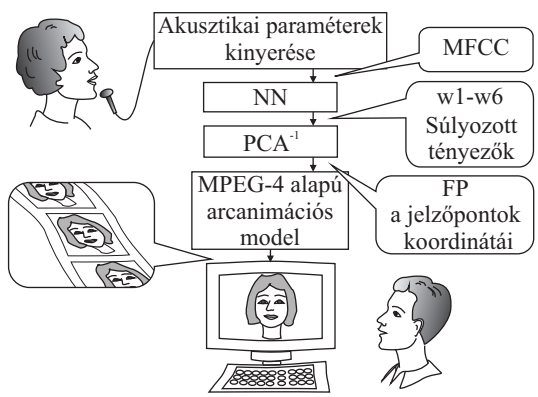
12.15. ábra. Az MPEG-4 jellemző pontok választott részhalmaza a száj környékének leírására

torzítottuk, hogy a jellemző sárga pontok minél jobban kiemelődjenek. A sárga pontokat végül az RGB komponensek komparálásával detektáltuk. A binarizált képen először dilatációs műveleteket végeztünk, hogy biztosan összefüggő képponthalmazt nyerjünk, majd lépésenként kívülről eróziós folyamattal szedtünk le képpontokat, amíg egyetlen pixel maradt, amit a jellemző pont közepének tekintettünk. Ez az automatikus eljárás legfeljebb 1–2 pixel eltérést eredményez a manuálisan kiválasztott középponthoz képest.

Tekintettel arra, hogy az egyes FP jellemző pontok vízszintesen 40–60, függőlegesen 80–140 pixel tartományban mozognak, az FP meghatározás fenti hibája elfogadható. A koordinátarendszert úgy választottuk meg, hogy középpontja az orr két oldalára helyezett (9.1 és 9.2 a képen) pontok között középen legyen, mivel ezek a pontok mozognak a 15 közül legkevésbé (12.15. ábra).

A beszédjelet hangcsatornában rögzítettük 48 kHz mintavételezéssel, 16 bites mintákkal. A tanító és tesztelő adatbázis szövegét a korábban leírt követelmények szerint választottuk ki. Eszerint a felvételek kétjegyű számokat, hónapok neveit, a hét napjait tartalmazták.

A beszédjel átalakítása szájmozgás képévé. A fejlesztés állapotában a rendszer lényegében egy személyi számítógépen futó programrendszer. A 12.16. ábrán az alapelemek feladata és kapcsolódása szerepel. A mintavételezett beszédjelen minden 40

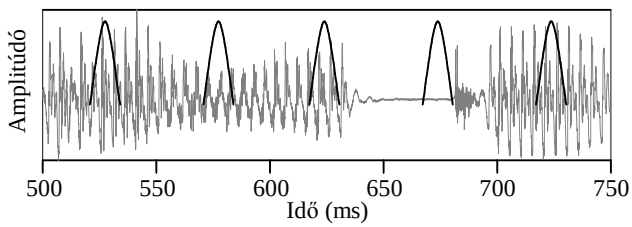


12.16. ábra. A beszéd-szájmozgás átalakító rendszer elemei

ms keretben meghatároztuk a mel-skála szerinti kepsztrumegyüttható-vektort (Mel-Frequency Cepstrum Coefficients, MFCC). Ezeket a jellemző vektorokat vezettük a neurális hálózat (NN) bemenetére, amely a kimenetein kiadja a szájmozgás pillanatnyi állapotát tömörítetten leíró súlytényező vektort [w1-től w6-ig]. A főkomponenselemzés (Principal Component Analysis PCA) inverz műveletével nyerjük a fejmodell vezérléséhez ténylegesen szükséges FP koordinátaértékeket. Ez egy lineáris kombinációs műveletet jelent csupán. Az FP koordinátákat meghatározzuk

minden időkeretre. A rendszer utolsó eleme a nyílt forráskódú LUCIA beszélőfejrendszernek egy enyhén módosított változata (Cosi et al. 2003). A modellt az FP koordinátákkal vezéreljük és a mozgó kép megjelenik a kijelzőn (lásd később).
Akusztikai lényegkiemelés. A bejövő beszédjelen először egy magasemelő szűrési műveletet hajtunk végre H(z) = 1-0,983z-1 karakterisztikával. Ezután 21,33 ms időtartamú Hamming-ablakkal súlyozzuk a jelet. Az ablakban lévő jelből 16-elemű melfrekvenciás kepsztrumegyüttható-vektort számolunk.

A koartikuláció jelenségének a beszédfolyamat képi ábrázolásánál legalább akkora jelentősége van, mint a hangjelek feldolgozásakor. A beszédszervek pillanatnyi állása szempontjából vannak domináns és változó fonémák. A domináns fonémák kifejezetten megszabják a száj és környezete képét, viszont a változó típusok képét a környező domináns fonémák nagyban befolyásolják. Ebből fakadóan a beszédjelből a beszédszervek képét becslő algoritmusnak a szomszédos kereteket is felölelő környezetre is tekintettel kell lennie. A siket partnerek számára a lassabb beszédtempó

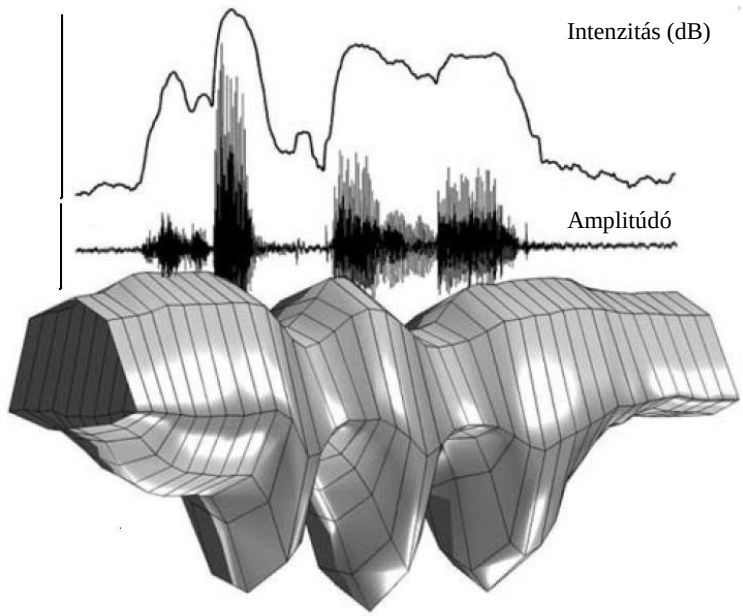


12.17. ábra. Egy hangátmenet jellemzése öt egymás utáni keret alapján

a kedvező. Gyakorlott jeltolmácsok a beszédhangok tiszta fázisú részét világosan és kiemelve képzik. Másodpercenként 5–10 beszédhangot ejtve és 40 ms hosszú elemzési időkereteket tekintve 5 elemzési ablak egyike bizonyosan ráesik a beszédfolyamat legalább egy domináns fonémájára (12.17. ábra). A neurális hálózat bemenetére tehát mindig 5 egymás utáni elemzési ablak kepsztrumvektora kerül.
A neurális hálózat. A visszacsatolt neurális hálózatot a hagyományos hibajel-visszaterjedéses módszerrel tanítottuk (Anguita–Back 1993). A hálózat három rétegben 80 csomópontot tartalmaz. A bemeneti réteg fogadja 80 ponton 5 egymás utáni időkeret 16–16 MFCC értékét. A rejtett réteg 40 csomópontot tartalmaz. A kimenő réteg 6 csomóponton szolgáltatja a 6 főkomponens súlyértékét, amelyekből előállítható a 15 jellemző pont (FP) x-y koordinátaértéke a középső időkeretben.

A tanító-adatbázis 5450 időkeretet tartalmazott. A hálózat tanítását 100 000 ciklusban végeztük. A neurális hálózatmodell a bemeneti és kimeneti változók értékeit a -1, 1 értéktartományba normálta. Az MFCC és PCA változókat mind ebbe a tartományba transzformáltuk lineárisan az MFCC vektorenergia-összetevőjének kivételével. A már betanított neurális hálózat programja igen gyorsan futtatható, mivel az egész adatbázist képviseli a hálózat súlytényező vektor, amely mindössze 3440

elemből áll. A hálózat kimeneti értékeinek valós idejű számolásához tehát egy alkalmas mobiltelefon erőforrásai elegendőek.



12.18. ábra. A 8.1–8.8 jelű jellemző pontok x-y koordinátái az idő függvényében a „szeptember" szó kiejtésekor. A felső folyamatos vonal a keretenkénti energiát ábrázolja dB-skálán, a középső görbe a hullámforma időfüggvénye. Az alsó ábrán látható felület az ajakkontúrokat mutatja.

Főkomponens-analízis. A képfelvétel minden időkeretében 15 jellemző pont írja le a száj és környékének pillanatnyi alakját. A kétdimenziós ábrázolás alapján ez egy 30 dimenziós térben egy ponttal jellemezhetô. A rendszer tanítása sokkal hatékonyabbá vált azáltal, hogy a 30 dimenziót 6-dimenziós rendszerré tömörítettük. A dimenzióredukció végrehajtására a főkomponens-analízis módszerét (Principal Component Analysis, PCA) alkalmaztuk. Ez felfogható mozgáskomponensek szerinti felbontásra. Az első 6 PCA vektort választottuk a száj és környékének leírására az alábbi egyenlet szerint

$$w_{1\ldots6} = P^{-1}B\Big|_{p_1^{-1}\times\ldots\times p_6^{-1}}, \tag{12.6}$$

ahol P jelöli a PCA vektorok (30x30) méretű sajátértékvektorát, B a 30 dimenziós vektorkészlet, c pedig a választott origó, amely a zárt ajkakkal semleges arc súlytényezőinek 0 értékét jelenti. Ez az adattömörítés mindössze 1–3% hibát eredményezett, ami az adott megjelenítő eszközön a jellemző pontok 1–2 pixeles változását eredményezi akár x, akár y koordináta szerint nézve. Ez teljesen elfogadható közelítés. Mivel a hálózat tanításához használt w súlytényező 0 értéke a semleges archoz tartozik, ezért a súlytényező előjele is egy nagyon fontos információt hordoz:

megmutatja, hogy a pont merre mozdul el. A betanított hálózat kimenő értéke egy 6-dimenziós térben jelenik meg. Ebből a jellemző pontok koordinátái a következő egyenlet segítségével határozhatók meg:

$$\overline{B}_k = (\underline{w}_k + \underline{c}) \cdot P. \tag{12.7}$$

Mivel P értékét a tanítás során határozzuk meg, ezért ez a művelet mindössze 180 szorzást igényel keretenként. A főkomponens-analízis ebben az esetben több, mint egy egyszerű mechanikus tömörítő eljárás. A PCA vektorok értékes információt hordoznak a bemondó beszédstílusáról is és a felvétel minőségéről is. A PCA vektorok – bár automatikus eljárás eredményeként adódnak – az egyes vizémák jól azonosítható megkülönböztető jegyeihez kapcsolódnak. Az állkapocs függőlegesen látszó mozgása adja a legerősebb PCA komponenst. A száj vízszintes széthúzása adja a második főkomponens nagy részét (erre a mozgásra kéri fel a fényképész az érintetteket azzal, hogy mondják: „csíííz"). A harmadik főkomponens az ajakkerekítés mértékéhez kapcsolódik. Ezek miatt állítható, hogy a PCA vektorok eredendően kapcsolódnak a vizéma megkülönbözető jegyekhez. Ezen nézőpontból a PCA vektorok dimenziósorrendje rendelkezik kiemelt jelentőséggel. Képzett jeltolmácsoknál az első néhány főkomponens tartalmazza a vizéma megkülönböztető jegyeket. Gyakorlatlan bemondóknál azt tapasztaltuk, hogy a korrektív komponensek sorrendben megelőzik a vizémákat megkülönböztető komponenseket (korrektív komponens például az érzelmet kifejező összetevő).

Beszélő fejmodell. A szabad forráskódú programmal közzétett LUCIA fejmodell némileg módosított változatát használtuk a rendszerben. Ezt más célra, az érzelmeket is kifejező vizuális beszédmodell céljára fejlesztették (Cosi et al. 2003). A LUCIA modell az MPEG-4 szabványra épült. Az eredeti fejmozgató (FAP) paraméterek vizémaalapú rendszert figyelembe véve lettek kialakítva, a szájról olvasás igényrendszerét nem vették tekintetbe a fejlesztésnél. Ezért volt szükség némi módosításra, hogy a modell képes legyen a jellemző pontkoordináták közvetlen fogadására. A közvetlen vezérlés bőrön látható pontok mozgási lehetőségeinek anatómiai alapú megkötöttségeinek finomabb figyelembevételét követelte meg.

Kísérletek. Hasznosnak bizonyultak az előzetes méréseink a rendszer tökéletesítése és az adatbázis kialakítása szempontjából. Ennek során derült ki például, hogy képzett jeltolmácsokat célszerű alkalmazni a rendszer tanításánál. Az előzetes vizsgálatok mutattak rá arra is, hogy a szavak közötti szünetekre is különös figyelmet kell fordítani. Egy küszöbszint alatti háttérzaj nem okoz gondot. Nagyobb háttérzaj óhatatlanul elkezdi mozgatni szavak között is picit a szájat, és ez nagyon megzavarja a pusztán szájról olvasásra épülő beszédfelismerést. Az előzetes vizsgálatok során a siket kísérleti személyektől összegyűlt észrevételeket, javaslatokat gondosan figyelembe vettük a rendszer tökéletesítésénél és a vizsgálati módszerek finomításánál.

Mérési módszerek és eredmények. Pusztán szájról olvasás alapján nem lehet azo-

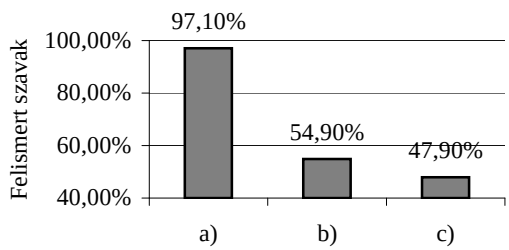
nos képzési helyű és módú fonémapárokat megkülönböztetni (például *baba-papa*). Természetes módon az észlelő személy a szövegösszefüggésre alapozva automatikusan korrigálja vagy kiegészíti a szájról leolvasott információt. Párbeszéd esetén a visszakérdezés tisztázni képes a többértelmű üzenetet. Vizsgálatainkban kizártuk a visszakérdezés lehetőségét, ezért olyan vizsgáló szöveget állítottunk össze, amely lehetőleg kizárja a kétértelműséget. A siketek az előzetes információk alapján mindig erősen leszűkített készletű lehetséges üzenetek közül egy kiválasztására összpontosítanak a szájról olvasott beszéd megértése során. Ezt a természetes mechanizmust célszerű volt követnünk a rendszer vizsgálata során is. Mindig megadtuk, hogy milyen zárt halmazból kell a lehetséges választ várniuk. A mérések során a modell teljes fejét, szájmozgását mutatta a kivetített mozgókép nagy méretű vetítővásznon. Természetesen hang nélkül. Így a töredékes hallással rendelkező vizsgálószemélyek sem hallhattak semmit a beszédjelből. A vizsgálati anyag véletlen rendben az alábbi eseteket tartalmazta:

a) a jeltolmács eredeti képfelvétele (hang nélkül),

b) a fejmodell mozgóképe, ahol a 15 vezérlő paraméter (FP) koordináták értékei jeltolmács képfelvételeiből származtak (hang nélkül),

c) a fejmodell mozgóképe, ahol a 15 vezérlő paraméter (FP) koordinátáit a rendszer a beszédjel paramétereiből számolta ki (a megjelenítés hang nélkül történt itt is).

A siket vizsgálószemélyek válaszaikat írásos formában adták meg egy előkészített űrlapon. A végső eredményeket adó vizsgálat részvevői már több alkalommal részt vettek az előzetes vizsgálatokban, így mindegyikük gyakorlott mérőszemélynek volt tekinthető. A végső vizsgálat 70 szó megértését regisztrálta és mintegy 30 percig tartott. Amikor jelezték, akkor a képfelvételt kérésükre megismételtük. A végső vizsgálatban 18 siket személy vett részt. Az eredmények a 12.19. ábrán láthatók.



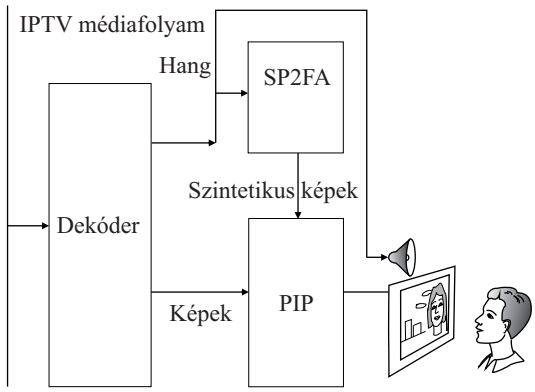
12.19. ábra. A helyesen megértett szavak aránya a) jeltolmács képfelvétele alapján, b) jeltolmács FP koordinátáival vezérelt fejmodell képe alapján, c) beszédjelből számolt FP koordinátákkal vezérelt fejmodell képe alapján

Értékelés. A jeltolmácsok eredeti képfelvételei alapján a szavak szájról olvasása körülbelül 3% felismerési hibát eredményezett. A 15 FP pont koordinátáival vezérelt fejmodell, ha a vezérlő paramétereket közvetlenül a jeltolmács képfelvételein megje-

lölt pontok koordinátáiból származtattuk, akkor 42% felismerési hibát adott. A méréseket követő megbeszéléseken a vizsgálószemélyek olyan szóbeli megjegyzéseket tettek, hogy hiányzott bizonyos helyzetekben a modellből a nyelv képe és néha a szájtól távolabbi részek mozgása is. Emiatt a fejmodell árnyaltabb vezérlése esetleg megfontolandó. A pusztán hangelemzésből számolt vezérlő paraméterekkel vezérelt fejmodell alapján mért szóérthetőség az előző esethez képest csak 7%-kal csökkent. Ez mutatja a rendszer alapvető eredményét, azaz annak igazolt tényét, hogy a hangjelből számolt vezérlő paraméterekkel jól megközelíthető a képjelből származtatott paraméterekkel vezérelt modell felismerési aránya. Mindez épít a siket személyek kifinomult felismerési képességeire, és kizárólag erre az esetre érvényes az előző megállapítás.

További alkalmazási lehetőség: IPTV speciális hanginformáció siketek számára. A siketek számára létkérdés a televízióműsorokban is a szájmozgás követése az események megértése szempontjából. A leggyakoribb probléma Magyarországon a szinkronizált filmek tömege. A színész szája az eredeti (legtöbbször angol) nyelv rendszere szerint mozog, amiről a magyar siketek nem tudnak leolvasni semmit. A felirat sem tökéletes megoldás, mert nagyon leköti a néző figyelmét, s ha elég részletes, már alig marad figyelem a film élvezetére. További esetek is problémát okoznak a siketeknek: a politikai, a magazin- és a hírműsorokban gyakran betétrészletek láthatók alámondott hanginformációval, amelyek a háttérhang nélkül érthetetlenek. A természetfilmekben, városok, tájak ismertetését tartalmazó műsorokban narrátor mondja az alapvető információt, amelyet a képek, mozgóképek színesítenek, tesznek élvezetessé. Ezek üzenetének lényege nem érheti el a siket vagy erősen nagyothalló nézőket.

Az általunk kidolgozott, beszédjelet szájmozgássá átalakító rendszer alkalmas arra, hogy egy IPTV médiafolyamban (streamben) érkező televízióműsor műsorhangjának felhasználásával egy szintetikusan előállított emberifej-modell száját a beérkező (akár szinkronizált vagy narrátor) hangnak megfelelően mozgassa. A szintetikusan előállított képet hozzáadja az eredeti műsor képtartalmához, és azt együttesen jeleníti meg a felhasználó képernyőjén (kép a képben, PIP). A rendszer blokkjait a 12.20. ábra ismerteti, ahol az SP2FA egység azonos a 12.16a ábrán bemutatott rendszerrel (SPeech to FAcial Animation). A megvalósítás Direct Show keretben volt a legcélszerűbb, amelyet média kezeléséhez fejlesztett a Microsoft. A keretrendszer a médiafeldolgozásban már jól ismert, hálózatba szervezhető alapvető funkciókra épül. Kísérleti eredményeink igazolták, hogy lehetséges beszédjelből közvetlenül szájmozgást leíró jellemzők származtatása olyan pontossággal, ami lehetővé teszi siket személyek számára a beszéd gyakorlati hasznosságú megértését. Erre alapozva segédeszköz készíthető siketek számára, hogy megértsék csak telefonon vett beszédjelből a beszédüzenetet. A rendszer alapelemei olyan számítástechnikai erőforrással megvalósíthatók, amely rendelkezésre áll a mai legfejlettebb mobiltelefonokban. A fejmodell további finomításától reméljük a teljes rendszer olyan fejlődését, amely



12.20. ábra. IPTV speciális hanginformációt siketek számára olvashatóvá konvertáló rendszer alapelemei

révén elérhető a 20% alatti vizuális felismerési hiba, amely szint minden szempontból elfogadható értéket jelent. Emlékeztetünk arra, hogy a mobiltelefonok áldásaiból gyakorlatilag kirekesztett siketek közösségének ez forradalmi előrelépést jelentene jelenleg még fennálló akadályaik leküzdésében.

A fejlesztések során több érdekes új tudományos eredmény is született. A hangi és képi jeltartalom kölcsönös információja alapján mérhető, hogy egy fél fonéma időtartamával is előbbre járhat a száj mozgása, mint a beszédszerveinkkel keltett hang.

12.9. Beszédtanítás és beszédtechnológia

Vicsi Klára

A beszédterápiával, és a beszédkutatással foglalkozó szakembereket már sok évtizeddel ezelőtt foglalkoztatta, hogy miként lehet a technika vívmányait a beszédterápiában felhasználni. A megvalósított eszközök bonyolultak és nehezen használhatók voltak. Ahogy a számítástechnika és a beszédtechnológia fejlődik, a beszédoktató rendszerek megoldási lehetőségei nőnek. Az új kutatásokat a beszédfelismerés, a beszédszintézis, a beszédelemzés, a vizuális megjelenítés legújabb kutatási eredményei támogatják (Vicsi 2004). A beszédhibás vagy hallássérült emberek beszédoktatásán kívül egy egészen új irányzat annak a vizsgálata, hogy az idegennyelv-oktatásban is lehetne-e hasznosítani a számítógépes támogatás adta lehetőségeket (Computer Aided Language Learning, CALL).

Fontos az is, hogy a gyors technikai fejlődés mellett figyelembe vegyük a fonetikai, fonológiai, oktatási szempontokat, a beszédfejlesztés különböző lépéseit, a felhasználók károsodási mértékét vagy az életkor szerinti szellemi képességeket. Sajnos

az utóbbi években kialakított beszédoktató rendszerek legtöbbje csak átveszi változtatás nélkül a legújabb beszédtechnológiai eljárásokat és nem alakítja azokat a speciális alkalmazáshoz, feladathoz. Erre példa a gépi beszédfelismerők széles körű alkalmazása a kiejtés helyességének megítélésére. A helytelen kiejtést azonban ezek a rendszerek nem képesek detektálni, holott éppen az lenne a feladatuk (pongyola, renyhe ejtés, hadarás vagy esetleges hangkihagyás).

A beszédfelismerők használatán alapuló automatikus visszajelzés hatékonyságát illetően a tanárok véleménye sem igazán pozitív (Wallace 1998). Tapasztalataik szerint vagy nem megfelelőek ezek az automatikus ítéletek, vagy pedig nem elég érzékenyek a gépi megoldások az apróbb különbségek észrevételéhez, ami félrevezeti a tanulókat. Ezáltal a felhasználók rosszabb eredményeket érhetnek el, mint az automatikus visszajelzés használata nélkül. Ilyen automatikus visszajelzésen alapuló kiértékeléses eljárással dolgozik az ISTRA és ISLE (Interactive Spoken Language Education) nyelvoktató rendszer. E program is a fonémaalapú rejtett Markov-modelleket alkalmazó beszédfelismerési technológiát használja fel a kiejtés megítélésére. Hasonlóan működnek továbbá a TALK TO ME, TELL ME MORE angol nyelvoktató programok (`http://www.auralog.com/us/schools.html`) (Nouza 1999). E programok használhatóságát segítené, ha valamilyen más, például vizuális visszacsatolást is alkalmaznának. Néhány program hullámforma-megjelenítést ugyan használ, és a fonetikus, akusztikus szakember el is igazodik a hullámformán, de egy gyermek biztosan nem.

A kialakított rendszerek egyik csoportjánál magát az artikulációt mutatják be a tanulóknak a beszéd közbeni artikulációs mozgás grafikai megjelenítésével. Közvetlenül a beszédképző szervek pontos beállítására teszik a hangsúlyt. Ez az úgynevezett folyamatorientált megközelítés. Az előállított szintetikus arc artikulál a beszédhanggal szinkronban. Az artikuláció azonos idejű modellezése technológiailag nehéz feladat (Hardcastle et al. 1999, Gibbon–Hardcastle 1998). A paraméterekkel vezérelt vizuális beszédszintézis az arc 3D-s poligonális modelljén alapszik (Massaro 1998b, Cole et al. 1998). Ezek a rendszerek a beszéd vizuális képsorozatának jellemzéséhez, megjelenítéséhez nyomon követik és meghatározzák a beszélő szájmozgását (lásd a 9.12. fejezetet). Egy kedves beszélő fej van beépítve a Massaro és munkatársai (Massaro 1998b) által fejlesztett oktató rendszerbe, melynek neve CSLU Speech Toolkit. Ez egy kutatói segédeszköz, melynek honlapja a következő címen található: `http://cslu.cse.ogi.edu/toolkit`. Ezek az artikulációs mozgást modellező oktató rendszerek arra a vizuális visszacsatolásra építenek, ami a természetes beszédkommunikációban is jelen van. Hallássérült gyermekek esetén problémát jelent, hogy a belső hangképző szervek pozícióját nem lehet látni a képernyőn. A fejlesztők igyekeznek a modellekben a rejtett beszédszerveket is láthatóvá tenni, de egyelőre ezek a rajzok még elég riasztóak.

A beszédoktató rendszerek másik csoportjánál különböző akusztikai paramétereket jelenítenek meg. Tehát nem az artikulációs szervek beállítását hangsúlyozzák,

hanem azt az akusztikai produktumot, amit a beszélő előállít és felhasznál a kommunikációban. Ez az úgynevezett produktumorientált megközelítés. Rögzítik a beszédjelet, visszajátsszák, és közben megjelenítik a képernyőn valamely formában. Ezeknek az eszközöknek vagy módszereknek a sikeressége azon múlik, hogy milyen paramétereket mérnek, és milyen megoldást vezetnek be a mért paraméterek vizuális visszacsatolására. Ebbe a csoportba tartozik az IBM Speech Viewer programja, amely kifejezetten hallássérült gyermekek számára készült. A megjelenített képek vagy ábrák kétfélék. Egy részük a beszéd fizikai megjelenítését használja valamilyen formában. Azonban ezeket az ábrákat a fonetikus szakértő értelmezni tudja, de a gyermek nem.

A 12.21. ábrán látható, gyermekeknek érthető képi megjelenítés viszont valójában inkább egy játék. Például az alma leesik a fáról, ha a kiejtés helyes. A program

a b

12.21. ábra. Az [u] hang helyes a) és helytelen kiejtésben b) (IBM Speech Viewer)

használatakor problémát jelent, hogy a gyermek nem tudja felmérni, hogy milyen a jó kiejtés, és azt sem, hogy az ő kiejtése milyen messze van az optimálistól, mit kell csinálnia, hogy az adott hang képzése jó legyen. Egy oktatásra jól használható rendszernél a megjelenített hangkép érdekes, de ugyanakkor fonetikailag helyes kell, hogy legyen. Meg kell adni az oktatáshoz azt az információt, hogy az egyik kiejtés miért helyes, a másik miért helytelen. Rá kell vezetni a gyemeket a helyes hangképzés és artikuláció kialakítására.

Magyarországi kutatások. Az 1950-ben Groningenben rendezett Siketnémaügyi Nemzetközi Kongresszuson terjesztették elő az oszcillográf alkalmazását siketek hangosbeszéd-tanítására. Az alacsony frekvenciás, hosszú utánvilágítású oszcilloszkóp 12 cm átmérőjű kerek képernyőjén rövid ideig megmaradt a folyamatosan ejthető a magánhangzók jellemző hangnyomás-időfüggvény görbéje, ez a megjelenítés a beszédritmus és a beszédtempó vizuális ábrázolását is szolgálta. Több éves tudományos kutatómunka eredményeként – az előzőekben tárgyalt tapasztalatokból kiindulva – Gáspár Árpád, a Gyógypedagógiai Tanárképző Főiskola Hallássérültek Pedagógiája Tanszékének tanára összeállította a Chordoscop műszercsoportot. Lényegében hangfrekvenciás erősítőből, alacsony frekvenciás oszcilloszkópból, száj- és gégemikrofonból, valamint fejhallgatóból állt. A Chordoscop műszercsoport a rezgéseket kristálymikrofon segítségével feszültséggé alakítja át, majd megfelelő erősítés után egy alacsony frekvenciás hosszú utánvilágítású oszcilloszkópra, illetve süllyesz-

tett képernyőre viszi fel, ahol a rezgéskép látható (vizuális) formába jelenik meg az idő függvényében. A tanár által adott példa mint norma a tanuló számára leutánzandó feladatként jelentkezik. A rezgéskép fotórekorder segítségével filmre rögzíthető. A szájmikrofon alkalmazása lehetővé tette, hogy a tanár és a tanuló a megfelelő levegőáramoltatást regisztrálja. A száj-, illetve a gégemikrofon alkalmazása azt is lehetővé tette, hogy a hangszalagok működtetését is kontrollálják (Gáspár 1971).

A Szegedi József Attila Tudományegyetem (SZTE) Informatikai Tanszékcsoportja, együttműködve a Hallássérültek Szegedi Intézetének igazgatójával, számítógépes segédeszközt fejlesztett ki, a Beszédmester szoftvercsomagot (Paczolay et al. 2004). A program azzal a céllal készült, hogy segítse az iskolások betűtanulását, az olvasást, illetve az otthoni gyakorlást, másrészt, hogy a hallássérült, siket és logopédiai kezelésben részesülő gyermekek számára kínáljon játékos számítógépes programot. Előnye, hogy játékosan, színes képekkel, a számítógép motivációs erejét felhasználva próbálja meg a kisiskolásokat az olvasás rejtelmeire megtanítani, az olvasni tanuló kisiskolás olvasástanulását, olvasásterápiáját, fejlesztését szolgálja. A program és didaktikája letölthető a www.inf.u-szeged.hu oldalról.

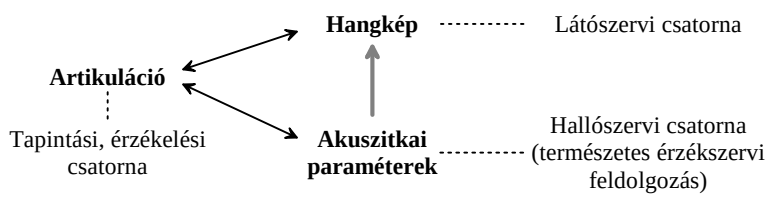
12.9.1. Beszédoktató varázsdoboz

A többnyelvű, és a hallási és látási visszacsatolás együttesével kialakított, tudományosan megalapozott beszédoktató és rehabilitációs rendszer a Beszédkorrektor (SPEech COrrector, SPECO) 1999 és 2001 között készült el magyar, angol, német, svéd és szlovén nyelvre (Vicsi et al. 1999), és ma is több országban használják. A Beszédkorrektor olyan kétszenzoros beszédoktató módszerre épül, amelynél a hallási visszacsatoláson kívül a vizuális visszacsatolás is segíti a tanulót a beszédterápiában. A beszéd akusztikai elemeit vizuális (játékos) képekkel jeleníti meg (Vicsi 2005). A program segítséget nyújt éphalló beszédhibás, valamint nagyothalló gyermekek és felnőttek helyes beszédképzésének kialakításában. A kiejtés és beszédfejlesztés során segíti a gyermekek artikulációs bázisának megteremtését, a magyar beszédhangok helyes kiejtésének kialakítását, rögzítését és agyi automatizálását. Lehetőséget nyújt továbbá az alapvető, általános beszédjellemzők helyes kialakítására, gyakorlására is. Ilyen jellemzők a hangosság, a hangmagasság, a ritmus, a hanglejtés és a hangszín. A készülék a gyermek által kiejtett beszédet elemzi és színes képeken megmutatja a képernyőn, hogy helyes volt-e az artikuláció (azaz a kiejtés). A Beszédkorrektor magyar nyelvű verziója a magyar gyermekek számára a VARÁZSDOBOZ-ként vált ismertté. Használata Magyarországon széles körben elterjedt. Logopédusok, szurdopedagógusok és foniáter orvosok munkáját teszi hatékonyabbá és változatosabbá, valamint a program lehetővé teszi, hogy a gyermekek otthon is önállóan gyakorolhassanak, játékos módon. Alkalmazási területek:

- pösze beszéd korrekciója
- hallássérültek beszédfejlesztése
- megkésett beszéd terápiája
- implantált betegek rehabilitációja
- egyéb beszédsérülések kezelése

Alapkoncepció. A gyermekeket a hangképző szervek helyes beállítására a vizuális visszacsatolás segíti. A beszéd akusztikai tulajdonságainak képi megjelenítése játékos formában tereli őket a helyes hangképzés felé. Az eljárás a hallási és látási egyidejű visszacsatoláson alapul. A gyermek hívóképeken kapja meg a feladatot, hogy mit kell kiejtenie.

A hagyományos beszédterápia közvetlenül a beszédképző szervek pontos beállítására összpontosít. A Beszédkorrektor nem az artikulációs szervek beállítását hangsúlyozza, hanem az akusztikai produktumot, a végeredményt mutatja meg. A természetes beszédtanulás során a gyermekek szintén az utóbbi megközelítést használják, vagyis az előállított beszéd akusztikai tulajdonságai alapján, saját akusztikai visszacsatolásuk segítségével találják meg a hangképző szerveik optimális beállítását. Minden beszédterápia lényeges része a modelltanulás, amelynél a referenciabeszélő (például a terapeuta) beszéde utánzásának van fontos szerepe. Nemcsak a jó beszédhang funkcionális képzésmódjának utánzásáról van szó, hanem hallószervi, akusztikus utánzásról is. Az utánzáshoz használt természetes érzékszervi csatorna elsősorban a hallás. A hatás sokkal intenzívebbé tehető a látószervi csatorna bekapcsolásával (Campbell et al. 1998). A Beszédkorrektor módszerében a visszacsatolás



12.22. ábra. A beszédoktatásba bevont érzékszervi csatornák

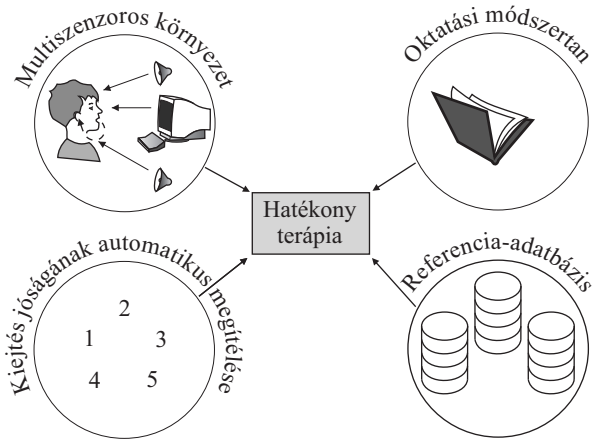
mind a két csatornán biztosított. A látószervi csatornán keresztül úgy, hogy az elhangzott beszédet vizuálisan, hangképek formájában jeleníti meg a rendszer, alapozva arra a tényre, hogy az artikuláció és a megjelenített beszédparaméterek, azaz a hangképek között egyértelmű kapcsolat van (12.22. ábra).

Az oktatórendszer négy különböző építőelem harmonikus, együttes alkalmazására épül (12.23. ábra).

1. A hallási és látási visszacsatoláson alapuló beszédterápiás környezet
2. Részletes oktatási módszertan az új környezetre építve

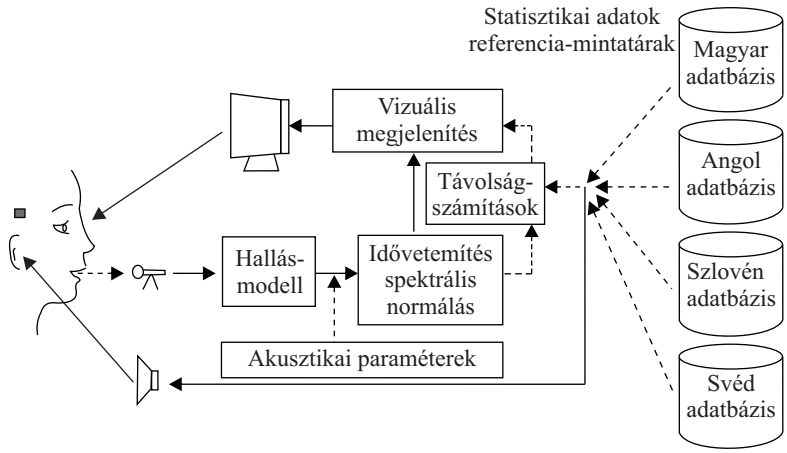
3. Megfelelő mennyiségű beszédmintát tartalmazó referencia-adatbázis, a gyakorlóanyag összeállításához, a fiziológiás szenzomotoros csatolások kiépítésére
4. Automatikus értékelés a gyakorló személy kiejtési jóságának megítélésére

Ezek az összehangoltan tervezett és felhasznált elemek kölcsönösen felerősítik egymás hatását a beszédterápiában. Ezt egyértelműen alátámasztják a Beszédkorrektorral szerzett közvetlen tapasztalatok (Vicsi et al. 1999, Váry 2001). A Beszédkorrek-



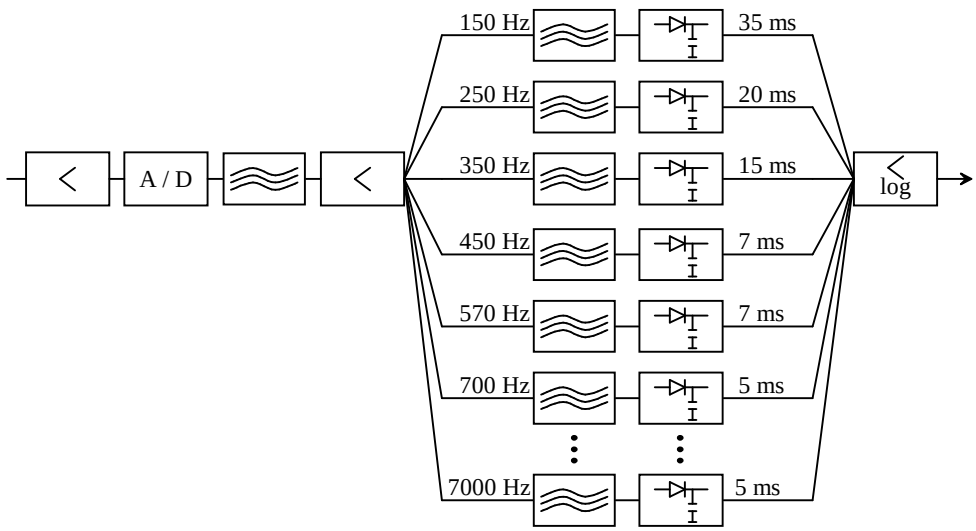
12.23. ábra. A Beszédkorrektor rendszer építőelemei

tor blokkvázlata a 12.24. ábrán látható. Az akusztikai előfeldolgozás, a vetemítő, távolságszámító algoritmusok és a képi megjelenítés nyelvfüggetlenek. A referencia hang- és képszótárak, valamint az adott nyelvre vonatkozó, nagy adatbázisból nyert statisztikai adatok nyelvfüggőek, azaz minden nyelvre külön kell őket létrehozni.
Akusztikai előfeldolgozás. A gyermek által bemondott hang, hangsor akusztikai előfeldolgozása egyszerűsített hallásmodell alapján történik, amely az emberi hallórendszer perifériás szintű feldolgozási mechanizmusát modellezi (lásd a 3.2. fejezet) (Zwicker 1982, Zwicker–Terhardt 1980). A frekvenciaelemzést 20 aszimmetrikus, az elfedési görbéknek megfelelő meredekségű szűrő (bark-sáv) végezi 80 Hz és 8000 Hz között. A szűrőkből kijövő jel egyenirányítás, rövid idejű átlagolás után logaritmizálásra kerül (12.25. ábra). A rövid idejű átlagolás időállandója 1 kHz-ig a szűrő középfrekvenciájának megfelelő periódusidő ötszöröse, 1 kHz felett állandó, 5 ms. A modell kimenő paramétereit a továbbiakban hallási spektrum, illetve hallási spektrogram névem adjuk meg, mivel a szinképelemzést a perifériás hallási rendszer frekvenciafelbontása szerint ábrázoljuk. A gyermek számára így könnyebb a kiértékelés, mint a hagyományos spektrális módszrekkel. A modell kimenő paramétereire alapított vizuális ábrázolás adja meg a tanulónak azt a lehetőséget, hogy a beszédet a vizuális látvány alapján feldolgozza, értelmezze, és ezután a szükséges kor-



12.24. ábra. A Beszédkorrektor mérő- és feldolgozórendszerének blokkvázlata

rekciót (ha kell) megtegye (másképp artikuláljon az új próbálkozásnál). A mintavétel: 22 025 Hz, 16 bit, a dinamikatartomány 50 dB. Az alapfrekvencia-meghatározás a rendszerben a rövid idejű AMDF-módszert használja (7.1.3. fejezet).



12.25. ábra. A mérőrendszer működési blokkvázlata

Vizuális ábrázolás. A gyermek vizuálisan hatásosan tud feldolgozni képi információt. Ezért hangképek formájában adjuk tudtára, hogy milyen a beszéde. A hangképek megformálásánál az volt a cél, hogy olyan információt kapjon a gyakorló személy a hangról, amely rávezeti őt a hang helyes megformálására. A kép megformálását

részletes vizsgálatok előzték meg (képi felbontás, színek és paraméterek összhangja a hangosság, a spektrális tartalom és az alapfrekvenciaváltozás esetében). A cél az volt, hogy a vizuális megjelenítés minél informatívabb legyen. A következő kérdésekre kellett választ keresni. Hogy lehet a gyermek figyelmét a spektrum maximális energiájú pontjaira irányítani? Hogyan lehet felismerni, ha rossz ritmusban ejtik ki a hangokat?

Néhány gyakorlatban a referenciamintával való összehasonlításon alapul a tanulás. Ilyenkor az ablak felső részében a referenciamintát mutatjuk, az alsó részben pedig megjelenik a saját kiejtésből készített ugyanolyan szerkezetű kép. Így össze lehet hasonlítani kiejtést a jó referenciamintával. A megjelenített hangképek a következők: a hangintenzitásszint időbeli változása, az alapfrekvencia (hanglejtés, azaz a beszéddallam), a hallási spektrum, a hangspektrogram, a spektrális különbségek és a ritmikai képletek.

12.9.1.1. Adatbázisok és modellezés

A kialakított multiszenzoros beszédterápiás környezet egyben adatbázis-szerkesztő is, amely lehetőséget ad a különböző beszédoktatási feladatok (beszédjavítás, nyelvtanulás) ellátásához szükséges szöveg, hang- és képadatbázisok előállítására (különböző nyelvekre). A beszédhibás gyermekek beszédjavítására létrehozott rendszerben a következő adatbázisok szerepelnek.

Szövegadatbázisok. A gyakorlandó hanganyag szövegadatbázisa nemzeti betűkészlettel, illetve SAMPA (`http://www.phon.ucl.ac.uk/home/sampa/hungaria.htm`) hangszimbólum-jelölésekkel (Vicsi–Víg 1998b) készült (4.2. fejezet).

Referencia-beszédminták tára. Minden gyakorlandó anyag jó kiejtéssel tárolva van a rendszerben (hang, hangkapcsolatok, szavak, mondatok valamint minimálpárok például: *szár - sár, szár - zár*). Ezeket a szótárban lévő helyes kiejtésű beszédmintákat nevezzük referencia-beszédmintáknak, a hozzájuk rendelt beszédkészletet pedig referencia-beszédmintatárnak. A referenciamintákat egy gondosan kiválasztott személy mondta be, akinek a beszédje szép, érthető. A tanulónak a terápia során ezekhez a referencia-beszédmintákhoz hasonló beszédet kell előállítania. Külön adatbázis készült a magyar Beszédkorrektornál a gyermek és a felnőtt nő, valamint férfi beszélőkre. Egyéb nyelvekre csak gyermekadatbázis készült. A felvételek süket, vagy csendes helyiségben, kis zajú hangkártyával készültek. A jel-zaj viszony minimum 40 dB volt minden nyelv esetén.

Referencia-hangképtár. Minden referencia-beszémintához tartozik valamilyen képi megjelenítés is (háttérkép). A magyarázó háttérképek létrehozásához minden nyelvre külön-külön gyermekbeszéd-adatbázist hoztak létre. Ezen adatbázisok statisztikai vizsgálatával készültek el az adott nyelv beszédhangjainak spektrális modelljei (lásd később). Ezeknek a spektrális modelleknek az adatai mint referencia-háttérképek ke-

rülnek rajzos formában megjelenítésre minden nyelv minden gyakorlandó beszédhangjára egyrészről izolált ejtésben mint az adott hang referenciaspektruma, másrészről folyamatos beszédben mint az adott hang referencia-spektrogramja. Ezeket a képeket a hangképtárban helyeztük el.

Gyermekbeszéd-adatbázis. Az adott nyelv beszédhangjainak statisztikai vizsgálatához külön, sok bemondóval készített adatbázisra volt szükség. Meg kellett határozni a helyes kiejtésű beszéd szórását, szélső értékeit. A 4 különböző nyelvű adatbázis szöveganyaga tartalmazta a gyakorlásra külön kiemelt beszédhangokat, azaz a réshangokat, affrikátákat és a magánhangzókat izolált ejtésben, hangkapcsolatokban, szavakban és mondatokban, különböző pozícióban. A bemondók ezeket olvasták fel. Egy beszélő 10–15 percig olvasott. A bemondók életkora 5 és 10 év között változott. Magyar nyelvre 72 beszélő mondott be egyenként 80 szót és 29 mondatot. Az 5–6 éves gyerekeket óvodából, a 7–10 éveseket egy általános iskola alsó tagozatából kérték fel (10 ötéves, 10 hatéves, 18 hét-, 16 nyolc-, 6 kilenc- és 12 tízéves). Azoknál a gyerekeknél, akik nem tudtak még olvasni, a felvételkészítő előmondta a felveendő anyagot.

A nyelv beszédhangjainak spektrális modelljei. A rendszerben fontos szerepe van a fonémák statisztikus színképi modelljeinek, amelyeket az alábbi paraméterekkel jellemezzük. Legyen egy fonéma k-adik helyes kiejtésében lévő i-edik időkerethez (például 10 ms hosszúságú) tartozó hallási spektrumvektor a_{ki}. Ennek f-edik eleme az f-edik bark sávban a k-adik kiejtés i-edik időkeretében mért energia (intenzitás): a_{fki} legyen $1 \leq k \leq K$, $l \leq i \leq I$, $1 \leq f \leq F$. (Megjegyezzük, hogy F értéke tipikusan 19). A statisztikus spektrális modellben

- a beszédhang átlagos hallási spektrumvektora (átlagvektora)

$$a = \frac{\sum_i \sum_k a_{ki}}{K + I} \tag{12.8}$$

melynek elemei $(a_1, a_2, ..., a_f, ..., a_F)$,

- a beszédhang hallási spektrumának kiterjesztett maximumvektora (röviden: max)

$$a_{max} = \max_{k,i} a_{ki} \tag{12.9}$$

- a beszédhang hallási spektrumának kiterjesztett minimumvektora (röviden: min)

$$a_{min} = \min_{k,i} a_{ki} \tag{12.10}$$

- a beszédhang hallási spektrumának átlagos pozitíveltérés-vektora (röviden: poz)

$$a_{poz} = \frac{\sum_{k,i} (a_{ki} - a)}{K_1 + I_1} + a \tag{12.11}$$

azon k és i indexekre, amelyekre $a_{ki} > a$ és amelyek számossága K_1, illetve I_1.

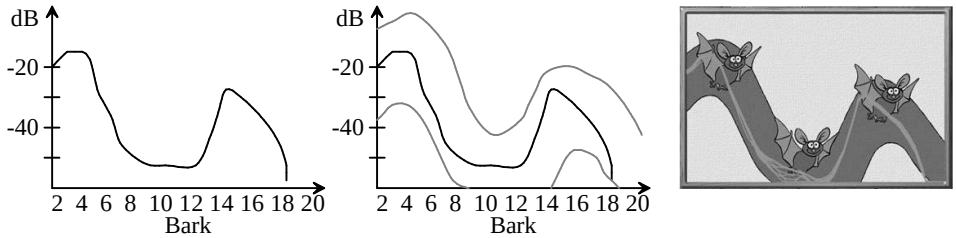
- a beszédhang hallási spektrumának átlagos negatív vektora (röviden: neg)

$$a_{neg} = \frac{\sum_{k,i}(a - a_{ki})}{K_2 + I_2} + a \tag{12.12}$$

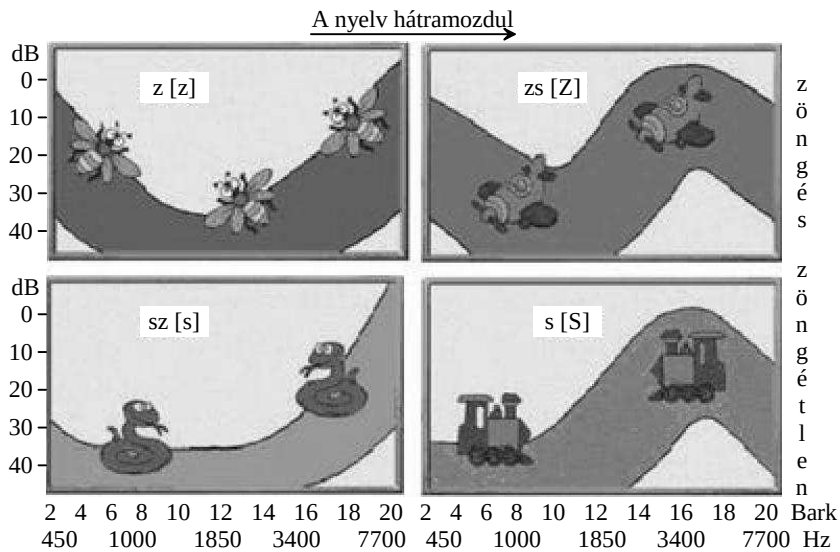
azon k és i indexekre, amelyekre $a > a_{ki}$ és amelyek számossága K_2, illetve I_2.
(Megjegyzés: az a_{ki} mennyiségek hisztogramja az a-ra mint átlagra nem szimmetrikus, ezért célszerű a *pos* és a *neg* használata, például a szórás helyett). A táblázatokban legtöbbször a mennyiségek szintértékre számítva, dB-ben kifejezve jelennek meg (pédául $a^{dB} = 10\lg a$).

12.9.1.2. Képi megjelenítés

A rendszer képi megjelenítés formájában adja vissza a hangrezésből kinyert parmétereket színes képekben. Négy fő kategória működik a rendszerben.
A hallási spektrum. A képek vízszintes tengelyén a frekvenciasávok (1–20 bark), a függőleges tengelyen az intenzitásszintek (50 dB-es dinamikatartományban) van feltüntetve. Méréskor a szűrőkimeneti szintértékekhez egy simított görbét illesztettünk. Minden 10 msec-os időkeretben új spektrumgörbe jelenik meg. A megfelelő vizuális ábrázolás érdekében minden spektrumgörbe a korábbi 9-cel együtt jelenik meg. Példaként a 12.26. ábrán bemutatjuk egy kitartott magyar [e:] hang (gyermekhangon) hallási spektrumát (bal kép). A méréskor a mérési pontokra illesztett görbe, a referencia-hangképtárból kiválasztott, az éppen gyakorlandó beszédhang spektrális modelljének spektrum típusú háttérképe jelenik meg az adott beszédhangra jellemző a_{max} és az a_{min} értékekre illesztett görbék feltüntetésével, amelyek a megengedett spektrális ingadozási sávot mutatják a felhasználó számára (középső kép). A 12.26. ábra bal oldali és középső képe helyett a jobb oldalit látja a gyermek, így számára érthetővé válik a kép és a hang összerendelése. A helyes kiejtéskor a két görbe közötti területen kell a mérési pontoknak lenniük. A megjelenített állatoknak is figyelemfelhívó szerepük van, a mérési pontoknak ugyanis érinteniük kell ezeket az állatokat. Az állatok a hallási spektrum jellemző pontjait (a jellemző energiamaximum- és minimumhelyeket) hangsúlyozzák. A 12.27. ábrán a vizsgált sziszegő hangok spektrális modelljének spektrum típusú háttérképeit mutatjuk be. Helyes ejtés esetén a szűrő kimeneti energiaértékeinek az ábrázolt sávon belül kell maradni.
A hallási spektrogram ábrázolása. A beszéd időbeli változását a hallási spektrogramból készített játékos színes képeken lehet nyomon követni. A kép itt két részből tevődik össze, a felső részében a referenciát mutatjuk, az alsó részében a gyermek által ejtett hangsorból kiszámított képet lehet látni A két kép vizuális összehasonlításával megállapítható, hogy a tanuló milyen mértékben ejtette helyesen a hangsor



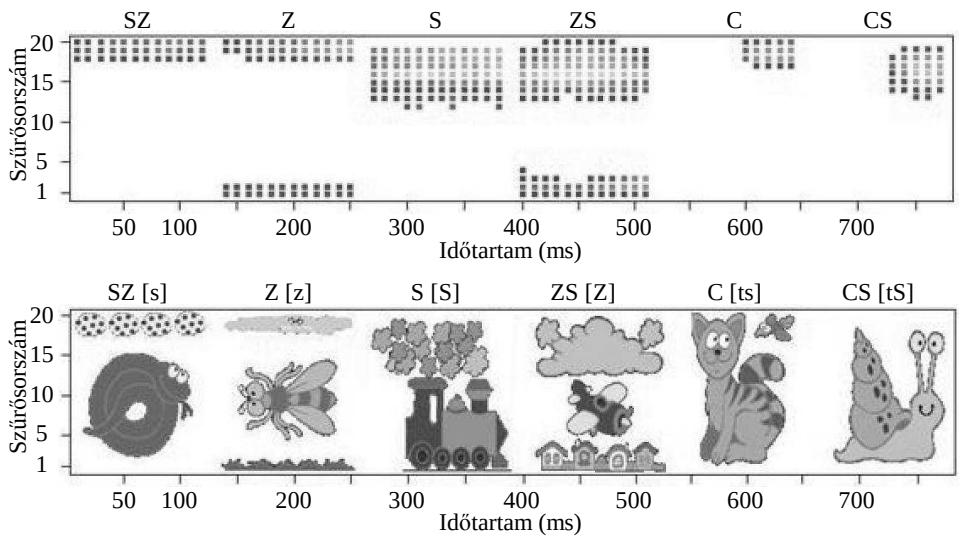
12.26. ábra. Egy gyermek által bemondott kitartott [eː] hang mért görbéje (bal); a helyesen képzett hang megengedett hallási spektrumának ingadozási sávja, benne az ejtett hang egyedi görbéjével (középen); a gyermekek számára érthető képforma az [eː] hang gyakorlására (akkor helyes a kiejtés, ha az állatkát a sötétszürke hullámon belül érinti a saját ejtésből megjelenő görbe)



12.27. ábra. Képek a magyar zöngés és zöngétlen réshangok gyakorlására

hangjait. A vízszintes tengely az időt, a függőleges a 20 szűrő (azaz a bark-sávok) frekvenciasávját mutatja. Szótagok, szavak és mondatok gyakorlásánál a szűrőkben mért intenzitásszintek a frekvencia függvényében jelennek meg színes kép formájában 10 msec-os időosztásban. A színek 6 dB-enként változnak, a világos árnyalatok a nagyobb energiát jelentik. A zörej jellegű hangok a piros szín árnyalataival, a zöngés hangok a kék különböző árnyalataival jelennek meg. A gyakorló a szavakat, mondatokat más ritmusban ejti ki, mint a referenciát bemondó személy, tehát a beszédhangok hossza különbözik a két mintában. A hatásos vizuális összehasonlításhoz a 9.4. fejezetben bemutatott szimmetrikus dinamikus idővetemítést alkalmaztuk (Sakoe–Chiba 1978). A háttérképpel és állatfigurákkal kiegészített hallási spektrogram típu-

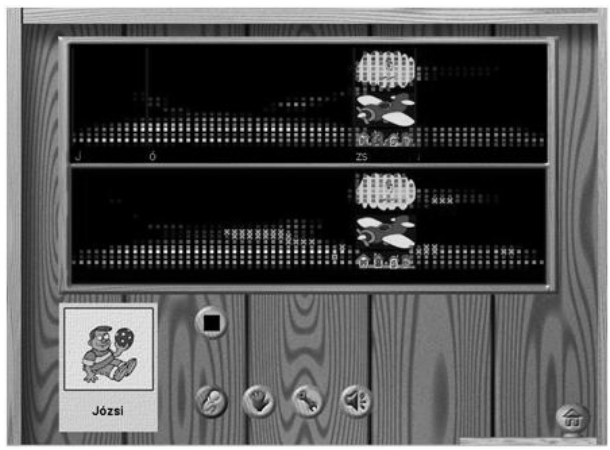
sú ábrázolás használhatósága döntő mértékben a megjelenítési küszöb helyes megválasztásán múlik. Ez sok kisérleti munkát igényelt. A sziszegő hangok esetében az egyes szűrősávokban a kvázistacionárius részben mért energiaértékeket a 12.28. ábrán bemutatott példaképpeni hallási spektrogram típusú ábrázolásban mutatjuk be, valamint az annak értelmezését segítő háttérrajzokat. Helyes ejtésben a [s] hang esetén a kígyó tojásait kell elrejteni a spektrumpontokkal, a kígyót tisztán kell hagyni; [z] hang esetében a felhőt és a füvet kell elrejteni, de a méhecskét tisztán kell hagyni; az [ʃ] képzésekor a vonat füstjét kell elrejteni, de a vonatot tisztán kell hagyni; a [ʒ] hangnál a felhőt, a házakat kell elrejteni, a repülőt tisztán kell hagyni; a [t͡s]-nél, [t͡ʃ]-nél a cicát és a csiga testét, házát tisztán kell hagyni, a madarat, a csiga szemeit viszont el kell rejteni. Egy példa az [s] hang rossz, laterális ejtésének a megjelenítése a 12.29. ábrán látható. A hangkörnyezet jelentősen befolyásolja a spektrumot. Az [u] hang például a sziszegők jellegzetes energiamaximumait a mélyebb frekvenciák felé mozdítja, az [i] hang viszont felfelé, a magasabb frekvenciák irányába húzza el az energiamaximumot (lásd 5.1.3. fejezet, 5.23. ábra). A VARÁZSDOBOZ lehetősé-



12.28. ábra. A spektrális modellek spektrogram típusú megjelenítése (fent) a megfelelő játékos háttérképekkel (lent), magyar sziszegőkre, folyamatos szövegben történő gyakorláshoz

get ad arra, hogy a gyakorlandó beszédhangot a hangkapcsolat vagy a szótag elején, belsejében, végén minden magánhangzóval összekötve gyakoroljuk a hívóképeken megadott szövegvariációkban. A hívóképek sorrendben először a könnyebb, majd az egyre nehezebb hangkapcsolatokat mutatják, amit célszerű követni, de a sorrend szabadon választható. A hallási spektrogram aktuális megjelenítésére a hívóképpel, kezelőgombokkal együtt a 12.30. ábrán látható példa.

12.29. ábra. Az *iszi* szótagokban a réshangot vizsgáljuk. Fent: referenciakiejtés, lent a réshang hibás ejtésben



12.30. ábra. Magyar [ʒ] hang gyakorlása a *Józsi* szóban. Fent: a referenciabemondás hallási spektrogramja, lent: a gyermek bemondása, balra lent a hívókép

Intenzitásváltozási kép. Az intenzitásszinteket 50 dB-es dinamikatartományban méri a rendszer. Az integrációs idő 10 msec. A megjelenített színes oszlopsor a beszédintenzitás 10-es alapú logaritmusának időbeli változását követi. Jól mutatja az egyes beszédhangok helyes intenzitásviszonyait. E megjelenítésnek például a megfelelő szóhangsúlyok kialakításánál van jelentősége. Minden fejlesztett nyelv esetében nagyon fontos a helyes hangsúly kialakítása a szavakon belül, hiszen ez adja meg az adott nyelv sajátos ritmusát. A magyar nyelvre jellemző, hogy a szavakban mindig az első szótagon van a hangsúly. Vizuális megjelenítésnél a színek 6 dB-enként változnak, világos árnyalatok a nagyobb intenzitást jelentik.
Hanglejtési kép. A rendszer az alapfrekvencia értékét egyenletesen temperált hangmagassági skálán ábrázolja, vagyis az alapfrekvenciát az első szótag magánhangzójának kvázistacionárius részén mért értékhez viszonyítja. A kép azt mutatja meg, hogy hogyan változik az alaphang a kiejtés során az első szótag alaphangjához vi-

szonyítva. E módszer választásának két oka volt. Először is az intonáció hallási érzékelésénél, agyi feldolgozásánál nem az alaphang konkrét frekvenciaértéke az, ami meghatározó, hanem a változás maga, vagyis az, hogy az alapfrekvencia nő, vagy csökken, és azt milyen ütemben teszi. Másodszor, vizuális feldolgozás szempontjából is könnyebb a kiértékelés, ha az alapfrekvencia-görbe mindig ugyanazon a ponton kezdődik. Ugyanakkor előnyös a kép elhelyezése szempontjából is.
Hangvisszajátszás. A referencia-adatbázis minden eleme egyenként behívható és hangszórón keresztül lejátszható. A hallási, látási érzékeltetést mindig együtt alkalmazzuk, vagyis a hangképek megjelenésével szinkronban mindig elhangzik az adott kifejezés. A gyermekekkel történt tesztelések alapján világossá vált, hogy nagy szükség van arra is, hogy a gyemek a saját bemondott hangját is vissza tudja hallani. A rendszer ezt is lehetővé teszi.

12.9.1.3. A kiejtés jóságának automatikus megítélése

Gyakorláskor a gyermek először a pedagógus segítségével, majd önállóan dönti el, hogy a gyakorlandó szótagot, szót helyesen ejtette-e ki, vagy nem. A rendszerben lehetőség van automatikus értékelésre, amelyet automatikus visszajelzésnek nevezünk. Ez az automatikus visszajelzés segíti a gyakorlót a kimondás helyességének megítélésében. A kiejtés helyességét értékelő automatikus módszerként a rendszer a spektrális statisztikus modellektől való módosított négyzetes (euklideszi) spektrális eltérés mérését használja (Vicsi et al. 1999). Itt végeredményben a gyemek kiejtésének helyességét 5 fokozatban ítéljük meg (választhatóan osztályzattal vagy egy kacsa mozgásával vagy virág kinyílásával). Önálló gyakorlásnál ez a funkció sokat segít. Hallássérült gyermekeknél lélektani jelentősége van az automatikus visszajelzésnek, hiszen nincs rászorulva a gyermek egy külső személy segítségére ahhoz, hogy megfelelő visszajelzést kapjon. Hangsúlyozandó azonban, hogy a gyakorlónak elsősorban a hangképet kell figyelni, ez az elsődleges visszacsatolás, nem pedig az automata visszajelzés. A visszajelzés ugyanis csak annyit mond, hogy jó, vagy nem jó a kiejtés. A hangképre figyelve viszont azonnal látja a gyermek, hogy mi a rossz, például ha túl hangosan beszél, csökkenteni kell a hangerőt, vagy a gyakorlás során megtanulja látni, hogy nem képez zöngét a zöngés réshangoknál, vagy nem áramoltat elég levegőt, vagy például a magánhangzóknál nem eléggé nyitott az állkapcsa egy hangnál, vagy nem eléggé kerekít az [u] hangnál stb. Gyakorlás során ugyanis a gyermek meg kell, hogy tanulja a hangképek értelmezését. Közös gyakorlásoknál javasolt az automata visszajelzés kikapcsolása, hogy ne vonja el a tanuló figyelmét a hangképről.

12.9.1.4. Beszédoktatási módszertan a használathoz

A beszédterápia általános módszertana gyermekek és felnőttek kezelésére is alkalmas. A játékos háttérképek kikapcsolásával a rendszer felnőttek számára is jól használható, mivel azonban a helyes beszéd kialakítását minél előbb el kell kezdeni, jellemzően még beiskolázás előtt, ezért a megjelenítésnél, a gyakorlatok kialakításánál, a szótárkészlet összeállításánál elsősorban az 5–7 éves, szellemileg egészséges gyermekek átlagos teljesítményszintjéhez alkalmazkodtunk. Ebben a korban még ki lehet alakítani stabil fiziológiás szenzomotoros csatolásokat, így a megtanult helyes hangadás, artikuláció, hanglejtés stabilan megmaradhat. Részletes oktatási módszer került kidolgozásra vezető logopédus, szurdopedagógus szakemberekkel a hangkiadás fejlesztésétől a folyamatos beszéd kialakításáig. Az oktatási módszer kidolgozásánál a hagyományos beszédterápia már bevált módszereit mindig figyelembe vették (Kovács-Vass 1974, 1983, Csányi 1990, Kassai 1998, Farkas 1996). A többnyelvű rendszereknél pedig a nemzeti sajátosságokat mindig szem előtt tartottuk a fejlesztésnél. A tanulás során állandó gyakorlásra, ismétlésre van szükség. A rendszerben minden gyakorlandó hang izolált formában, szótagokban, szavakban, mondatokban és minimál párokban szerepel különböző helyzetben és különböző hangkapcsolatokban. Az elegendő és változatos gyakorlás érdekében nagy szótárkészlet került beépítésre. A gyakorlatok felépítésénél mindig a legkönnyebb kiejtéstől kell haladni az egyre nehezebb felé. A gyakorlatok sorrendje is egymásra épülve az egyre bonyolultabb szöveg kiejtése felé halad. A végcél a szép folyamatos beszéd automatikus kiejtésének az elsajátítása. Külön gyakorlatsorozat van egy-egy szupraszegmentális beszédjellemző (a hangerő, a ritmus és a hanglejtés) helyes kialakítására is. A kialakított teljes menürendszer minden nyelvre közelíti a 12.2. táblázatban bemutatott szerkezetet.

12.2. táblázat. A tanításhoz kialakított menürendszer

Előkészítés	**Sziszegők**	**Magánhangzók**	**Hanglejtés**
Hangerősség	Beszédhang-kialakítás	Beszédhang-kialakítás	Ereszkedő
Ritmus	Artikuláció	Artikuláció	Eső
Színkép	Kitartott hang	Kitartott hang	Emelkedő-eső
Zönge	Hangkapcsolat	Hangkapcsolat	Eső-ereszkedő
Alaphang	Elején	Elején	Lebegő
Zöngés-zöngétlen	Belsejében	Belsejében	Szökő
	Végén	Végén	
	Szavak	Szavak	
	Elején	Egy szótagú	
	Belsejében	Több szótagú	
	Végén	Mondatok	
	Mondatok	Szópárok	
	Szópárok		

A gyakorlatok sorrendjének jelentősége. A különböző beszédjellemzők gyakorlására változatos lehetőségeket ad a rendszer. A használt hangképekkel egymás után fokozatosan kell megismertetni a gyermeket, hiszen meg kell tanulnia a látottakat értelmezni, és miután ez az értelmezés automatikussá válik, akkor lehet már a hangképzésre összpontosítani. Ez a folyamat, ahogy haladunk a gyakorlatokkal, automatikusan végbemegy, hiszen az egyszerű vizuális megjelenítéstől haladunk a bonyolultabb felé. A legösszetettebb képi megjelenítés a szavak, mondatok gyakorlásánál van, de mire elérünk ide a terápia során, addigra már semmi problémát már nem jelent a látott hangkép értelmezése. Például, amikor egy új hang kialakítása a feladat, célszerű a kitartott hangokkal indulni, majd a hangkapcsolatokkal folytatni. A szavak gyakorlása akkor célszerű, amikor a hangkapcsolatokban (szótagokban) már jól be lettek gyakorolva az adott szó hangjai, és a gyermekek jól megismerték a hang és a hanghoz tartozó hangkörnyezet hangképét, a háttérrel együtt.
Szabad gyakorlás. A terápia során mindig felmerülhet olyan gyakorlat, ami előre nincs a rendszerbe beépítve, de a felhasználó oktató kívánatosnak tartaná a használatát. Ilyen esetben javasoljuk, hogy válassza a szabad gyakorlást, ahol maga dönthet, hogy mit fognak gyakorolni. Itt lehet új hangcsoportokat, például a nazális hangokat gyakorolni, vagy olyan új szavakat, mondatokat, amelyek a program szótárában nincsenek. A terapeuta bemond egy szót vagy kifejezést, amelynek spektrogramja a képernyő felső részén jelenik meg, majd a gyemek mondja be ugyanazt a kifejezést, amelynek spektrogramja a képernyő alsó részében látható. A két kép összehasonlításával eldönthető, hogy jó volt-e a kiejtés, vagy nem.
Hallásfejlesztés. A képernyőn megjelenő referenciaminták vagy a gyermek által kimondott minták tetszés szerint többször is meghallgathatóak. A hang és a kép együttes használata alapeleme a kidolgozott módszernek, és együttes használatuk a leghatékonyabb. Hallássérült vagy implantátumot viselő gyermekeknél ez az együttes használat is fejleszti a hallást. Ezeknél a gyermekeknél viszont külön hallástréningre van szükség. A program lehetőséget ad hallástréningre a teljes szótárkészletével oly módon, hogy a képernyőtől elfordulva, vagy a hangképet lefedve, csak a hallásra koncentrálunk, és visszamondatjuk a gyermekkel a hallott kifejezést. A gyermek próbálkozásai ilyenkor is lementhetőek a tárba és később visszahívhatóak. A közeli beszédhangok hallás alapján történő megkülönböztetésének fejlesztésére a szópárokban végzett gyakorlás kiválóan alkalmas.
Kapcsolat a betűkkel. A módszer hangsúlyozottan közvetlenül iskolakezdés előtt álló gyermeknek ajánlott, erre optimalizálták. Azokra, akik még nem tudnak olvasni, de szellemileg már érettek a különböző szimbólumok megkülönböztetésére és memorizálására. Minden gyakorlandó beszédhangnak a programban állandó szimbólumképe van. Az [s] hang szimbólumképe például a kígyó, az [i] hangé a kiscsibe stb. Mindig ezek a szimbólumképek jelennek meg az adott hang hívóképeként, vagy a háttérképecskéken. Így a gyermek megtanulhatja csak a hangképeket látva, hogy melyik beszédhangról van szó. Mivel a kimondott hang, szótag, szó, kifejezés betűje, betű-

sora is mindig megjelenik a képernyőn, a terápia végére a gyerekek megismerkednek a magyar betűkészlet nagy részével. Az iskolai oktatás során ezt a kialakult új ismeretet tudatosan megerősíthetjük. A szavak hívóképei a szavak betűsorát, valamint a szavak értelméhez kötődő, azt kifejező rajzokat mutatják. A szavak választékát adó kártyák kiváló alkalmat adnak arra, hogy a kártyán lévő rajzokról beszélgetni lehessen.

A Beszédkorrektor (SPEech COrrector, SPECO) rendszerről bővebb információ kapható, és demoprogramok is letölthetők az alábbi honlapokról: `http://alpha.tmit.bme.hu/speech/research.php`, `http://rcs.hu/sc.htm`

12.10. Beszédkommunikátor beszédsérültek segítésére

Tóth Bálint–Németh Géza

Az Európai Unióban a beszédképességüket teljesen vagy szinte teljesen elvesztett emberek száma körülbelül kétmillióra tehető. A beszélni tudás hiánya bekövetkezhet bizonyos betegségek (például agyvérzés), idegrendszeri zavarok, külső behatások (például baleset, műtét) következtében. A beszédsérült emberek jelentős része idős ember, akik sokszor egyéb fogyatékosságtól is szenvednek (például rossz látás, koordinációs zavarok). A beszélni nem tudó emberek számára nagyon nehéz a mindennapi életben való helytállás; problémát jelent mind az élő, szemtől szembeni, mind a telefonon keresztül való kommunikáció. Ez nemcsak magánéletükben jelent súlyos gondot, hanem a munka szempontjából is hátrányban vannak az egészséges emberekkel szemben. Ezentúl számos kellemetlen helyzetbe kerülhet a beszélni nem tudó ember, de sajnos akár halálhoz is vezethet a beszédképesség hiánya (például érzi a szívroham előjeleit, de nem tud senkit sem felhívni telefonon segítségért). Ezért számukra nagyon fontos egy, az elvesztett beszédképességüket pótló rendszer, mely által a beszéddel való kommunikáció valamilyen szinten ismét elérhetővé válik. A beszédkészség elvesztése – főleg a kezdeti szakaszban – a legtöbb esetben kisebb és nagyobb mértékű lelki betegségeket (depresszió, szorongás, fóbiák) is okozhat. A beszédkészségét nemrég elvesztett ember nehezen tudja elfogadni azt a tényt, hogy a továbbiakban a természetes beszéd mint kommunikációs csatorna nem létezik számára. Ezért egy speciális eszköz, amely pótolja az elvesztett képességet, nem csak kommunikációs, hanem mentális problémákon is képes segíteni. Sokkal kedvezőbb a helyzetük azoknak az embereknek, akik beszédképességüket csak átmenetileg vesztik el, például gégeműtét miatt. Számukra nem létfontosságú, azonban igen hasznos és kényelmes lehet egy beszédet generáló segédeszköz használata a kritikus időszak alatt. Ma már reális célkitűzés lehet, hogy egy olyan kommunikációs segédeszközt készítsenek, mely lehetővé teszi a gyors, kötetlen, szemtől szembeni és telefonos pár-

beszédet a beszédsérült személy és partnere között. A megoldást a gépi beszédkeltés alkalmazása jelenti, speciális kezelői felülettel együtt.

Magyarországon már a 90-es években készítettek ilyen céleszközöket, amelyeket tesztelési szinten kórházakban alkalmaztak (Olaszy–Németh 1993). Erre a kis méretű, hordozható számítógépek megjelenése adott módot (laptop). Az akkor Voxaid névre keresztelt beszédkommunikátor laptop beépített hardver beszédszintetizátorral rendelkezett. A szövegbevitel rugalmassá tételére célprogramot készítettek (textcreator), amely több szinten tette lehetővé a felhasználó számára a kimondandó szöveg elkészítését, tárolását, gyors előhívását. A gyakorlati tapasztalatok azt mutatták, hogy leginkább a telefonáláshoz volt nélkülözhetetlen a készülék (személyes beszélgetésnél még van mód az „én leírom, te elolvasod" kommunikációra). A céleszközt, német és spanyol nyelvre is elkészítették, és mintegy 15 helyszínen tesztelték Európában. Az itt ismertetett megoldás már korszerű mobileszközökre épül.

A beszédkommunikátor funkciói. A funkciók definiálása az első lépés. Fontos, hogy ne az alkalmazható műszaki megoldás (hardver) szabja meg az eszköz szolgáltatásait, hanem a kívánt szolgáltatások alapján választhassuk ki a legmegfelelőbb hardver(eke)t. A segédeszköz alapvetően szövegbevitelre épül, a szöveget alakítja át beszéddé (feltételezzük, hogy a sérült személy képes szöveget készíteni a mobileszközzel). Az eszköz használatakor a beszédsérült ember szövegben adja meg, amit el akar mondani. A gép a szöveget alakítja át beszéddé. A szövegbevitelt ebben az esetben nem feltétlenül úgy kell elképzelni, hogy karakterenként beírjuk a szöveget a felhasználás pillanatában, hiszen az nagyon sok időt venne igénybe, ezt nem lehetne dialógusnak nevezni. Olyan megoldásokra törekszenek a fejlesztők, amelyekkel a beszédsérült ember gyorsan tud szöveget kimondásra előkészíteni. Ennek egyik lehetősége az előtervezés biztosítása (szövegek előzetes beírása és eltárolása). Egy másik szempont, hogy szűkíteni lehessen a témaköröket, az azokhoz tartozó szövegeket. A felhasználónak biztosítani kell, hogy adott témakörhöz tartozó szövegeket el tudjon tárolni, és azokat egy-két gombnyomással elő tudja hívni, kissé módosítani és ki tudja mondatni.

Szabad szövegbevitel. A felhasználónak lehetőséget kell adni kötetlen szövegbevitelre. A szöveget lehessen tárolni (célzottan is), és később újra előhívni. Így hosszabb párbeszédek esetén a felhasználó előre eltárolhatja a lehetséges válaszokat, később előhívhatja azokat, a beszélgetés során pedig csak választania kell közülük. A szabad szövegbevitel teremti meg a beszélni nem tudó ember számára a kötetlen beszéd lehetőségét.

Sablon (kötött) szövegek. Fontos, hogy a felhasználó tudjon az előre megírt mondatokból gyorsan, egyszerűen választani, mondatokat összeállítani. Hasznos a mondatokat tartalmuk szerint kategóriákra bontani (gyógyszertár stb.). A mondatokat és a kategóriákat külön programból lehet szerkeszteni. Az előre megírt mondatok hasznosak a helyzetek gyors megoldásában (például *Fáj a jobb lábam! Rosszul érzem magam.*, de a párbeszédet is gördülékenyebbé teheti: *Örülök, hogy találkoztunk!*).

Szerkeszthető sablon (félig kötött) szövegek. A párbeszédek gyorsítása érdekében érdemes ötvözni a szabad és kötött szövegmódokat a következőképp: a mondatokat a kötött szöveghez hasonlóan kategóriába rendezett módon tároljuk, azonban amikor kiválasztunk egy mondatot, a mondat bizonyos, előre meghatározott részét még szerkesztheti a felhasználó (például *kérek szépen x dkg sajtot* stb.).
Szöveg felolvasása. A programnak fel kell tudni olvasni a felhasználó által beírt szöveget (az egészet vagy a felhasználó által meghatározott részeit), illetve az előre eltárolt mondatokat. A szövegeknek természetesen láthatónak kell lenni a képernyőn.
Gyors szövegbeviteli módszer. Mivel a beszélő alkalmazást a felhasználó más emberekkel való kommunikációra használja, lényeges a megfelelő sebességű szövegbeviteli módszer.
Telefonálás. Fontos a telefonálás megoldása, hiszen ez olyan kommunikációs csatorna, mely egy beszélni nem tudó ember számára elérhetetlen marad megfelelő technológiai segítség nélkül. Míg szemtől szembeni kommunikáció során papírra le tudja írni, vagy el tudja mutogatni közlendőjét, telefonon képtelen segítség nélkül kommunikálni. Ezért mind az alkalmazott eszköznek, mind a hozzá készített programnak támogatnia kell a telefonálás lehetőségét.
Kisegítő lehetőségek, testreszabhatóság. Mivel az alkalmazást elsősorban idős emberek fogják használni, ezért fontos, hogy be tudják állítani a számukra legjobban olvasható formát. Ezért szükséges, hogy a betűk színét, háttérszínét és méretét változtatni lehessen. Ezáltal a gyengén látók is tudják használni a programot. A beszéd paramétereit is állíthatóvá kell tenni (sebesség, hangmagasság, hangerő).
Eszközfüggetlen működés. A beszédsérült felhasználó számára meg kell adni a választás lehetőségét a készüléket illetően (például Smartphone, PDA, Tablet PC, asztali számítógép), ugyanis maga a program az összes eszközön hasonló támogatást nyújt. A felhasználónak csak a programot kell telepítenie a gépére. Ebben a tekintetben a kérdés az volt, hogy milyen fejlesztői hátteret válasszunk? A legrugalmasabbnak a .NET Framework / Compact Framework mutatkozott több platformos kód hordozhatóságának tekintetében. A Java MIDP is felmerült mint alternatíva, azonban a fejlesztés kezdetekor Java MIDP-ben több funkció csak nagyon nehezen volt elérhető (például telefonálás), továbbá a Microsoft saját eszközein jobban támogatja a Microsoft-alapú fejlesztéseket, több dokumentumot, segédanyagot és hibajavítást adnak ki hozzá.

A .NET keretrendszerben készült programokat megfelelő körültekintés mellett viszonylag könnyen lehet a mobileszközök és az asztali számítógép között hordozni, de természetesen mindig figyelembe kell venni az adott platform teljesítményéből, felhasználói felületének méretéből és adatbeviteli módjából adódó sajátosságokat.
Egy gyakorlati megoldás ismertetése. A BME TMIT-en készült el a VoxAid szoftvercsalád, ami beszédképtelen embereknek adja meg a szintetizált beszéddel való hangos kommunikáció lehetőségét kis méretű, hordozható, kereskedelemben kapható eszközökön (PDA, mobil stb.).

Amennyiben a hordozhatóság, a gyors, rövid információközlés, segítség kérése a cél, fontos, hogy a hardver kis méretű legyen, de ezzel szemben a nagy kijelző és a könnyű szövegbeviteli mód is lényeges szempont. Ezek alapvetően ellentmondásos tulajdonságok, ezért egyfelől a hardverválasztás kompromisszumokkal járt, másrészt azon kihívást kellett megoldani, hogy az eszköz hiányosságait, hátrányait az alkalmazás funkciói kompenzálják. A választás először a Pocket PC alapú PDA készülékekre esett (12.31. ábra). A Pocket PC alapú PDA készülékek esetében gyorsabb

12.31. ábra. A beszédkommunikátor PDA változata, szabad szöveg (bal), félig kötött szöveg (jobb)

a szövegbevitel, és nagyobb terület áll rendelkezésre a felhasználói felület számára, azonban a készülék sérülékenyebb, mérete nagyobb, és a program vezérlése is lassabb. A beszédkommunikátor PDA-s változata elsősorban közepesen hosszú, közepesen összetett beszélgetések kivitelezésére javasolt (Tóth et al. 2004).

A legösszetettebb változata a rendszernek a beszédkommunikátor asztali/laptopos, Tablet PC-s verziója (Tóth–Németh 2006). Ez a változat már nem csupán beszédsérült, hanem beszéd- és/vagy hallássérült embereknek készült, a korábbi rendszereket beszédfelismerővel és új funkciókkal egészítették ki.

Ez a beszédkommunikátor egy ISDN modem segítségével küldi és fogadja a hanganyagot (szintetizált vagy természetes beszédet) a telefonvonalról. A szoftver képernyője két fő részből áll: a bejövő és a kimenő szövegből. A bejövő szövegablakban a beszédfelismerő modul által felismert szavak, mondatok jelennek meg, míg a kimenő szöveg ablakban a korábban megismert három lehetőség áll a felhasználó rendelkezésére (a szabad, a kötött és a félig kötött szöveg).

12.11. Hallásmérés szintetikus beszéddel

Olaszy Gábor

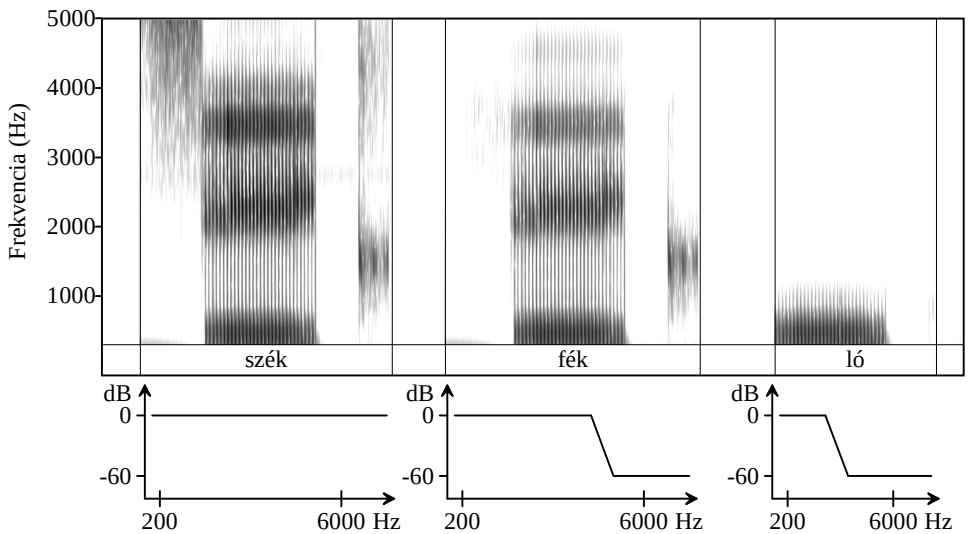
A hallásnak és a beszédészlelés szintjének a mérése már kisgyermekkorban nagyon fontos. A beszédtechnológia fejlődése lehetővé tette, hogy a hagyományosnak tekinthető – főleg felnőtteknél alkalmazható – szinuszos hanggal végzett hallásvizsgálaton túlmenően újfajta eljárások is megszülessenek. Ezek egyike a GOH rendszerű beszédhallás-ellenőrző eljárás, amely alapvetően kisgyermekek szűrésére szolgál, akár 3 éves kortól (Gósy et al. 1984). Az elve teljesen más, mint a szinuszos módszeré. Kihasználja egyrészről azt, hogy a beszéd egyes hangjaiban egyszerre több frekvencián van jelen nagy energiájú frekvenciakomponens, másrészről azt, hogy vannak olyan hangok, amelyek hangzását meghatározó frekvenciakomponensek csak egy meghatározott sávon belül lehetnek jelen. Amennyiben a fenti frekvenciakomponensekből nem jut el mindegyik az agyba, a beszédhang minősége megváltozik, más hangot hall a vizsgált személy, mint amit a fülébe közvetítettek, mondjuk egy fejhallgatón keresztül. Ha arra kérjük a vizsgált gyermeket, hogy ismételje meg a hallott szót, a válaszából következtetni lehet a hallásának ép, avagy sérült voltára. Az eljárás előnye, hogy a szűrés lefolytatása csak minimális előképzettséget kíván, a mérést az óvónő, a logopédus, a szülő könnyen és gyorsan el tudja sajátítani. A gyermekek szívesen vesznek részt a vizsgálatban, azt játéknak tekintik. Az eljárás magyar szabadalom. Lajstromszáma 193211.

12.11.1. A Mondom-2000 beszédhallást ellenőrző eljárás

Egy hallásvizsgálat során arról akarunk meggyőződni, hogy a fizikai rezgés minden frekvenciakomponense eljut-e az agyba és ott szabályosan feldolgozódik-e. A hagyományos, szinuszos jeleket használó hallásvizsgálati eljárásnál adott frekvenciájú és intenzitású hangot vezetnek a fülbe, és megkérik a vizsgált személyt, hogy jelezze (kézfeltartással), hogy mikor hallja meg a jelet. A visszajelzések alapján térképezik fel a hallási rendszer frekvenciaátvitelét. Ezt úgynevezett audiogramon ábrázolják. Az audiogram képe egy elektronikus szűrő átviteli karakterisztikájához hasonlítható, amelynek vízszintes tengelyén a frekvencia, függőleges tengelyén a hang intenzitása szerepel. A halláskárosodás egyik formája, ha valamely frekvenciatartományban a hallási rendszer levágja, elnyeli a hang eme komponenseit, tehát azok nem jutnak el az agyba. A másik forma, amikor az agy nem tudja feldolgozni megfelelően a hozzá érkezett jeleket. Az emberi beszéd frekvenciakomponensei a 100–8000 Hz-es frekvenciatartományban helyezkednek el, tehát ezt a frekvenciasávot kell vizsgálni.

Akusztikai elv. A hallásmérés akusztikai elve a Mondom-2000 elnevezésű, GOH rendszerű beszédhallást ellenőrző eljárásnál szintetizált beszéden (szavak) alapul,

nem szinuszhangokon. A vizsgálandó érzetisávot (fül és agy) a beszédhangokra jellemző frekvenciakomponensekkel ingereljük. Az eljárás fontos jellemzője, hogy csak minimalizált redundanciájú beszédjellel lehet a hatásos mérést elérni (közvetlen emberi bemondás hangfelvétele nem használható, beszédtechnológiai eljárásokkal kialakított beszédjelre van szükség). Egy elvi példán mutatjuk be az akusztikai hatásmechanizmust (12.32. ábra). Legyen a fülbe közvetített szó a *szék*. A hangsor



12.32. ábra. A GOH hallásmérési elv bemutatása. A halláscsökkenés mértéke határozza meg, hogy a fülbe adott ingerből mennyi jut el az agyba, és ott milyen szóhoz társítódik

hangjainak akusztikai szerepe a vizsgálatban a következő. Az első mássalhangzó a magas frekvenciás jelet képviseli 5000 Hz-től felfelé. A magánhangzó a 400, 2000, 3000 Hz-es frekvenciákkal (formánsokkal) és azok szűk formánssávszélességével jellemezhető). Az utolsó mássalhangzó kis energiájú zöngétlen zárhang, tehát a hallási rendszer hangerőátviteli csökkenését lehet vele mérni.

Amennyiben a vizsgált személy hallási mechanizmusa enyhe magas frekvenciás károsodású (az 5000 Hz feletti frekvenciákat nem engedi át, ez enyhe hallászavarnak tekinthető), akkor az agyi percepcióban a *fék* szó fog megjelenni (az első mássalhangzó érzeti megjelenése az agyban a labiodentális zöngétlen réshangot eredményezi a dentialveoláris zöngétlen réshang helyett). Ennek akusztikai magyarázatát a mássalhangzókról szóló fejezetben részletesen kifejtettük. Amennyiben a halláskárosodás még alacsonyabb frekvenciára tolódik lefelé (például 2500 Hz-re), akkor a zöngétlen réshang el is tűnhet az érzetből, és az *ék* szó azonosítása jelenik meg az agyban. Nagyfokú halláskárosodásnál (1000 Hz feletti vágás) a magánhangzó magasabb formánsai is eltűnhetnek és az érzet átcsaphat egy másik magánhangzóba,

olyanba, amelyiknek a jellemző formánsai az 500 Hz-ig tartó frekvenciasávban vannak jelen. A beszédhangok akusztikai jellemzőinek ismeretében ilyenkor az *ó, ló* hangsor érzete jelenik meg az agyban (a megadott válaszok valós mérésekből származnak). A fentiekből látható, hogy egyetlen szóval több frekvenciasávot és ezzel több frekvenciasávra vonatkozó halláskárosodási fokozatot lehet prognosztizálni.

Az eljárást az 1980-as évek elején dolgozták ki a formánsszintézis elvének alkalmazásával, több évi kutató munkával. A mérés hatásosságát több száz éphalló és halláskárosodott gyermek vizsgálatának összehasonlításával igazolták. A mérési elv véglegesítése során kialakult az optimális mérési tér, ami azt jelenti, hogy 10 szóval lefedhető a teljes vizsgálandó beszédfrekvencia-tartomány. A beszéd redundanciájának minimalizálása azt jelenti, hogy a magánhangzókat keskeny formánssávszélességgel valósítjuk meg, az ezeken felüli frekvenciakomponensek tehát nem lesznek jelen a jelben, a réshangok frekvenciasávjait pedig egy szűkebb frekvenciatartományra korlátozzuk. Például a dentialveoláris zöngétlen réshang létrehozható akár három fajta szűk frekvenciasávval is az 5000–8000 Hz-es tartományon belül (5500 Hz, 6500 Hz és 7500 Hz) anélkül, hogy a percepciójában valamilyen zavar következne be ép hallás esetén. Ez a tény lehetővé teszi, hogy a hallásmechanizmus magas frekvenciás átvitelét finom frekvenciasávokra osszuk és mindegyikhez „gyártsunk" egy-egy külön szót, amelyben a dentialveoláris zöngétlen réshang a beállított frekvenciasávval van jelen. A fenti elvet más hangoknál is alkalmazzuk.

A GOH mérési módszer. A mérés elve hasonlít az anyanyelv elsajátítási folyamathoz, ezért sikeresen használható kisgyermekek hallásának szűrésére is. Nem klinikai, diagnosztikai eszköz, hanem gyors szűrésre, előrejelzésre alkalmas (rövid ideig terheljük a gyermeket). A szűrés eredménye, hogy kell-e klinikai vizsgálat, vagy minden rendben van a gyermek hallásával. A hallásszűrést játékként vezetjük be. A gyermeket megkérjük, hogy a fejhallgatóból kiadott szót hallgassa meg, majd mondja vissza hangosan, mivel mi nem halljuk, hogy mit mond a mókus. A választ a tesztlapon rögzítjük. A hallásmérés célkészülékkel történik, ez a GOH beszédhallás-vizsgáló eszköz (digitális, kis méretű hanglejátszó a tesztszavak megszólaltatására). Az eszközhöz tartozik egy fejhallgató, valamint a tesztlap, a mérési eredmények rögzítésére (12.33. ábra).

1. Szócsoportkijelző
2. Az elhangzó szó sorszámának kijelzése (az első szónak minden csoportban a nulla szám felel meg)
3. A szó elhangzásának kijelzése (fénysor)
4. Jelzőfény a jobb (piros) vagy bal fül (zöld) átkapcsolásának jelzésére
5. Bal fü - jobb fül átkapcsoló nyomógomb
6. Állj gomb a szavak elhangzásának megállítására
7. Indít gomb a szavak elhangzásának indítására. A lejátszás indulását a gomb feletti piros fény jelzi.

12.33. ábra. A GOH mérőkészülék

8. Vissza gomb. Visszalépés az előző szóra, az elhangzott szó ismétlésére
9. Csoport gomb a szócsoport kiválasztására (1–6 csoport)
10. Hangerőbeállító gomb. Az erősebb hangerőnél (szobai környezet) a felső, piros fény világít, gyengébb hangerőnél (vizsgáló csendesszoba) az alsó, sárga. A két hangerő között 10 dB különbség van
11. Tápegység-csatlakoztató
12. Fejhallgató-csatlakoztató

A lejátszó készülék automatikusan nyolc másodpercnyi szünetet iktat be az elhangzó szavak közé, ez elég arra, hogy a gyermek válaszát meghallgassuk és lejegyezzük. Amennyiben nem elég, megállítjuk a lejátszást az állj gombbal, majd ismételt megnyomással folytatjuk.

A mérés szóanyaga. A készülék memóriájában 40 db formánsszintetizátorral előállított egyszótagú szó van tárolva. Mindegyik speciális akusztikai szerkezetű, a méréshez van igazítva. (Figyelem! Emberi ejtésű ilyen szavakkal nem lehet mérni, a beszéd redundáns tulajdonsága miatt!) Az eltárolt szavakból 10-es szócsoportok választhatók. Hat ilyen szócsoportot állítottunk össze. A csoportok között sorszám szerinti funkcióegyezés van. Egyetlen 10-szavas csoport meghallgattatásával a beszéd teljes frekvenciatartománya vizsgálatra kerül. Külön szócsoportot kell használni a jobb, illetve a bal fül mérésénél. A szavak nagyobb része az óvodáskorú gyermekek számára ismert, kisebb részük értelmetlen hangsort jelent nekik: ez a visszamondásban azonban nem okoz problémát. A szavak első négy csoportjához tesztlap, az 5. és 6. csoporthoz képsor tartozik.

1. *meggy, sír, bú, ász, gáz, szél, méz, zsír, bőr, szú*
2. *mos, csíp, bab, szűz, bál, sző, bor, csók, kút, ész*
3. *mák, sín, bók, szív, ágy, cél, busz, só, zsák, szög*
4. *dob, sír, bús, szék, gyík, cím, gép, sár, gyík, ősz*
5. *mos, síp, bab, szív, ágy, szék, gép, só, kút, szög*
6. *meggy, síp, kút, szív, bot, cél, busz, zsák, dob, ősz*

A teszt úgy van összeállítva, hogy az 1.–4. oszlopok megfelelő sorszámú helyein levő szavak azonos frekvenciákat szűrnek, tehát egymással helyettesíthetők. Ha például a gyermek az első szócsoport alkalmazása esetén nem mondja vissza hibátlanul az *ász* szót, akkor – ha nem fáradt még –, használhatjuk – második kísérletként – a második csoportban levő *szűz*, a harmadikban a *szív* vagy a negyedikben a *szék* szavakat. Ugyanígy a *meggy* azonos frekvenciákat vizsgál a *mos*, a *mák*, a *dob* szavakkal.

A tesztlap. A tesztlapon jelöljük a gyermek által mondott válaszokat. Négy oszlopot lát a vizsgálatot végző: jó hallás, enyhe hallászavar, hallászavar és súlyos hallászavar, ez esetben azonnal orvoshoz kell fordulni. Minden oszlopban megtaláljuk azokat a jellemző válaszokat, amelyek a leggyakrabban fordulnak elő. A tesztlapon aláhúzással, bekarikázással, illetve beírással jelöljük a gyermek válaszait: ennek alapján azonnal elsődleges értékelést is kapunk (lásd később).

Kétféle tesztlap van: a) „A válaszlehetőségek fokozatai” című: ezt a tesztlapot két és fél évnél idősebb, jól beszélő, értelmes gyermek mérésére használjuk (12.3. táblázat), b) Képsoros tesztlap: ezt a hozzá tartozó képsorral együtt az erősen beszédhibás, szorongó, beszélni alig akaró gyermekek mérésére használjuk (részletes alkalmazását lásd később).

12.3. táblázat. A válaszlehetőségek fokozatai tesztlap részlete. A gyermek válaszait be kell karikázni

I. Ép	II. Enyhe zavar	III. Zavar	IV. Súlyos
Meggy(megy)	begy legy negy vegy	egy ety eny	bó e ó u
Síp(sik)	sít sűt sűp szíp szép	zúg su só fut	kút út
Bú	dú bók bot pók pú púk	tú tó pó út	ó ú
Ász	ház pász	ás ágy	áf áh át ó

A tesztlapon megkeressük azt a szót vagy hangsort, amelyet a gyermek az elhangzás után visszamondott vagy a képsoron megmutatott: a választ aláhúzzuk vagy bekarikázzuk. Ha olyan szót vagy betűsort mondott a gyermek, ami a tesztlapon nincs leírva, akkor a gyermek válaszát abban a rubrikában kell feltüntetni, amelyiknek a szavaihoz az elhangzott szó leginkább hasonlít. A jobb fülben hallott és visszamondott szavakat piros színnel, a bal fülben hallottakat zöld színnel jelöljük.

A képsor. A képsornak megfelelő hanganyag 20 különböző szót tartalmaz. Az egy-egy fülbe kerülő 10 szóhoz 17 kép tartozik. Ez azt jelenti, hogy egy fül mérésekor 17 képet kell a gyerek elé tennünk, hogy a megfelelőt kiválassza. Azért tartozik 10 szóhoz 17 kép, mert azok nemcsak az elhangzó 10 szó jelentését ábrázolják, hanem olyan szavakét is, amelyeket a beszédhallás esetleges zavara esetén mond vissza a

gyermek. (Például: a szív szót a gyermek beszédhallás zavara esetén fül-nek hallhatja (és mondja vissza), ezért a „fül” képe is megvan a képsorban.)
Értékelés. Az értékelést a vizsgálatot végző személy végzi a tesztlap segítségével. Meg kell nézni, hogy a gyermek válaszait (a gyermek által visszamondott szavakat, hangsorokat a tesztlap melyik függőleges oszlopába jelöltük be. Az oszlop tetején található az elsődleges értékelés és tennivaló (I. Ép hallás; II. Enyhe zavar; III. Zavar; IV. Súlyos zavar). A kiértékelésre életkor szerinti kritériumok szerint kerül sor.
Felhasználói kör. 2010-ben a GOH beszédhallás vizsgáló rendszert óvodák, logopédiai rendelők, nevelési tanácsadók és szülői lakóközösségek széles körben használják. Mintegy 800 készülék működik országszerte, és a gyermekek is és gondozóik is nagyon szeretik egyszerű kezelhetősége miatt.

13. fejezet
Interfészek, szabványok, honlapok, programok

13.1. VXML

Csapó Tamás Gábor

A VoiceXML egy szabványos, XML formátumú leíró nyelv, melynek célja, hogy a beszédalapú ember-gép dialógusok fejlesztését meggyorsítsa és könnyítse (Sharma–Kunins 2002). A VoiceXML a HTML nyelvhez hasonló elveken alapul, de utóbbi a vizuális tartalmak megjelenítését definiálja, előbbi pedig párbeszédek leírására szolgál. Céljai közé tartozik a webalapú interfészek és tartalomszolgáltatás előnyeinek bevitele interaktív hangvezérelt alkalmazásokba oly módon, hogy az utóbbiak készítéséhez ne legyen feltétel az alacsony szintű programozói szaktudás. A webfejlesztők felhasználhatják korábbi Java, XML és egyéb webfejlesztési tapasztalatukat, nem szükséges speciális IVR programozási nyelvek ismerete. A szabványos leírás előnye, hogy az elkészített dialógus (ideális esetben) gyártótól függetlenül ugyanúgy működik, így egy megvalósított alkalmazás mögötti infrastruktúra cseréje könnyen lehetséges.

A VoiceXML-t olyan hangalapú dialógusok készítéséhez hozták létre, amelyek tartalmaznak szintetizált beszédet, előre rögzített hanganyagot, valamint beszédet és DTMF kódot tudnak felismerni. A célok közé tartozik, hogy hangalapú hozzáférést nyújtson bizonyos alkalmazásokhoz: vezetékes és mobiltelefonról, PDA-ról vagy VoIP kapcsolattal rendelkező számítógépről.

A (beszédalapú) dialógusrendszerek célja, hogy az ember-gép kapcsolat megfelelő működését a beszédtechnológia eszközeivel segítsék. A párbeszéd- vagy más néven dialógusalapú rendszer létrehozásához egy olyan fejlesztőkörnyezet szükséges, amelynek segítségével kialakítható egy adott feladathoz legjobban megfelelő dialógus. A rendszer beszédszintetizátor és beszédfelismerő integrálásával tudja a gyors alkalmazásfejlesztést támogatni.

A VoiceXML előnye abban is felfedezhető, hogy a VoiceXML alkalmazás beszéddel vezérelt böngészőprogramként hozzá tud férni az internethez. Felhasználásával

lehetőség van arra, hogy adatokat küldjünk és fogadjunk webszerverekhez kapcsolódva. Amennyiben a VoiceXML-t egy webes alkalmazás előtétjeként használjuk, jelentősen lecsökkenthető a megírandó VoiceXML kód. Ilyenkor az alkalmazás nagy része alapulhat olyan ismert protokollokon, melyekhez már léteznek hatékony fejlesztőeszközök (például HTML és PHP).

Már a web születésekor is felmerült az igény olyan egyszerű leíró nyelvre, amellyel automatizált, beszédalapú rendszerek fejleszthetőek. A VoiceXML 1.0 változatot 2000-ben hozta nyilvánosságra a VoiceXML Forum (AT&T, IBM, Lucent, Motorola). A 2.0 változat, amely már szélesebb körben elterjedt, 2004-ben született meg W3C ajánlás formájában. Néhány apró javítást elvégezve a könyv írásakor a legfrissebb elérhető változat a 2.1, melyet 2007-ben adtak ki (VoiceXML 2007). A VoiceXML Forum közeli tervei közé tartozik a 3.0 változat létrehozása, mely számos új funkcióval fog rendelkezni. A következő leírásban a VoiceXML 2.0 változatára láthatunk példát (VoiceXML 2004).

13.1.1. VoiceXML alkalmazásfejlesztés

Egy tipikus VoiceXML alkalmazásban a felhasználó először csatlakozik a rendszerhez, vagyis tárcsázza a megfelelő telefonszámot. A VoiceXML értelmező fogadja a hívást, és elkezdi végrehajtani a VoiceXML programkódot. Az értelmező különböző utasításokat hajthat végre a program hatására. Ezek közé tartozhat az előre felvett üzenetek (promptok) vagy egyéb hanganyag (zene, illetve hangeffektek) lejátszása a felhasználónak, a felhasználó által beütött DTMF kód fogadása, a felhasználó hangbemondásainak fogadása és a beszéd felismerése, a felhasználó hangbemondásainak fogadása és felismerés nélküli tárolása, a felhasználó kéréseinek továbbítása egy webszerver felé, illetve adatok fogadása webszervertől és azoknak továbbküldése a felhasználó felé.

A VoiceXML alkalmazások képesek bizonyos előre programozott funkciók megvalósítására, mint például az aritmetikai műveletek vagy egyszerű szövegfeldolgozás. Ezeknek segítségével az alkalmazás ellenőrizni tudja a felhasználói bemenet érvényességét. Egy-egy dialógus nem csak egy statikus sorozat lehet, ami mindig ugyanúgy megy végbe minden híváskor. Lehetőség van dinamikus elemek hozzáadására is (például feltételes szerkezetek alkalmazásával), melyek az aktuális állapottól függően módosítják a dialógus menetét.

Egy példaalkalmazáson keresztül röviden bemutatjuk a VoiceXML működését.

```
1 <?xml version="1.0"?>
2 <vxml version="2.0" xmlns="http://www.w3.org/2001/vxml">
3   <form>
4     <field name="selection">
```

```
5           <prompt>
6               Kérlek válassz: Hírek, mozi vagy sport.
7           </prompt>
8           <grammar mode="voice">
9               <rule>
10                  <one-of>
11                      <item>hírek</item>
12                      <item>mozi</item>
13                      <item>sport</item>
14                  </one-of>
15              </rule>
16          </grammar>
17      </field>
18      <block>
19          <submit next="process.vxml"/>
20      </block>
21  </form>
22 </vxml>
```

Egyszerű VoiceXML alkalmazás forráskódja

A fenti VoiceXML kód eredménye: a megfelelő telefonszám felhívása esetén a rendszer felolvassa a „Kérlek válassz: hírek, mozi vagy sport.” részletet. Ezután a felhasználó válaszol, majd a rendszer továbbítja őt a „process.vxml” oldalra, ahol a válasz feldolgozása történik.

A példaalkalmazás néhány alapvető dolgot mutat be a VoiceXML-ből. A program nyers szövegből és *címkék*ből áll. Utóbbi a kulcsszavakat és kifejezéseket jelenti, amelyeket zárójel határol (< és >).

A fejléc (1–2. sor) és a záró </vxml> címke (22. sor) minden VoiceXML dokumentumba kötelezőek. Ezek megfelelnek a szabványos VoiceXML leírásnak.

A fejléc után következik a VoiceXML dokumentum fő része, a törzs. A törzsben lévő <form> címke egy VoiceXML űrlapot definiál, ami a dokumentumnak a felhasználóval történő interakciót végző része. Ez a VoiceXML-beli megfelelője a HTML űrlapnak.

Az űrlap általában *mező*kből áll, amelyek a felhasználótól érkező információt kezelik. Ebben a példában az űrlap egy selection nevű mezővel rendelkezik (<field>, 4–17. sor), amely a felhasználó választását tárolja.

A mezőn belül van egy <prompt> címke (5–7. sor), amely nyers szöveget tartalmaz. A VoiceXML értelmező a hozzá tartozó beszédszintetizátor moduljával a szöveget beszéddé alakítja át, és a létrejött hanganyagot lejátssza a felhasználónak.

A <prompt> után egy <grammar>, vagyis nyelvtanelem található (8–16. sor), amely definiálja a három lehetséges választ, amelyeket a beszédfelismerő felis-

merhet, és az értelmező elfogad. A <rule> nyelvtani szabály megadására szolgál, mely jelen esetben a <one-of>-on belül <item> elemek felsorolását jelenti. A <grammar> címke bonyolultabb formáival lehetőség van komplex nyelvtanok megadására is. A nyelvtanok általában kimondandó szavakat vagy DTMF kódokat tartalmaznak, az alkalmazás mindegyiket tudja kezelni. A VoiceXML rendszerekben több különböző formátumú nyelvtan megadására is lehetőség van.

Miután az értelmező elküldi a promptot a felhasználónak, várakozik a válaszra és a beszédfelismerő megpróbálja összehasonlítani a választ a megadott nyelvtannal. Ha a válasz megfelelő, akkor az értelmező az adott mező selection változóját a válasznak megfelelőre állítja.

Végül az űrlap a <block> elemmel zárul, amely egy <submit> címkét tartalmaz (18–20. sor). Ezek a sorok átadják a vezérlést a process.vxml-nek, a selection változó értékét is továbbítva a feldolgozó szkriptnek.

A fenti példában látható, hogy a VoiceXML programozási nyelv felhasználásával egyszerűen és gyorsan készíthetünk interaktív dialógusokat. További, bonyolultabb példaprogramok a következő hivatkozásokon találhatóak:

```
http://cafe.bevocal.com/resources/voicexml_samples/index.html
https://studio.tellme.com/library2/code/
http://developer.voicegenie.com/examples_VoiceGenie.php
```

13.1.2. VoiceXML alapú alkalmazások

A VoiceXML specifikációhoz számos szabad felhasználású és kereskedelmi forgalomban kapható platform készült: *BeVocal, JVoiceXML, OpenVXI, PublicVoiceXML, Tellme Studio, VoiceGenie* (Rehor 2007). Ezek mindegyike tartalmaz VoiceXML böngészőt, amely a programok értelmezését és végrehajtását végzi; valamint különböző webes és offline fejlesztőeszközök állnak rendelkezésre az alkalmazások elkészítéséhez.

A beszédtechnológiával foglalkozó nagy cégek közül a legtöbb (*Avaya, AT&T, IBM, Loquendo, Lucent Technologies, Motorola, Nuance*) megvalósította, hogy beszédszintetizátoraik és beszédfelismerőik a VoiceXML interfészen keresztül is használhatóak legyenek. Ezáltal lehetővé tették, hogy a beszédalapú alkalmazások fejlesztéséhez ne legyen szükség speciális szaktudásra. A rengeteg valós VoiceXML alkalmazásból néhányat kiemelünk és röviden bemutatunk a következőkben.

Tipikus felhasználás az utastájékoztató rendszer (például az Amerikai Egyesült Államokban elérhető *Utah's 511 Traveller Information*), melyet felhívva megtudhatjuk a legfrissebb közlekedési információkat, lekérdezhetjük a tömegközlekedési menetrendeket, valamint időjárás-előrejelzést hallgathatunk meg. A mobiltelefonszolgáltatók (például *mobilkom austria, Telecom Italia Mobile, Orange UK*) önki-

szolgáló telefonos rendszerei is sokszor VoiceXML menü alkalmazásával készülnek el. Az *OnStar* autóba épített kommunikációs rendszere, mellyel baleset esetén juthatnak gyors segítséghez az amerikai előfizetők, szintén a VoiceXML technológián alapul. Ezen beszédalapú információs rendszerek előnye a könnyű telepítés és módosítás mellett, hogy a nap 24 órájában elérhetőek, hiszen emberi erőforrás nélkül, gépi vezérléssel működnek.

13.2. Programozói interfész beszédtechnológiai alkalmazásokhoz (SAPI)

Kiss Géza

A könyvünkben részletesen tárgyalt beszédfelismerő és beszédszintézis-rendszerek alkalmasak arra, hogy emberekkel természetes nyelvű kommunikációban vegyenek részt; de ez csak úgy lehetséges, ha számítógépes alkalmazásokba beépítik őket. A beszédtechnológia fejlesztői és a felhasználói programok készítői ritkán ugyanazok a személyek vagy akár azonos cég munkatársai, mivel merőben eltérő szaktudás szükséges e kétféle munkához. Mégis, az általuk készített moduloknak együtt kell működniük. A beszédtechnológia készítői számára cél, hogy fejlesztéseik eredménye minél több alkalmazásba beépíthető legyen; hasonlóképpen az alkalmazások készítőinek is fontos, hogy ne legyenek egy megoldáshoz kötve, hanem könnyen tudjanak fejlettebb technológiára váltani az alkalmazás jelentős átírása nélkül. Ahhoz, hogy az alkalmazások használni tudják e beszédtechnológiai eszközöket, kommunikálniuk kell velük; e kommunikáció lebonyolításhoz szükség van közöttük egy közös számítógépes „nyelvre”, egy számítógépes interfészre. A beszédinterfész céljára legjobban egy olyan beszédtechnológiai lingua franca felel meg, amelyet minden beszédtechnológiai modul (más néven beszédmotor) és minden alkalmazás ismer. Ebben a fejezetben két beszédinterfész-technológiát mutatunk be, amelyek azzal a céllal készültek, hogy ezt az összekötő szerepet betöltsék. Mivel egymással bizonyos mértékig konkuráló két technológiáról van szó, ezért egyiket sem nevezhetjük univerzális összekötő nyelvnek, de mindkét technológia nagy jelentőségű. Az egyik a Microsoft Speech API (MS SAPI), a másik a Java Speech API (JSAPI); az API jelentése Application Programming Interface, azaz alkalmazás programozói felület. A következőkben először további motivációt adunk e technológiák használatára. Ezután röviden bemutatjuk mindkettőt: történetüket, összetevőiket, használatukat és néhány rájuk alapuló alkalmazást. Végül beszélünk a két technológia egymáshoz való viszonyáról.

Szabványos interfészek használata. Szeretnénk néhány szemponttal megvilágítani az egységes beszéd interfész használata mellett, illetve ellene szóló érveket. Az itt leírtak megértését nagyban segíti, ha észrevesszük a természetes emberi nyelvekkel való

párhuzamot. Az emberi nyelvek között megkülönböztetünk nemzeti nyelveket, valamint nemzetközi státuszt kapott nyelveket; az utóbbiak között vannak mesterségesek (például eszperantó), és olyanok, amelyeket gazdasági-politikai-kulturális jelentőségük tett ilyenné (például angol). A nemzeti nyelvek elválaszthatatlanul összefonódnak az adott nép történetével, kultúrájával, életmódjával, gondolkozásmódjával. Az összekötő nyelveken általában kevésbé tudjuk magunkat kifejezni, és más-más szinten beszéljük őket; de jelentős haladást tesznek lehetővé azáltal, hogy széles körű együttműködést biztosítanak. E céljuk betöltéséhez szükség van még a kölcsönös érdekeltségi területre és olyan platformokra, ahol a kommunikáció megvalósulhat. Ugyanezeket elmondhatjuk a beszédtechnológiai komponensek vonatkozásában egyedi nyelveikre, a szabványos vagy kváziszabvány beszédinterfészekre, és az alkalmazásokban való felhasználhatóságukra.

A beszédfelismerők és beszédszintetizátorok rendszerint rendelkeznek egy egyedi interfésszel (native interface). Ez lehetőséget ad arra, hogy a komponenst teljes mértékben ki tudjuk használni: lehetővé teszi, hogy minden szokványos és nem szokványos funkcióját, beállítási lehetőségét el tudjuk érni és kihasználni. Például a 10.3.6.1. fejezetben bemutatott ProfiVox szintetizátor is rendelkezik ilyennel. Viszont egy ilyen interfészt változatlan formában rendszerint csak egy konkrét termékben találunk meg, sőt annak egyes verziói között is változik. Ezért ha egy felhasználói alkalmazás készítője a natív interfészt használja, azzal hozzáköti magát egy konkrét beszédtechnológiai modul használatához. Ha szeretne egy másik használatára átváltani – mert beszüntetik a korábban használt fejlesztését, vagy egy sokkal jobb elérhetővé válik –, ehhez jelentős módosításokat kell végeznie az alkalmazásában.

A beszédfelismerés, illetve beszédszintézis feladata elég jól körülhatárolható, ezért meg lehet határozni a funkciók és beállítások egy olyan halmazát, ami elégséges e feladat megvalósításához egy alkalmazásban. Egy szakértői csoport által meghatározott ilyen művelethalmaz adja a beszédtechnológiai modulok *szabványos interfész*einek alapját. Egy modul nem feltétlenül valósítja meg a teljes művelethalmazt, mivel a szabványos interfészeket annyira általános célúvá tervezik, hogy megfeleljenek egyrészt sok különféle tervezési szempontú modulnak, másrészt a közeljövőben várható kutatási eredmények is megvalósíthatók legyenek benne. Például a könyv írásának időpontjáig nincs kereskedelmi forgalomban jó minőségű beszédszintetizátor, amely többféle stílusban (hivatalos, lezser, gépies, izgatott, affektáló) képes beszélni, mégis a stílusok közötti váltás lehetősége több mint 15 éve megtalálható az MS SAPI interfészben. A szabványos interfészek specifikációjában pontosan meghatározzák, hogy egy programnak milyen funkciókat kell biztosítania ahhoz, hogy a specifikációnak megfelelő (compliant) legyen, illetve minimálisan mit kell biztosítania ahhoz, hogy kompatibilis (compatible) legyen. Tehát a kompatibilitás alacsonyabb szintű megvalósítást takar, mint a megfelelés. Egy modul összes, az egyedi interfészén keresztül elérhető funkciójához, beállításához általában nem található a

szabványos interfészben közvetlen funkcióhívás. Ez nem jelent gondot, ha a modul jó alapértelmezésekkel, illetve a beállítások automatikus megállapításával áthidalja ezt a hiányt. Az összes lehetőségre valószínűleg egyébként is csak a fejlesztés korai fázisában van szüksége a fejlesztőknek, egyébként pedig a beszédkutatást végzőknek fontosak. De szükség esetén meg lehet oldani, hogy minden lehetőség elérhető legyen a szabványos interfész használata esetén is: a szövegbe helyezett vezérlő szekvenciákkal, vagy az interfész által nyújtott bővítési lehetőségen keresztül.

A beszédtechnológiai modulokat szabványos interfészen keresztül használó alkalmazás készítőinek el kell dönteniük, hogy csak az interfésszel való kompatibilitást (a kötelező funkciókat) várják-e el, vagy a megfelelést (az összes funkció megvalósítását), esetleg egy modul egyedi funkcióit is igénybe veszik-e. Minél több funkciót várnak el egy modultól, a modulok annál szűkebb köre lesz használható számukra. Egy intelligens, de nagy körültekintést igénylő megoldás, ha az alkalmazás csak a kompatibilitást várja el, az ezen kívül eső funkciók meglétét ellenőrzi, és meglétük esetén felhasználja azokat extra szolgáltatásokban; sőt ha egy általa ismert beszédmodult csatlakoztatnak hozzá, akár annak egyedi funkcióit is kihasználja.

A szabványos interfész használata esetén bármely felhasználói program, amely betartja a szabvány rá eső részét, tetszőleges beszédtechnológiai modul szolgáltatását igénybe veheti, amely a szabvány másik felének betartását vállalja. Ez a megoldás előnyös a beszédmotorok készítőinek, mivel számos alkalmazásban használhatóvá válik a munkájuk. Hasznos továbbá a felhasználói alkalmazások készítőinek is, mivel a modulok széles körével működhet az alkalmazásuk, nincsenek egy céghez kötve; sőt a felhasználóik maguk választhatnak tetszésük szerinti modult a programhoz. Továbbá a felhasználóknak is hasznos ez a rugalmasabb konfigurálhatóság miatt, sőt közvetve a cégek termékei közötti szabadabb verseny révén is. Ha egy alkalmazás mégis csak egy bizonyos beszédmotort szeretne használni valamilyen okból (például mert csak ennél látják garantálva a minőséget), vagy fordítva, egy beszédmotort egy alkalmazáshoz szeretnének kötni (például mert a vele való terjesztésre szól a megállapodás), ez is megoldható: A beszédprogram ellenőrzi az őt hívó alkalmazás „kilétét”, és csak akkor működik vele együtt, ha ez a várt változat.

A szabványosítás a beszédtechnológiai modulok erőforrásaira, a kiejtési szótárakra és a beszédfelismerők nyelvtanaira is kiterjed; ezeknek a cserélhetősége, alkalmazások közötti hordozhatósága is hozzájárul a technológia rugalmasságához.

Az MS SAPI és a JSAPI viszonya. A két leglényegesebb beszédinterfész, a JSAPI és az MS SAPI között számos hasonlóság van. Mindkettő célja szabványos interfész nyújtása a beszédfelismerés- és beszédszintézis-technológiákhoz, PC-s és telefonos környezetben. Az MS SAPI született meg előbb, 1995-ben, a Windows 32 bites operációs rendszerein való használatra. Ezt kváziszabványnak nevezhetjük, mivel a Microsoft szoftvercég szakértői készítették, széles körű egyeztetés nélkül, de a Windows operációs rendszer elterjedtsége miatt nagy jelentőségű. A JSAPI létrehozásában a Sun Microsystems játszotta a főszerepet, hét céggel együttműködve, melyek a követ-

kezők: Apple Computer, AT&T, Dragon Systems, IBM Corporation, Novell, Philips Speech Processing és Texas Instruments.

Az MS SAPI-t elsősorban Windows környezet alatti futásra tervezték, de számos programnyelven keresztül elérhető: a C++ mellett Visual Basicből, szkriptnyelvekből (például JavaScript), és a .NET futtató környezetben működő úgynevezet managed code-ot eredményező nyelvekből (például C#). Az utóbbi valójában tetszőleges platformra hordozhatóvá teszi, amelyre a .NET futtató környezetet megvalósítják, és már jelenleg is számos operációs rendszeren működik.

A JSAPI a Java programnyelvhez készült könyvtár, amely a platformfüggetlen Java nyelvből használható. A JSAPI tartalmaz egy XML-alapú nyelvet, amely lehetővé teszi szövegek részletes felolvasási utasításokkal való ellátását; ugyanígy az MS SAPI újabb változatai is (az MS SAPI 5-től kezdve) rendelkeznek egy ugyanilyen célú XML-alapú nyelvvel.

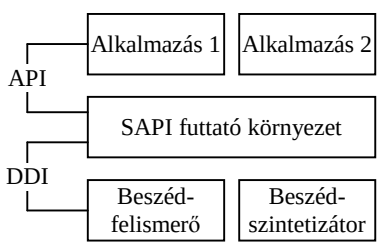
Ahogy látjuk, mindkét technológia célja azonos, mindkettő gyakorlatilag platformfüggetlenül használható, és szabványos technológiák használatára épít (például XML). A kettő között van bizonyos szintű átjárás is: egyes MS SAPI interfésszel rendelkező beszédmotorok elérése lehetséges JSAPI-n keresztül egy kiegészítő szoftver (Cloud Garden) segítségével. Most egyenként részletesebben bemutatjuk a két technológiát.

13.2.1. Microsoft Speech API

A Microsoft Speech API (MS SAPI) beszédinterfésznek két fő változata létezik. Az első az 1995 és 1998 között létrehozott változat, amelyek az 1-től 4-es verziószámig terjednek, a legutolsó változat az MS SAPI 4.1. A különböző változatai között csak bővítéseken ment keresztül, ezért ezek egymás között visszafelé kompatibilisek. Ez csak a Windows 32 bites változatai alatt használható, a Windows 95 és Windows NT 3.51-től kezdve. Az MS SAPI 5-ös verziójú változatát 2000-ben adták ki: ez az előző verzióktól teljesen független, számos szempontból módosított elképzelésen alapuló interfész (lásd részletesebben később). A Windows 98 és Windows NT 4.0-tól használható, egy új változata a Windows CE mobil operációs rendszer alatt is képes futni. A könyv írásának időpontjában a legfrissebb verziója az MS SAPI 5.4 változat. Ennek a különböző változatai szintén visszafelé kompatibilisek egymással.

A két fő MS SAPI változat, a 4-es és 5-ös verziók alapelvei között sok a közös, de az utóbbi lényeges újításokat is hozott. Közös bennük, hogy lehetővé teszik a beszédmotoroknak, hogy regisztrálják magukat az operációs rendszerben, a felhasználói alkalmazások pedig különböző tulajdonságaik alapján megkereshetik és használhatják az elérhető változatokat. A kiválasztás szempontja lehet a beszédmotor számos tulajdonsága, mint például a nyelve, az elérhető funkciók, a gyártó cég

neve, beszédszintetizátor esetén a beszédhangot nyújtó személy neme, kora, vagy a beszédhang megnevezése. Az alkalmazás szolgáltatásokat tud igényelni beszédmotortól a függvényei meghívásával, valamint átadhat eseménynyelőket (event sink), amelyen keresztül a beszédmotor értesítheti az alkalmazást a művelet elvégzésének fázisairól. Az MS SAPI 4-ig az interfész csak az alkalmazás és a beszédmotor összekapcsolódásában segített, ezután a beszédmotor és az alkalmazás közvetlenül kommunikáltak egymással. Ennek számos következménye van: többek között a beszédmotornak védekeznie kellett az alkalmazás hibái ellen, ellenkező esetben instabillá válhatott, ha az alkalmazás rossz bemenetet adott; valamint a minden beszédmotorra közös feladatok megvalósítását is minden komponensgyártónak egyenként újra és újra el kellett végeznie. Ilyen minden beszédszintetizátorra közös feladat például a bemeneti szövegbe elhelyezhető vezérlő címkék értelmezése, vagy egy kimondandó hangsorozat átadásakor a fonémaszimbólumok értelmezése. Az MS SAPI 5 változatban újdonság, hogy az interfészhez tartozó futtatható állomány (sapi.dll) közvetítőként végig jelen van a beszédmotor és az alkalmazás között; lásd a 13.1. ábrát. Így el tudja végezni az átadott adatok ellenőrzését, és csak a specifikációnak megfelelőeket adja át, ezért a szoftverek készítői számíthatnak arra, hogy mindig a specifikációnak megfelelő bemenetet kapnak. A bemeneti szöveget fel is dolgozza a sapi.dll, így például az XML formátumú beszédszintetizátor bemenetben a vezérlő parancsok elemzését (a 4-es verzióhoz képest újítás az XML formátumú bemenet is). Az inicializálás és vezérlés egyes feladatait is ellátja, így például ha a beszédmotor esetleg lefagy, észrevétlenül újraindítja. Továbbá teljesebb szétválasztást követel meg a beszédmotor és a felhasználói alkalmazás között, így gátolva a fejlesztés során használt változatok túlzott egymásra épülését, hogy biztosan együttműködjenek tetszőleges másik SAPI 5 kompatibilis szoftverrel. Ha a beszédfelismerőt egyidejűleg több alkalmazás szeretné használni, azt egy megosztott objektumként működteti, így erőforrás-takarékosabb, és nincs versengés a mikrofonért. Több más újítás is található a SAPI 5 változatban, amely az előző változatnál könnyebben programozhatóvá és erőforrás-takarékosabbá teszi. Az MS SAPI 4 számos COM (Component Object



13.1. ábra. Az MS SAPI 5 futtatható állományának szerepe: az alkalmazások felé az API (Application Programming Interface), a beszédmodulok felé a DDI (Device Driver Interface) megvalósítása. Forrás: az MS SAPI 5 dokumentációja

Model) objektumot tartalmaz, amelyeket nyolc interfészcsoporton keresztül lehet elérni. Ezek magas szintű interfészek parancsszavas vezérléshez (Voice Command), diktáláshoz (Voice Dictation), beszédszintézishez (Voice Text) és telefonos alkalmazásokhoz (Voice Telephony), valamint alacsony szintű interfészek a beszédfelismerő, illetve beszédszintézis motorok nagyobb mértékű eléréséhez (DirectSpeechRecognition, DirectTextToSpeech), valamint bemeneti és kimeneti audioobjektumok (Audio Objects) és egyéb beszédes segédobjektumok (Speech Tools) kezeléséhez.

Az MS SAPI 5 interfész COM objektuma, és az ezekhez tartozó interfészek két csoportba vannak osztva: az egyik csoport a beszédmotorokat jeleníti meg az alkalmazások számára (Application-Level Interfaces, alkalmazásszintű interfészek), a másik csoport tagjait a beszédmotor készítőinek kell megvalósítaniuk (Engine-Level Interfaces). Ahogy fentebb írtuk, a SAPI 5 futtatható kódja ellenőrzéseket és feldolgozásokat követően közvetíti a hívásokat a kettő között. A beszédmotor-szintű interfészek: beszédfelismerő (Speech Recognition), beszédfelismerő motor (Speech Recognition Engine), beszédszintetizátor motor (Text-to-Speech), nyelvtan-fordító beszédfelismerőkhöz (Grammar Compiler), erőforráskezelő (Resource). Az alkalmazásszintű interfészek között szintén megjelennek az előzők, de az alkalmazás számára szükséges tartalommal, valamint: kiejtésiszótár-kezelő (Lexicon), eseménykezelő (Eventing), audio (Audio) interfészek. Ezen kívül tartalmaz számos adatstruktúrát, konstanst és segédfüggvényeket, amelyek használhatók a programok elkészítésekor.

Az MS SAPI egyes verziói külön telepítést igényelnek, az újabb változatai a Windows operációs rendszer részét képezik. Ha telepítésre van szükség, telepítheti az ezt használó alkalmazás vagy beszédmotor telepítője, vagy telepíthetjük a szoftverfejlesztéshez használható SDK (Software Development Kit, szoftverfejlesztői-csomag) használatával. Az MS SAPI 4-hez az MS SAPI 4 SDK használható. Az MS SAPI 5 első változatai a Speech SDK 5-ös verziójának részei, az újabb változatai közül pedig a gépi kódú változat a Windows SDK 5-nek és magasabb verzióinak része, míg a managed code API a .NET 3-nak és magasabb verzióinak része.

A szoftverfejlesztéshez használható SDK-kban a beszéd vezérlését végző állományok és interfészleírók mellett vannak példa forráskódok beszédmotorokhoz és felhasználói alkalmazásokhoz, tesztprogramok az újonnan készülő motorok megfelelési szintjének ellenőrzésére, részletes dokumentáció, működő beszédmotorok és vezérlőpultba épülő kezelőprogram, hogy csak a lényegesebbeket említsük.

Források: `http://www.microsoft.com/speech/` `http://en.wikipedia.org/wiki/MS_SAPI`

13.2.2. Java Speech API

A Java Speech API (JSAPI) első változatát 1998-ban adták ki. A célja egy több platformon működni képes (cross-platform) interfész létrehozása volt, az MS SAPI-hoz hasonlóan beszédszintetizátorok, beszédfelismerők és telefónia használatára.

A Java objektum hierarchián belül a javax.speech csomagban, ezen belül a javax.speech.recognition és javax.speech.synthesis csomagokban találhatók az interfész objektumai. A beszédfelismerő-csomag tartalmaz interfészeket a nyelvtanok kezeléséhez (Grammar, DictationGrammar, RuleGrammar), a beszédfelismerőhöz (Recognizer, RecognizerProperties), a felismerés eredményének tárolásához (Result, FinalResult, FinalDictationResult, FinalRuleResult, ResultToken) és eseményfigyelőkhöz (ez felel meg az MS SAPI eseménynyelőjének; GrammarListener, RecognizerAudioListener, RecognizerListener, ResultListener) és a beszélők profiljainak kezeléséhez (SpeakerManager). A beszédszintézis-csomag interfészei: a felolvasandó szöveg lekérése (Speakable), a szintetizátor kezelői (Synthesizer, SynthesizerProperties) és az eseményfigyelők (SpeakableListener, SynthesizerListener).

A beszédszintetizátorok bemeneti szövegét Java Speech API Markup Language (JSML) segítségével lehet felcímkézni, ami egy XML-alapú nyelv. Ez tartalmaz elemeket a dokumentum struktúrájának leírására, a szavak, kifejezések kiejtésének megadására, kifejezések határának és hangsúlyoknak a jelölésére, a hangmagasság, beszédtempó, és egyéb beszédjellemzők előírására. A beszédfelismerők nyelvtanának leírására szolgál a Java Speech API Grammar Format (JSGF).

Források: `http://java.sun.com/products/java-media/speech` `http://en.wikipedia.org/wiki/Java_Speech_API`

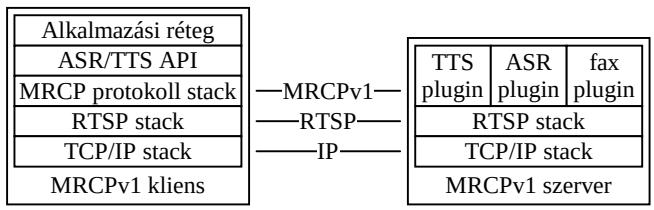
13.3. MRCP

Fegyó Tibor

Az MRCP (Media Resource Control Protocol) a gépi beszédfelismerők (ASR) és beszédszintetizátorok (TTS) médiaszerverben történő alkalmazásához nyújt segítséget, mint kifejezetten erre a célra kifejlesztett hálózati protokoll. Az alábbiakban röviden ismertetjük az MRCP alapvető felépítését és működését.

Az MRCP (rfc4463) a Cisco Systems, a Nuance és a SpeechWorks közös fejlesztésének eredménye, egy olyan beszédszerverek (Speech Server) által használt kommunikációs protokoll, melyen keresztül a szerver beszédszolgáltatásokat tud nyújtani a hozzá kapcsolódó klienseknek. Ezek a szolgáltatások tipikusan a gépi beszédfelismerés és a beszédszintézis. A kliensek hálózaton keresztül MRCP üzenetet küldenek a szerver számára egy másik protokoll segítségével. Ezek az RTSP – Real Time Streaming Protocol (rfc2326) és a SIP – Session Initiation Protocol (rfc3261). Az

MRCP-nek két verziója fejlődött ki az egyik verzió (13.2. ábra) RTSP-t, a másik (13.3. ábra) SIP-et alkalmaz.



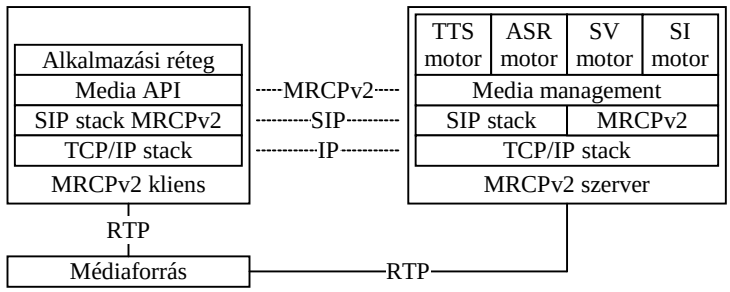
13.2. ábra. Az MRCP első változatának felépítése

Beszédfelismerő (ASR). A szerverben az ASR erőforrás feladata, hogy beszédjelet adott szabályok szerint (nyelvtanok) szöveggé alakítsa, az eredményt pedig visszaküldje a kliensnek. A felismeréskérés a RECOGNIZE üzenettel valósul meg, amely tartalmazza a nyelvtanok definícióját. A nyelvtan megadható külön is (DEFINE-GRAMMAR), ekkor statikus nyelvtanról beszélünk, ilyenkor a RECOGNIZE üzenetben erre a nyelvtanra kell hivatkozni. Inline nyelvtan esetén a RECOGNIZE üzenet törzsében kötelező definiálni a nyelvtan(oka)t. A szerver a RECOGNITION-COMPLETE üzenetben jelzi, amikor a nyelvtanban definiáltaknak megfelelően a felismerés megtörtént. Kétféle nyelvtantípus definiálható: beszédhangra vonatkozó (voice) és DTMF. Az MRCP a DTMF detektálást/felismerést a beszédfelismeréssel teljesen azonos módon kezeli.
Szövegfelolvasó (TTS). A szerverben a TTS feladata – az ASR fordított funkciójaként –, hogy a kliens által elküldött szöveget beszédjellé alakítsa, majd visszaküldje. A SPEAK üzenet paramétere tartalmazza a szintetizálni kívánt szöveget, majd a teljes beszédjel elküldése után a szerver SPEAK-COMPLETE válasszal jelzi kliensnek, hogy az utolsó beszédkeretet is elküldte.

A fentiek szerint az MRCP szerver alkalmas egyéb erőforrások kezelésére is: a felismerő jeldetektorként, a beszédszintetizátor pedig szintjelgenerátorként is használható, vagy akár fax kezelésére is kiterjeszthető. A 13.3. ábrán az MRCP a TTS és az ASR erőforrások kiterjesztett kezelését teszi lehetővé, biztosítva néhány új utasítást. Ezenkívül külön hangfelvevő (Recorder) és beszélőszemély-igazoló és -azonosító (Speaker Verification and Identification [SV/SI]) erőforrás-kezelést is magába foglal.

Az MRCP a HTTP-hez hasonlóan, kérés-válasz elven működik. A válasz lehet egyszerű nyugtázás, de különböző információkat is tartalmazhat. Formátumát tekintve egyszerű szöveg (text) formátumot használ, melynek felépítése három részre tagolható:

- *Első sor:* Az első sor jelzi az üzenet típusát és az MRCP verziószámát. A válaszüzenet egyéb információkat is tartalmazhat, mint például az állapotstátuszt, illetve annak kódját. Ezenkívül a sorban szerepel az üzenetkérés egyedi azonosító száma,



13.3. ábra. Az MRCP második változatának felépítése

melyet a válasznak is tartalmaznia kell. Ennek alapján a kérés-válasz összerendelés elvégezhető.

- *Üzenet fejléc:* A fejléc különböző paramétereket tartalmaz *<paraméter>:<érték>* formátumban. Ezek lehetnek kötelező paraméterek, mint például – amennyiben az üzenet tartalmaz törzs részt – a szövegtörzs típusa, mérete és opcionálisan megadható, az üzenettől függő beállítások.
- *Üzenettörzs:* Az üzenet törzse a fejlécnek megfelelő további szöveges információt tartalmaz, mint például beszédfelismerés esetén a nyelvtan leírását.

Minden MRCP alkalmazás értelemszerűen – mivel beszédfeldolgozásra fejlesztették ki – tartalmaz hang-, illetve beszédátvitelt, melynek megvalósításával a protokoll nem foglalkozik. Magát a beszédátvitelt erre alkalmas más kapcsolódó protokollok végzik, leggyakrabban az RTP – Real-time Transfer Protocol (rfc3550). A DTMF jel ugyancsak RTP keretekben kerül átvitelre, akár a beszédjelbe ágyazva (in-band), akár egyéb módon (rfc2833).

13.4. Intelligens beszédhang-időtartam mérő

Abari Kálmán

Az ismertetésre kerülő program (HIDOL) célja, hogy segítse a kutatói, oktatói munkát. Letölthető a `http:\magyarbeszed.tmit.bme.hu` honlapról (oktatási, kutatási célra ingyenesen használható). A HIDOL olyan eszköz, amellyel gyorsan és pontosan lehet hangidőtartamokat (hangkörnyezetfüggően is) megmérni egy előre elkészített, felcímkézett hangadatbázisban. Hangidőtartam-mérésekhez célszerű sok hangot tartalmazó beszédadatbázist készíteni, hogy a mérések hitelesek legyenek. A program csak a saját beszédadatbázisával működik (egy férfi beszélő). Az adatbázis hang- és szószinten kézzel címkézett beszéd- és szöveganyag (480 mondat, 2950 szó, 12 364 hang, 16 percnyi beszéd). Az adatbázison végzett mérések alapján már egy könyv is született (Olaszy 2006b). Jelenleg ez az egyetlen ilyen nyilvános adatbázis Magyar-

országon. A HIDOL program hangok, hangkapcsolatok, szavak és szünetek hosszát szolgáltatja különböző lekérdezési formákra adott válaszaiban. Működése során az eredményeket tabulált szöveges állományok (txt) formájában hozza létre, amelyek felhasználhatóak más statisztikai programokban az adatok összefüggéseinek kimutatására.

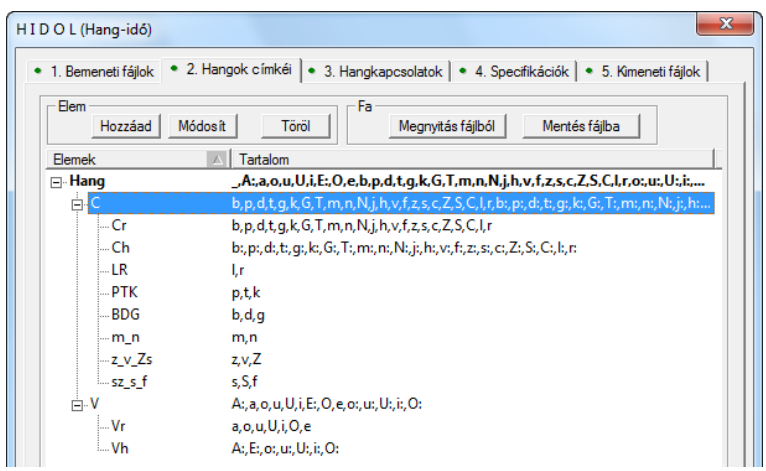
A program hármas hangcsoportok definiálását várja (megelőző hangok, mérendő, követő hangok) bemeneti specifikációként, és a középsőre vonatkozó időtartamot adja meg. A hang szót a továbbiakban a mérendő elemre vonatkoztatjuk, lehet például hangok csoportja is. A hangok jelölésére az E1-jelű szimbólumcsoportot alkalmazzuk.

Például a BBB_Vr_BBB hármas csoport 7 hangot definiál (B = bármilyen hang, Vr = rövid magánhangzó). Minden méréshez 7 hangra vonatkozó definíciót kell megadni, ezekből a középső hang (a negyedik) időtartamát méri a program. A program 5 funkcióját mutatjuk be röviden.

Mérés a programmal. A program használata 1–2 órán belül megtanulható, utána csak pár percet vesz igénybe egy-egy mérés konfigurálása. Windows környezetben használható (hidol.exe).

Bemeneti fájlok. A Hozzáadás gombra kattintva hívhatjuk be a felcímkézett beszédadatbázist (egészében vagy részleteiben) a kiválasztott könyvtárból. A Könyvtár teljes könyvtárakat jelöl ki, míg az Állomány bármelyik könyvtár egyes wav-fájljait. A feltöltés eredménye: az üres ablak tele lesz fájlnevekkel, amelyeken a méréseket végezzük. A kijelölt fájlt akár meg is hallgathatjuk.

Hangok címkéi. Itt kell meghatározni, hogy milyen hangok, hangcsoportok időtartamát akarjuk mérni. A leggyakoribb hangcsoport-meghatározások már be vannak építve (13.4. ábra), de bármilyen csoport (akár egyetlen hang is) megadható. Az Ele-



13.4. ábra. Lekérdezhető hangcsoportok meghatározása

mek oldalon láthatók a hangmeghatározások nevei (ezekkel hivatkozunk a csoportra, ezek tetszőlegesen adhatók meg. A Tartalom oldalon pedig azok a hangok jelennek meg, amelyek a név alatti csoportba tartoznak. Példa: a legmagasabb szinten lévő Hang jelzésű csoportba minden hang benne van, a Vr csopotban csak a rövid magánhangzók stb. A Hozzáad gombbal új csoportot definiálhatunk, a Módosít gombbal valamely meglévőt módosíthatjuk (erre ritkán, csak speciális mérések esetében van szükség).

Hangkapcsolatok. Itt kell meghatározni a mérendő hangot megelőző hármas csoportot, illetve a követő hármast (13.4. ábra). Az Elem vezérlőcsoportablak arra szolgál, hogy új hármas együttest határozzunk meg. Erre akkor van szükség ha nem találunk olyan hármast, amelyikre szükségünk van a mérés definiálásához. Az új elem mindig három hangot tartalmaz, ezeket kell definiálni. Ezután a Hozzáad gombra kattintva az új elemhármas bekerül a baloldali nagy listába. Új elem hozzáadása után célszerű eltárolni a hangkapcsolatok.txt fájlt, hogy legközelebb is meglegyen a meghatározott új elemhármasunk.

Specifikációk. Az előző pontokban csak a mérés előkészítését végeztük. A legtöbb esetben ezeken a pontokon kattintással át kell menni, mivel elég sokféle hangmérési variációt már kidolgoztunk a program számára. A Specifikáció fülre kattintva definiáljuk a tényleges mérést, azaz meghatározzuk azt a 7 hangot (3+1+3) amelyikből a negyediknek az időtartamát fogjuk megmérni.

A Közös jellemzők címszóval jelölt területen meg kell adnunk egy fájlnevet, amibe a program összegyűjti az adatokat. Legyen ez például CeC.txt (bármilyen név adható) a mérés bemutatására.

Kimeneti fájlok. Itt végzi el a program a mérést. A Konverzió indítása gombbal átfésüli az összes wav-fájlt és feltérképezi a mérendő e jelű hangokat. A Lekérdezés indítása gombbal megméri és kiírja a megadott fájlba az eredményeket. Az Eredmény megtekintése gombbal megnézhető az eredmény. Ennek különböző mélységi fokozatai vannak. A legtömörebb megjelenítés összesen négy adatot közöl az e jelű hangra a következők szerint:

SPEC.NEV: CeC
ÖSSZ(db): 1137
ÁTL.(ms): 80
MIN.(ms): 30
MAX.(ms): 163

Eszerint 1137 e jelű hangot talált mássalhangzók közötti helyzetben. Az átlagidőtartam ezekre 80 ms. A legrövidebb 30 ms, a leghosszabb 163 ms. A Részletek megtekintése bekapcsolásával a program részletezi az előbbi eredményeket, megmutatja, hogy melyik mondatban és melyik szóban találta meg a szélső értékeket.

SPEC.NEV: CeC
ÖSSZ(db): 1137
ÁTL.(ms): 80

MIN.(ms): 30
-=Mondat(478): moSdbefejezedamunkA:t_
Szó(2726): befejezed
Hang(11925): e
MAX.(ms): 163
-=Mondat(71): nem_
Szó(381): nem
Hang(1505): e

A mondatok, szavak, hangok sorszámozva vannak. Így akár meg is kereshetjük az adott mondatban, szóban a kérdéses hangot, meghallgathatjuk a mondatot stb. A legrészletesebb eredményt az Eloszlás megjelenítése jelölőnégyzetrevaló kattintással kaphatjuk meg:

SPEC.NEV: CeC
ÖSSZ(db): 1137
ÁTL.(ms): 80
MIN.(ms): 30
MAX.(ms): 163
ELOSZLÁSOK(db):
0–10: 0
10–20: 0
20–30: 1
30–40: 1
40–50: 38
50–60: 103
60–70: 218
70–80: 258
80–90: 206
90–100: 149
100–110: 85
110–120: 37
120–130: 18
130–140: 14
140–150: 7
150–160: 1
160–170: 1

Ilyenkor a program megadja a mért hang időtartam szerinti eloszlását. Ehhez az időtengelyt 10 ms-os sávokra osztja. Az adatokból látszik, hogy a legtöbb e jelű hang a 60–90 ms közötti tartományban van. Ha mindhárom lehetőséget egyszerre bekapcsoljuk akkor az eloszlásra vonatkozó mondatokat, szavakat is megkaphatjuk, így a hullámforma szintjéig visszamenve minden információ rendelkezésünkre áll a mért hangidőtartammal kapcsolatban.

13.5. Glottalizáló program

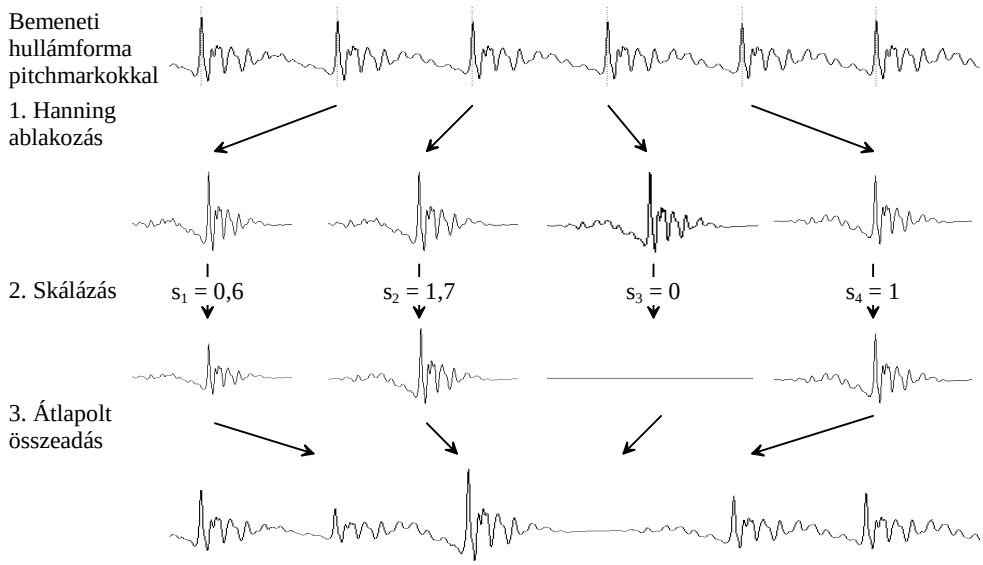
Bőhm Tamás

Ez a fejezet egy olyan ingyenesen elérhető programot mutat be (Glottalizer), amely reguláris zöngével képzett beszédrészleteket képes irregulárissá alakítani, azaz glottalizált hangzást előállítani (lásd az 5.3.3. fejezetet). Elérhető a `http://www.bohm.hu/glottalizer.html` címről. A program által alkalmazott feldolgozási módszer az egyes alapperiódusok amplitúdójának külön-külön történő skálázásán alapul. A grafikus felhasználói felület Nicolas Audibert (GIPSA-lab, Grenoble) munkája, míg a beszédfeldolgozó módszert a BME TMIT-en valósították meg (Bőhm et al. 2008).

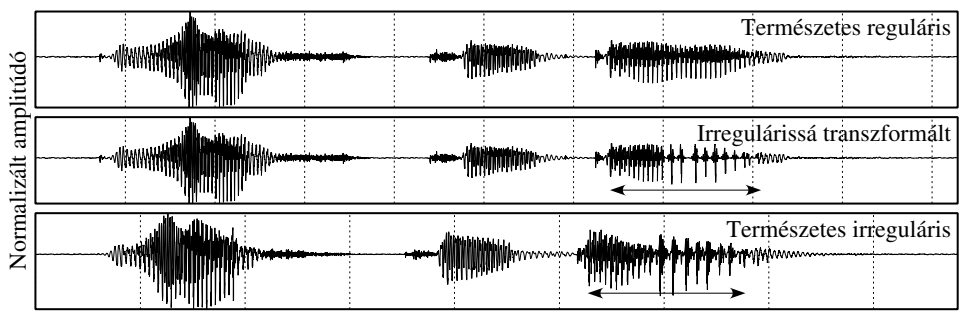
A működés elve. A szoftver zöngeszinkron-ablakozással megközelítőleg szétválasztja az alapperiódusokat, majd az egyes periódusok amplitúdóját egyenként skálázza egy periódusonként beállított faktorral (gyakran nulla szintre, azaz egyes periódusokat kitörölve a jelből), végül átlapolt összeadással (overlap-and-add) előállítja a kimenő beszédjelet (Bőhm 2009). Bemenete a beszédjel és a glottális impulzusok hozzávetőleges időpontjai a transzformálni kívánt régióban. Az első lépés az egyes alapperiódusok „szétválasztása" a glottális impulzusok környezetének Hanning-ablakolásával. Ehhez zöngeperiódus-jelzőket (pitchmark) kell elhelyezni az időfüggvényen. Az ablak ($w_i[n]$) csúcspontját a periódus időfüggvényének maximumánál (jelzőpont) éri el, és az előző ilyen jelzőponttól a következőig tart (tehát két alapperiódust fed le és aszimmetrikus is lehet). Ez az ablakozási eljárás – amely megegyezik a PSOLA algoritmus (7.1.4. fejezet) analízis szakaszával – az egyes alapperiódusok egy durva közelítését nyeri ki különálló jelekbe. Ezután minden egyes jelet mintánként beszorzunk egy kézzel beállított skálázó faktorral (s_i), majd az egyes jeleket átfedve és összeadva újra összeállítjuk a teljes beszédjelet.

A kimenő beszédjel. A skálázó faktorok az egyes periódusokat erősíthetik (s_i>1), csillapíthatják (s_i <1), kitörölhetik (s_i=0) vagy módosítás nélkül meghagyhatják (s_i=1). Az eljárás illusztrációja látható a 13.5. ábrán, míg a 13.6. ábra b) része egy transzformált beszédfelvételt ábrázol. Egy-egy alapperiódus csillapítása vagy éppen nullázása a periódusban jelen lévő háttérzaj szintjét is lecsökkenti (például több egymás utáni periódus törlésekor a zaj hiánya a természetesség érzetét csökkentheti). Ennek az elkerülésére az eljárás háttérzajt (b[n]) ad a csillapított és nullázott periódusokhoz. A zajt például a felvétel végéről lehet kivágni, majd ablakozás után (1 - s_i)-vel skálázva kompenzálja az eredeti periódus skálázásából adódó zajenergia-veszteséget.

A skálázó faktorok félautomatikusan beállíthatók „impulzusmintázat-másolással". Ennek lényege, hogy a kialakítandó impulzusmintázatot (a glottális impulzusok időbeli távolságai és amplitúdói) egy természetes beszédben előforduló irreguláris régió alapján modellezzük – azaz egy modellfelvételt használunk, amelynek impulzusmintázatát a javasolt módszerrel átmásoljuk a transzformálandó



13.5. ábra. A beszédjelben irreguláris periódusokat létrehozó transzformációs eljárás illusztrációja

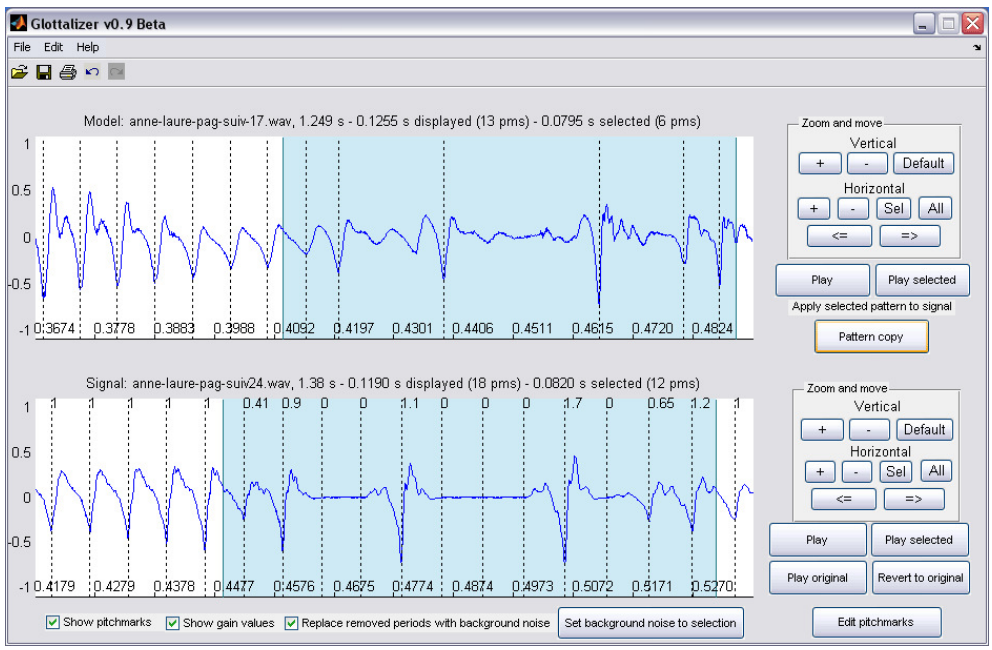


13.6. ábra. Egy reguláris zöngével végződő beszédrészlet (fent) és annak irregulárissá transzformált változata (középen), valamint egy természetes irreguláris beszédrészlet (lent). A vízszintes nyilak az irreguláris szakaszokat jelölik. A vízszintes tengely az időt 0,1 s-os felbontásban jelöli

beszédrészletre. Az impulzusmintázat kezdetben egy olyan vektor, amely a mintaként szolgáló irreguláris beszédrészlet glottális impulzusainak relatív amplitúdóját tartalmazza. Ha egy irreguláris alapperiódus jelentősen hosszabb, mint a $T0_{ref}$ referencia-periódusidő (például kétszer vagy háromszor olyan hosszú, mint a referencia, amit néhány megelőző periódus átlagaként számol), akkor az eljárás megfelelő számú nullát szúr be az impulzusmintázatba – ezeken a helyeken ugyanis egy vagy több periódust törölni kell a jelből. Érzeti és akusztikai tesztek eredményei azt mutatták, hogy a fenti transzformációs eljárás egyrészt képes bizonyos

mértékben reprodukálni az irreguláris zönge ismert akusztikai jellemzőit, másrészt a transzformált beszédet a kísérleti személyek természetes irreguláris zöngéhez hasonlónak ítélik (a rekedtség és természetesség szempontjából).

A *Glottalizer* nevű program a fenti eljárás hatékony alkalmazását teszi lehetővé. Windows operációs rendszer alatt futtatható és nem kereskedelmi célra ingyenesen elérhető (egy részletes használati útmutatóval együtt). A program lehetővé teszi (a) a módosítandó és a modellhullámforma párhuzamos megjelenítését, (b) impulzusmintázatok másolását és (c) a skálázó faktorok iteratív beállítását, amely során a paraméterváltozások hatása azonnal látható és hallható. Ezenfelül a programban elérhetőek a szokásos hangmegjelenítő és lejátszó funkciók, valamint a kiadott parancsok visszavonhatóak.



13.7. ábra. A Glottalizer program ablaka a modellfelvétel (felső panel) kijelölt szakaszáról a másik felvételre (alsó panel) történő impulzusmintázat-másolás után; a skálázó faktorok a jelzőpontok (szaggatott vonalak) felett láthatóak

A 13.7. ábrán a program működési felülete látható. Az alsó panel a módosítandó felvétel hullámformáját jeleníti meg. A felső panel a modellfelvételt ábrázolja, amely segíthet a transzformáció megvalósításában (akár kézi faktorbeállítás esetén, akár impulzusmintázat-másolással). Ebbe a panelbe egy irreguláris zöngét tartalmazó modellfelvétel tölthető be. Bár a másik felvétel transzformációját segítheti, a modellfelvétel nem módosítható. Ahhoz, hogy bármelyik panelbe betöltsünk egy hangfelvételt, a hozzá tartozó jelzőpontfájl is elérhető kell, hogy legyen (például Praat Point-

Process formátumban). A jelzőpontok a hullámformán megjeleníthetőek, szerkeszthetőek és elmenthetőek. Az alsó panelen egyszerű egérkattintásokkal az egyes periódusok amplitúdóskálázhatóak, kitörölhetőek (nulla faktorral skálázás) és visszaállíthatóak eredeti formájukba (azaz a skálázó faktor bármikor 1-re állítható). Az alkalmazott skálázó faktorok a módosított hullámforma felett láthatóak és egy külön fájlba menthetőek (ami később újra beolvasható). A transzformációt impulzusmintázat-másolással is elvégezhetjük. Ehhez ki kell választani a modell-hullámformán azt a tartományt, ahonnan az impulzusmintázat kinyerhető, valamint az alsó panelen azt a régiót, amire a mintázatot alkalmazni szeretnénk. Az impulzusmintázat-másoláshoz szükséges továbbá, hogy elegendő jelzőpont előzze meg a modellfelvételen kijelölt szakaszt, hogy a referenciaértékeket ki lehessen számítani.

13.6. A könyvben szereplő honlapok beszédkutatáshoz, oktatáshoz, fejlesztésekhez, döntéshozatalhoz

`http://magyarbeszed.tmit.bme.hu` 8.4.3. . 5.1.2. 10.3.5. 13.4. , 8.4.2. fejezet
`http://alpha.tmit.bme.hu/speech/hdbbabel.php` 8.1.1.1. fejezet
`http://alpha.tmit.bme.hu/speech/hdbspeechdt.php` 8.1.1.1. fejezet
`http://alpha.tmit.bme.hu/speech/hdbMTBA.php` 8.1.1.1. fejezet
`http://alpha.tmit.bme.hu/speech/hdbtesztelen.php` 8.1.1.1. fejezet
`http://alpha.tmit.bme.hu/speech/hdbMRBA.php` 8.1.1.1. fejezet
`http://alpha.tmit.bme.hu/speech/paperc013.php` 8.1.1.1. fejezet
`http://speechlab.tmit.bme.hu/hmm/` 10.3.8. 10.3.5.1. és 10.1. fejezetek
`http://fonetika.nytud.hu/cccc` 8.4.1. 5.2.3. 5.1.1.2. és 2.3. fejezetek
`http://fonetika.nytud.hu/cvvc` 8.4.1. 5.1.1.2. és 8.4.4. fejezetek
`http://hlt-platform.hu/offline-adatbazisok.html` 8.1.1.2. fejezet
`http://www.xbox.com/en-US/community/events/e3/kinect.htm` 11. fejezet
`http://catalog.elra.info/` 8. fejezet
`http://www.ldc.upenn.edu/` 8. fejezet
`http://www.squale.org/` 8.1.1.1. fejezet
`http://www.etca.fr/CTA/gip/Projets/Transcriber/` 8.1.1.1. fejezet
`http://www.xces.org/` 8.1.1.2. fejezet
`http://corpus.nytud.hu/mnsz/` 8.1.1.2. fejezet
`http://www.inf.u-szeged.hu/projectdirs/hlt/` 8.1.1.2. fejezet
`http://mokk.bme.hu/resources/webcorpus` 8.1.1.2. fejezet
`http://www.auralog.com/us/schools.html` 2.3. fejezet
`http://cslu.cse.ogi.edu/toolkit` 2.3. fejezet
`http://www.gyogyszervonal.hu` 12.3.5. fejezet
`http://vilaghallo.hu` 12.7.3. fejezet

http://www.mindroom.hu 12.4.2. fejezet
http://www.metnet.hu/ 12.4.1. fejezet
http://infoalap.hu 12.7.1. fejezet
http://www.phon.ucl.ac.uk/home/sampa/hungaria.htm 12.9.1. fejezet
http://alpha.tmit.bme.hu/speech/research.php 12.9.1. fejezet
http://rcs.hu/sc.htm 12.9.1. fejezet
http://hlt-platform.hu 12.4.3. és . fejezet
http://www.webforditas.hu 12.4.3. fejezet
http://www.bohm.hu/glottalizer.html 13.5. fejezet
http://www.microsoft.com/speech/ 13.2. fejezet
http://en.wikipedia.org/wiki/MS_SAPI 13.2. fejezet
http://java.sun.com/products/java-media/speech 13.2. fejezet
http://en.wikipedia.org/wiki/Java_Speech_API 13.2. fejezet
http://www.w3.org/2001/vxml 13.1. fejezet
http://cafe.bevocal.com/resources/voicexml_samples/index.html 13.1. fejezet
https://studio.tellme.com/library2/code/ 13.1. fejezet
http://developer.voicegenie.com/examples_VoiceGenie.php 13.1. fejezet
http://mazsola.iit.uni-miskolc.hu/~czap/mintak 10.3.1.6. fejezet
http://speechlab.tmit.bme.hu/hmm/ 10.3.8. 10.3.5.1. és 10.1. fejezetek
http://fonetika.nytud.hu/hitint 10.3.1.4. és 6.4. fejezet
http://fonetikai.nytud.hu 10.1. fejezet
http://blog.makezine.com/archive/2009/02/speech_synthesis_in_the_year_1939.html 10.1. fejezet
http://alpha.tmit.bme.hu/pub/multivox4/ 10.3.5.1. fejezet

14. fejezet
A beszédtechnológia jövője

Campbell Nick–Németh Géza

Ma már a beszédfelismerés és a beszédszintézis is megoldásokat kínál az iparnak és a társadalomnak, fontos mindennapi alkalmazásokról gondoskodnak. Ezen technológiák vegyítésének nagy piaca van, de a probléma még semmiképpen nem tekinthető megoldottnak. Egyelőre csak szűk, jól lehatárolt témakörökben lehet mindkettőt használni, távol állunk még attól, hogy a gépek képesek legyenek észlelni és érteni a mindennapi emberi beszélgetést.

A társalgás egy interaktív folyamat, amelyben kettő vagy több ember közösen alkotja meg osztott nézetét a világ egy apró részéről egy rövid időre. A beszélgetés a mindennapi társas együttműködés fontos része. A hétköznapi beszéd lehet egy-egy feladathoz kötött, de ez rendszerint nem egy olyan aszinkron folyamat, ahol egy ember beszél és a többiek csak hallgatják, vagy egymás után felszólalnak. Egy tipikus társalgásban mindkét fél egyszerre beszél az idő nagy részében, a mondatok darabokra hullanak, és úgynevezett „nyelvtaniatlan" darabkák képződnek, amint a résztvevők ténylegesen belemerülnek a beszélgetésbe.

Az egyszerű információcserével szemben a társalgásban együttműködő partnerek közös véleményt építenek gondolataik kölcsönös egyeztetésével. A folyamat egyik fő összetevője a visszacsatolás küldése, amellyel nyugtázzák a dialógus minden kis szakaszának megértését. Az állandó visszacsatolási folyamathoz a partner szoros megfigyelése szükséges, vagy ha távoli a beszélgetés (például telefonon keresztüli), akkor a hangszínezet és a beszédstílus apró változásainak fülelése segíthet. Ennek eredményeképpen a beszéd nagy része átlapolódó, és a beszélgető felek gyakran megszakítják egymást, hogy befejezzék a másik mondatát, vagy egyszerűen levágják, és nyitva hagyják a megnyilatkozás bizonyos részeit.

A beszédfelismerés alkalmas arra, hogy egy teljes elhangzott mondatot szöveggé alakítson, a beszédszintézis pedig megfelelően működik a jól formált szöveg beszéddé konvertálásában, de egyik technológia sem képes szembeszállni a természetes társalgás daraboltságával és befejezetlenségével. A beszédtechnológia egyelőre nem tartalmaz megfigyelő üzemmódot, habár a legfrissebb kutatások a multimodális

kommunikáció területén (mint például csoportos megbeszélések adatainak elemzése) kezdik beolvasztani a kamerákat és a mikrofonokat is a feldolgozó rendszerekbe. Azáltal, hogy megfigyeljük az emberek beszélgetésben történő együttes interakcióját, és hogy modellezzük a darabokra hullott és összefonódó folyamatokat, olyan technológiák tervezhetőek, amelyek kezdenek aktív szerepet vállalni az emberrel történő párbeszédben, vagy segíthetik az ember-ember beszélgetések feldolgozását.

A jelenlegi (korai) beszédtechnológiát aszinkron jellegű beszédfolyamat modellezésére tervezték. Például az ember-gép interfészek minden teljes információegységet egymástól függetlenül, felváltva dolgoznak fel, és teljes kifejezést várnak minden egyes lépésükhöz. Ezzel szemben a jövőbeli technológiáknak (például telefonos ügyfélszolgálat, robotok, játékok és embereket segítő eszközök) az emberi beszélgetőpartnerrel interaktívan kell majd kommunikálniuk. Az emberi beszélgetés „táncát" kell imitálniuk. Ehhez új, multimodális interfészparadigmákat is ki kell dolgozni. Talán úgy is fogalmazhatunk, hogy akár minden szoftveralkalmazás saját viselkedéssel rendelkezhet (például személyi interaktív naptár).

A beszédszintézisnek a jövőben rugalmasabbá kell válnia: a szöveg helyett inkább „cél" jellegű bemenet fogadása szükséges. A teljes rendszernek alkalmasnak kell lennie a beszéd kis szegmensekre darabolására, amelyek mindegyike a fenti cél egy kis különálló darabját hordozza. Emellett folyamatosan ellenőrizni kell, hogy a) a hallgató jelen van, b) a személy hallótávolságon belül van, c) figyelik a beszédet, d) sikeresen követik a beszélgetést, e) megértik és egyetértenek, vagy ha nem értenek egyet teljesen, akkor milyen mértékben teszik azt, és f) vajon szükséges-e visszalépni, ismételni, átfogalmazni, körülírni, egyszerűsíteni, vagy éppen kihagyni valamit annak érdekében, hogy a beszélgetés hatékonysága maximális legyen.

Az emberi beszélgetőpartnerek egy speciális elemet, a humort is hozzáadják a párbeszédhez, de talán a technológiának nem szükséges ilyen messzire mennie. A másik oldalon, a beszédfelismerés számára ez szinte kötelező. Ha egy társalgásra specializált beszédfelismerő nem ismeri fel a humort, csevejt és a beszélgetés háttércsatornáját vagy a visszacsatolási kísérleteket, akkor a bemeneti társalgási beszéd nagy részét nem fogja tudni helyesen feldolgozni.

Ezen fejlett beszédfeldolgozó eszközök támogatásához szükség lesz a társalgási beszéd szintaxisának kidolgozására. A töredék részek, félbehagyott és úgynevezett „nyelvtaniatlan" beszédrészletek, melyekbe akadozás és töltelékszavak is vegyülnek, nem egyszerűen a tökéletlen valós idejű emberi kognitív feldolgozás eredményei. Tulajdonképpen ez a kommunikáció leghatékonyabb formája, amely azért alakult ki, hogy a beszéd egydimenziós csatornáján összetett információt tudjunk továbbítani.

A nyelvtan fejlesztésének támogatásához szükség lesz olyan korpuszok gyűjtésére, amelyek jobban reprezentálják az emberi társalgásokat. Míg a beszédtechnológia jelentős része hatalmas korpuszokon történő statisztikai tanulás eredményén alapul, ezen korpuszok csekély hányada illusztrálja az emberi interakció társas aspektusait, ami pedig a való világban jóval több mint a felét kiteszi az interaktív helyzeteknek.

Alkalmazásokat lehet készíteni egy-egy feladatra specializált korpusz segítségével, de ahhoz, hogy igazán rugalmas technológiát hozzunk létre, amely a mindennapos emberi interakció észrevehetetlen és transzparens helyettesítésére képes, emberközelibb korpuszok szükségesek.

A következő években gyűjtött korpuszokat fel is kell majd címkézni. Jelenleg kizárólag az emberi hallgatók képesek a különféle beszédszegmensek megfelelő perceptuális osztályozására. Talán a nyers beszédanyag emberi felcímkézése az egyik legdrágább része a beszédtechnológiai fejlesztéseknek, de ahogy az igények előrehaladnak a pusztán lexikális feldolgozástól az inkább pszichológiai alapú feldolgozás felé, úgy kell ezt az elemet is automatizálni. Jelenleg is komoly munka folyik az érzelmek felismerésén, de az érzelmek egy-egy ember belső állapotától függenek, és nem feltétlenül a párbeszédben zajló kognitív interakció leglényegesebb vetületei feldolgozási szempontból. Emellett a társas és figyelmes attitűd állapotok – amelyek messzemenően eltérnek az érzelemtől – feldolgozása is szükséges.

A statisztikai módszereket már most is felhasználja a beszédtechnológiai fejlesztés, és ezek az erőforrások rendelkezésre fognak állni a beszédjel társas információjának feldolgozásához. Az eredmények nem csak a korpuszok annotálásában és kidolgozásában lesznek hasznosak, hanem magukban a beszédfeldolgozó eszközökben is. De a tanulóadatot ki kell egészíteni úgy, hogy lefedjen minden nyelvezetet, társadalmi osztályt és interakciótípust. Emellett elég rugalmasnak kell lennie ahhoz, hogy alkalmazkodni tudjon a nyelvhasználat megváltozásához, a stílus módosulásához és új közösségek létrejöttéhez. A beszédtechnológia jövője valószínűleg végtelen. Belátható időn belül nem várható, hogy az embert imitálni képes gépi megoldások létrejöjjenek. Ez a beszédtechnológián messze túlmutató kognitív modellezési kérdés. A beszédtechnológia ma körülbelül ott tart, mint a járműipar 1900-ban.

Irodalomjegyzék

Abari K.–Olaszy G. (2006): Internetes beszédadatbázis a magyar mássalhangzó kapcsolódások akusztikai szerkezetének bemutatására. In Alexin Z.–Csendes D. (szerk.) *IV. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, 213–222.

Abari K.–Olaszy G. (2007): A magyar beszéd hangkapcsolódásainak bemutatása az interneten. In Gósy M. (szerk.) *Beszédkutatás'2007*. Budapest, MTA Nyelvtudományi Intézet, 178–186.

Abari K.–Olaszy G.–Kiss G.–Zainkó Cs. (2006): Magyar kiejtési szótár az Interneten. In Alexin Z.–Csendes D. (szerk.) *IV. Magyar Számítógépes Nyelvészeti Konferencia*. 223–230.

Acero, A.–Stern, R. M. (1990): Environmental Robustness in Automatic Speech Recognition. In *Proc. ICASSP*. Albuquerque, New Mexico.

Adriaens, L. M. H. (1991): *Ein Modell deutscher Intonation*. PhD-dissertation (IPO Eindhoven).

Allauzen, C.–Riley, M.–Schalkwyk, J.–Skut, W.–Mohri, M. (2001): OpenFst: A general and efficient weighted finite-state transducer library. *Implementation and Application of Automata*, 11–23.

Allen, J.–Hunnicutt, M.–Klatt, D.–Armstrong, R.–Pisoni, D. (1987): *From text to speech: The MITalk system*. Cambridge University Press.

Alonso, A. M. (2004): *Spanish Text-to-Speech conversion system based on Profivox Technology*. M.sc. thesis (Budapesti Műszaki és Gazdaságtudományi Egyetem).

Alonso Frech, E.–Montesinos Guerrero, E.–Bernabé Martín, F. V.–Díaz-Guerra Vico, M. Á.–Bellido Pérez, F. J.–Becerra González, A.–Gavilán Casado, F. J.–Aranda Domínguez, Á. (1997): The Coverage Information System. 15. sz., *Comunicaciones de Telefónica I+D*.

Álvarez, A.–Cearreta, I.–López, J. M.–Arruti, A.–Lazkano, E.–Sierra, B.–Garay, N. (2007): A Comparison Using Different Speech Parameters in the Automatic Emotion Recognition Using Feature Subset Selection Based on Evolutionary Algorithms. In *Proc. TSD*. Pilsen, 423–430.

Andó É. (2002): *A történetmondás kommunikatív jellemzői*. PhD-értekezés (ELTE). Budapest.

Anguita, D.–Back, M. (1993): Propagation of an efficient implementation of the BP algorithm. Technical Report, University of Genova.

Arisoy, E.–Can, D.–Parlak, S.–Sak, H.–Saraclar, M. (2009): Turkish Broadcast News Transcription and Retrieval. 17. évf. 5. sz., *IEEE Transactions on Audio, Speech, and Language Processing*, 874–883.

Ascom Infrasys AG (1998): Speech Quality and its Objective Evaluation with PACE (Ascom White Paper Series Issue No. 103/98).

Ascom Infrasys AG (1999a): Quality of Service Acceptance Testing for Cellular Networks (Ascom White Paper Series Issue No. 101/96).

Ascom Infrasys AG (1999b): RXQUAL and Voice Quality (Ascom White Paper Series Issue No. 102/96).

Ascom Infrasys AG (2000): QVoice – System description (GSM).

Baayen, H.–Piepenbrock, R.–Van Rijn, H. (1993): *The CELEX lexical database — Dutch, English, German*. Linguistics Data Consortium, Philadelphia PA.

Baayen, R. (2001): *Word Frequency Distributions*. Kluwer.

Babarczy A.–Gábor B.–Hamp G.–Kárpáti A.–Rung A.–Szakadát I. (2005): Hunpars: mondattani elemző alkalmazás. In Alexin Z.–Csendes D. (szerk.) *III. Magyar Számítógépes Nyelvészeti Konferencia*. 20–28.

Bach I. (2002): *Formális nyelvek*. Typotex Kiadó.

Bahl, L. R.–Brown, P. F.–de Souza, P. V.–Mercer, R. L. (1986): Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proc. IEEE ICASSP*, vol. 1. 49–52.

Bánó M. (1916): Tetszőleges szöveg reprodukálására alkalmas beszélőgép. Magyar Szabadalmi Hivatal.

Barát T. (2001): Tolmács a hídon. egyetemi jegyzet, https://phigy.hu/node/849.

Bárczi G. (1963): *A magyar nyelv életrajza*. Gondolat Kiadó.

Batliner, A.–Burger, S.–Johne, B.–Kiessling, A. (1993): MüSLI: A Classification Scheme for Laryngealizations. In *Proc. ESCA Workshop on Prosody*. 176–179.

Baum, L. E.–Eagon, J. A. (1967): An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology. 73. évf. *Bulletin of the American Mathematical Society*, 360–362.

Bawab, Z.–Turicchia, L.–Stern, R.–Raj, B. (2009): Deriving Vocal Tract Shapes from Electromagnetic Articulograph Data via Geometric Adaptation and Matching. In *Proc. Interspeech*. Brighton, 2051–2054.

Becchetti, C.–Ricotti, L. P. (1999): *Speech Recognition*. Willey.

Beerends, J. G. (1998): Audio Quality Determination Based on Perceptual Measurement Techniques. In Kahrs, M.–Brandenburg, K. (szerk.) *Applications of Digital Signal Processing to Audio and Acoustics*. 1. fejezet. Kluwer.

Beerends, J. G.–Stemerdink, J. A. (1994): A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation. 42. évf. *Journal of the Audio Engineering Society*, 115–123.

Beesley, K. R.–Karttunen, L. (2003): *Finite State Morphology*. CSLI-Publications.

Beke A. (2008): A felolvasás és a spontán beszéd alaphangszerkezeteinek vizsgálata. In Gósy M. (szerk.) *Beszédkutatás'2008*. Budapest, MTA Nyelvtudományi Intézet, 93–107.

Beke A. (2009): A veláris magánhangzók stabilitása a spontán beszédben. In Gecső T.–Sárdi C. (szerk.) *A kommunikáció nyelvészeti aspektusai*. Tinta Könyvkiadó, 27–31.

Beke A.–Gráczi T. E. (2010): A magánhangzók semlegesedése a spontán beszédben. In Navracsics J. (szerk.) *NYELV - BESZÉD - ÍRÁS. Pszicholingvisztikai tanulmányok I*. Budapest, Tinta Könyvkiadó, 57–64.

Bekenstein, J. D. (2003): Information in the Holographic Universe. 289. évf. 2. sz., *Scientific American*, 58–65.

Békésy Gy. (1960): *Experiments in Hearing*. New York, McCraw-Hill.

Bellman R. E. (1957): *Dynamic Programming*. Princeton University Press.

Bennett, M. R.–Hacker, P. M. S. (2003): *Philosophical foundations of neuroscience*. Wiley.

Benoit, C.–Adjoudani, A.–Guiard-Marigny, T.–Le Goff, B.–Reveret, L. (1998): Multimodal integration for advanced multimedia interfaces, Reports ESPRIT III, Basic Research Project 8579.

Benoit, C.–Guiard-Marigny, T.–Le Goff, B.–Adjoudani, A. (1996): Which components of the face do humans and machines best speechread? In Stork, D. G.–Hennecke, M. E. (szerk.) *Speechreading by Humans and Machines*, vol. 150. NATO ASI, 315–328.

Bisani, M.–Ney, H. (2008): Joint-sequence models for grapheme-to-phoneme conversion. 50. évf. 5. sz., *Speech Communication*, 434–451.

Blauert, J. (1997): Binaural Technology: A technology with a view. In *Proc. Inter-Noise*, vol. II. Budapest, 1121–1128.

Boersma, P.–Weenink, D. (2009): Praat: doing phonetics by computer. Version 5.1.05. `http://www.praat.org`, Computer program.

Bogert, B. P.–Healy, M. J. R.–Tukey, J. W. (1963): The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In *Proc. Symposium on Time Series Analysis*. 209–243.

Bőhm, T.–Audibert, N.–Shattuck-Hufnagel, S.–Németh, G.–Aubergé, V. (2008): Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. In *Proc. Acoustics*. Paris, 6141–6146.

Bőhm T.–Ujváry I. (2008): Irreguláris fonáció előfordulása magyar beszédben, mint egyéni hangjellemző. In Gósy M. (szerk.) *Beszédkutatás'2008*. Budapest, MTA Nyelvtudományi Intézet, 108–120.

Bőhm T. (2009): *Irreguláris zöngével képzett beszéd vizsgálata és modellezése*. PhD-értekezés (Budapesti Műszaki és Gazdaságtudományi Egyetem).

Bőhm T.–Olaszy G. (2007): A magyar [v] hang szerkezetének és zörejességének fonetikai vizsgálata. In Gósy M. (szerk.) *Beszédkutatás'2007*. Budapest, MTA Nyelvtudományi Intézet, 19–34.

Bolla K. (1980): *Magyar hangalbum: A magyar beszédhangok artikulációs és akusztikai sajátságai*. Budapest, MTA Nyelvtudományi Intézet.

Bolla K. (1995): *Magyar fonetikai atlasz. A szegmentális hangszerkezet elemei*. Budapest, Tankönyvkiadó.

Bolla K. (1978): A magyar magánhangzók analízise és szintézise. In Bolla K. (szerk.) *Magyar Fonetikai Füzetek*. 1. köt. Budapest, MTA Nyelvtudományi Intézet, 53–68.

Bóna J. (2006): Tudunk-e változtatni spontán beszédünk tempóján? In Mártonfi A.–Papp K.–Slíz M. (szerk.) *101 írás Pusztai Ferenc tiszteletére*. Budapest, Argumentum, 560–566.

Bóna J. (2009): *A gyors beszéd*. Budapest, Lexica Kiadó.

Brandenburg, K.–Popp, H. (2000): An introduction to MPEG Layer-3. *EBU Technical Review*, 1–15.

Brants, T. (2000): TnT–a statistical part-of-speech tagger. In *Proc. Applied Natural Language Processing*. 224–231.

Burger, S.–Sloane, Z. A.–Yang, J. (2006): Competitive Evaluation of Commercially Available Speech Recognizers in Multiple Languages. In *Proc. LREC*. 809–814.

Burkhardt, F.–Paeschke, A.–Rolfes, M.–Sendlmeier, W.–Weiss, B. (2005): A Database of German Emotional Speech. In *Proc. Interspeech*. Lissabon, 1517–1520.

Cai, J.–Laprie, Y.–Busset, J.–Hirsch, F. (2009): Articulatory Modeling Based on Semi-polar Coordinates and Guided PCA Technique. In *Proc. Interspeech*. Brighton, 56–59.

Campbell, N. (2004): Getting to the Heart of the Matter. Keynote Speech. In *Proc. LREC*. 221–228.

Campbell, N. (2005): Getting to the heart of the matter; speech as the expression of affect; rather than just text or language. In *Proc. LREC*, vol. 1. 109–118.

Campbell, N. (2007a): Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse. In *COST Action 2102 International Conference on Verbal and Nonverbal Features*. Patras, Greece, 107–120.

Campbell, N. (2007b): The Role and Use of Speech Gestures in Discourse. 32. évf. 4. sz., *Archives of Acoustics*, 803–814.

Campbell, N.–Black, A. (1995): Prosody and the selection of source units for concatenative synthesis. In van Santen, J.–Sproat, R.–Olive, J.–Hirschberg, J. (szerk.) *Progress in Speech Synthesis*. Springer.

Campbell, R.–Dodd, B.–Burnham, D. (1998): *Hearing by Eye II.* Psychology Press.

Carney, E. (1994): *A survey of English spelling*. Routledge.

Chan, D.–Fourcin, A. (1995): A Spoken Language Resource for the EU. In *Proc. Eurospeech*, vol. 1. Madrid, 867–870.

Chomsky, N. (1965): *Aspects of the Theory of Syntax*. MIT Press.

Chu, M.–Zhao, Y.–Chang, E. (2006): Modeling stylized invariance and local variability of prosody in text-to-speech synhtesis. 48. évf. 6. sz., *Speech Communication*, 716–726.

Chu, W. C. (2003): *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. Wiley.

Cohen, M. M.–Massaro, D. W. (1993): Modeling coarticulation in synthetic visual speech. In Thalmann, N. M.–Thalmann, D. (szerk.) *Models and Techniques in Computer Animation*. Tokyo, Springer.

Cole, R.–Carmell, T.–Connors, P.–Macon, M.–Wouters, J.–Villiers, J.–Tarachow, A.–Massaro, D.–Cohen, M.–Beskow, J.–Yang, J.–Meier, U.–Waibel, A.–Stone, P.–Fortier, G.–Davis, A.–Soland, C. (1998): Intelligent Animated Agents for Interactive Language Training. In *ESCA-STILL 98*. Marholmen, Sweden, 163–166.

Collier, R. (1990): Multi-lingual intonation synthesis: Principles and applications. In *Proc. ESCA Workshop on Speech Synthesis*. Autrans, France, 273–276.

Collier, R.–Terken, J. (1987): Intonation by rule in text-to-speech applications. In *Proc. Europspeech*. Edinburgh, 165–168.

Content, A.–Mousty, P.–Radeau, M. (1990): Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. 90. évf. *L'Année Psychologique*, 551–556.

Cosatto, E.–Grafat, H. P. (1998): *Photo-realistic Talking Head. Computer Animation*. Philadelphia, Pennsylvania, 103–110.

Cosi, P.–Fusaro, A.–Tisato, G. (2003): LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. In *Proc. Eurospeech*. Geneva, Switzerland, 2269–2272.

Coulmas, F. (1999): *The Blackwell encyclopedia of writing systems*. Wiley.

Cowie, R.–Douglas-Cowie, E. (2001): Emotion Recognition in Human-Computer Interaction. 18. évf. 1. sz., *IEEE Signal Processing*, 32–80.

Creutz, M.–Hirsimaki, T.–Kurimo, M.–Puurula, A.–Pylkkonen, J.–Siivola, V.–Varjokallio, M.–Arisoy, E.–Saraolar, M.–Stolcke, A. (2007): Morph-based speech recognition and modeling of out-of-vocabulary words across languages. 5. évf. 1. sz., *ACM Transactions on Speech and Language Processing*, 1–29.

Crystal, D. (2003): *A Dictionary of Linguistics and Phonetics*. Oxford, UK, Blackwell Publishing.

Csányi Y. (1990): *Hallás-beszéd nevelés*. Budapest, Tankönyvkiadó.

Csapó T. G. (2009): Változatos prozódia megvalósítása szövegfelolvasó rendszerekben. IX. évf. 3. sz., *Akusztikai Szemle*, 16–18.

Csapó, T. G.–Bárkányi, Zs.–Gráczi, T. E.–Bőhm, T.–Lulich, S. M. (2009): Relation of formants and subglottal resonances in Hungarian vowels. In *Proc. Interspeech*. Brighton, 484–487.

Csapó T. G.–Németh G.–Fék M. (2008): Szövegfelolvasó természetességének növelése. LXIII. évf. 5. sz., *Híradástechnika*, 21–30.

Csatári, F.–Bakcsi, Zs.–Vicsi, K. (1999): A Hungarian Child Database for Speech Processing Applications. In *Proc. Eurospeech*. Budapest, 2231–2234.

Csendes D.–Hatvani C.–Alexin Z.–Csirik J.–Gyimóthy T.–Prószéky G.–Váradi T. (2003): Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. *II. Magyar Számítógépes Nyelvészeti Konferencia*, 238–245.

Csűri B. (1919): Hanglejtés. 78. sz., *Magyar Nyelvőr*, 71–78.

Czap L. (2004): *Audiovizuális beszédfelismerés és beszédszintézis*. PhD-értekezés (Budapesti Műszaki és Gazdaságtudományi Egyetem).

Dalton, B.–Kaucic, R.–Blake, A. (1996): Automatic Speechreading Using Dinamic Contours. In *Proc. NATO ASI Conference on Speechreading by Man and Machine*. 373–382.

Damper, R.–Marchand, Y.–Adamson, M.–Gustafson, K. (1999): Evaluating the pronunciation component of text-to-speech systems for English: A performance comparison of different approaches. 13. évf. *Computer Speech and Language*, 155–176.

Damper, R.–Marchand, Y.–Marsters, J.–Bazin, A. (2005): Aligning text and phonemes for speech technology applications using an EM-like algorithm. 8. évf. 2. sz., *Journal of Sol-Gel Science and Technology*, 147–160.

Damper, R. I.–Soonklang, T. (2007): Subjective Evaluation of Techniques for Proper Name Pronunciation. 15. évf. 8. sz., *IEEE Transactions on Audio Speech, and Language Processing*, 2213–2221.

Davis, K.–Biddulph, R.–Balashek, S. (1952): Automatic Speech Recognition of Spoken Digits. 24. évf. 6. sz., *J. Acoust. Soc. Am.*, 637–642.

Dedina, M.–Nusbaum, H. (1991): PRONOUNCE: A program for pronunciation by analogy. 5. évf. 1. sz., *Computer Speech & Language*, 55–64.

Deme L. (1962): A hanglejtés. In Tompa J. (szerk.) *A mai magyar nyelv rendszere II.* Budapest, Akadémiai Kiadó, 503–522.

Dempster, A. P.–Laird, N. M.–Rubin, D. B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm. 39. évf. 1. sz., *Journal Royal Statistical Society*, 1–38.

Deng, L.–Cui, X.–Pruvenokó, R.–Huang, J.–Momen, S.–Chen, Y.–Alwan, A. (2006): A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing. In *Proc. ICASSP*. 369–372.

Digalakis, V.–Rtischev, D.–Neumeyer, L. (1995): Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures. 3. évf. 5. sz., *IEEE Transactions on Speech and Audio Processing*, 357–366.

Dilley, L.–Shattuck-Hufnagel, S. (1996): Glottalization of word-initial vowels as a function of prosodic structure. 24. évf. 4. sz., *Journal of Phonetics*, 423–444.

Dong, M.–Lua, K. T. (2000): An Example-based Approach for Prosody Generation in Chinese Speech Synthesis. In *Proc. ICSLP*. Beijing, 303–307.

Douglas-Cowie, E.–Campbell, N.–Cowie, R.–Roach, P. (2003): Emotional Speech: Towards a New Generation of Databases. 40. évf. *Speech Communication*, 33–60.

Dreyer, M.–Eisner, J. (2009): Graphical models over multiple strings. In *Proc. Empirical Methods in Natural Language Processing: Volume 1*. Association for Computational Linguistics, 101–110.

Dudley, H.–Tarnóczy, T. (1950): The speaking machine of Wolfgang von Kempelen. 22. évf. *J. Acoust. Soc. Am.*, 151–166.

Díaz, F. C.–Banga, E. R. (2006): A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems. 48. évf. 8. sz., *Speech Communication*, 941–956.

É. Kiss K.–Kiefer F.–Siptár P. (1998): *Új magyar nyelvtan*. Budapest, Osiris.

Ekman, P.–Friesen, W. (1978): *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press.

Elekfi L.–Wacha I. (2003): *Az értelmes beszéd hangzása*. Szemimpex Kiadó.

Elliott, L. (1962): Backward and Forward Masking of Probe Tones of Different Frequencies. 34. sz., *J. Acoust. Soc. Am.*, 1116–1117.

Ericsson NetQual Inc. (2000): Auryst Audio Quality Measurement Solution.

Erjavec, T. (2004): MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. LREC*. 1535–1538.

Esposito, A. (2009): The Perceptual and Cognitive Role of Visual and Auditory Channels in Conveying Emotional Information. 1. évf. 3. sz., *Cognitive Computation*, 268–278.

Fackrell, J.–Vereecken, H.–Buchmann, J.–Martens, J.–Coile, B. (2000): Prosodic variation with text type. In *Proc. ICSLP*, vol. 3. Beijing, China, 231–234.

Fant, G. (1960): *Acoustic theory of speech production*. Mouton De Gruyter.

Fant, G.–Kruckenberg, A.–Nord, L. (1990): Prosodic and segmental variations. In *Proc. Speaker Characterization in Speech Technology*. 106–120.

Farkas M. (1996): *A hallássérültek kiejtés- és beszédfejlesztésének elmélete és gyakorlata*. Budapest, Bárczy Gusztáv Gyógypedagógiai Főiskola.

Fegyó, T.–Mihajlik, P.–Szarvas, M.–Tatai, P.–Tatai, G. (2003): Voxenter – Intelligent Voice Enabled Call Center for Hungarian. In *Proc. Eurospeech*. 1234–1237.

Fék M.–Olaszy G.–Szabó J.–Németh G.–Gordos G. (2005): Érzelem kifejezése gépi beszéddel. In Gósy M. (szerk.) *Beszédkutatás'2005*. Budapest, MTA Nyelvtudományi Intézet, 134–144.

Fék, M.–Pesti, P.–Németh, G.–Zainkó, Cs.–Olaszy, G. (2006): Corpus-Based Unit Selection TTS for Hungarian. In *Text, Speech and Dialogue*. Springer, 367–373.

Fék M.–Németh G.–Olaszy G.–Gordos G. (2004): Megértést segítő részletező gépi névfelolvasás magyar nyelvre. In Alexin Z.–Csendes D. (szerk.) *II. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, 301–306.

Fék M.–Zainkó Cs.–Németh G. (2007): Érzelmes beszéd gépi előállítása érzelem specifikus beszédadatbázisok felhasználásával. In Alexin Z.–Csendes D. (szerk.) *V. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, 34–43.

Fekete L. (1992): *Magyar kiejtési szótár*. Gondolat Könyvkiadó.

Feldhoffer, G.–Oroszi, B.–Takács, Gy.–Tihanyi, A.–Bárdi, T. (2007): Synchronization of acoustic speech data for machine learning based audio to visual conversion. In *Proc. Acoustics*. Madrid.

Ferenczy, T.–Németh, G.–Olaszy, G. (1997): A flexible Client-Server Model for Multilingual CTS/TTS Development. In *Proc. Eurospeech*, vol. 5. Rhodes, Greece, 693–697.

Field, J. (2003): *Psycholinguistics. A resource book for students.* Routledge.

Fischer, S. R. (1999): *A History of Language*. Reaktion Books.

Fischer, S. R. (2004): *A History of Reading*. Reaktion Books.

Fisher, R. (1922): On the mathematical foundations of theoretical statistics. 222. sz., *Philosophical Transactions of the Royal Society of London*, 309–368.

Fletcher, H. (1940): Auditory Patterns. 12. évf. 1. sz., *Revs. Modern Physics*, 47–65.

Fletcher, H.–Munson, W. A. (1933): Loudness, its definition, measurement and calculation. 5. évf. *J. Acoust. Soc. Am.*, 82–108.

Fónagy, I.–Magdics, K. (1960): The speed of utterance in phrases of different lengths. 3. sz., *Language and Speech*, 179–192.

Fónagy I. (1958): A hangsúlyról. In *Nyelvtudományi Értekezések 18.* Budapest, Akadémiai Kiadó.

Fónagy I. (1959): *A költői nyelv hangtanából*. Budapest, Corvina.

Fónagy I.–Magdics K. (1967): *A magyar beszéd dallama*. Budapest, Akadémiai Kiadó.

Fourcin, A.–Dolmazon, J. (1991): Speech knowledge, standards and assessment. In *Proc. International Congress of Phonetic Sciences*. 430–433.

Freitas, D.–Teixeira, J.–Olaszy, G.–Németh, G. (1998): Multivox – Conversor Texto-Fala para o Portugues. In *III Encontro para o Processamento Computacional de Portugues Escrito e Falado (PROPOR '98)*. Porto Alegre, Brazília, 88–98.

Fujisaki, H. (1992): Modelling the process of Fundamental Contour Generation. In *Speech Perception, Production and Linguistic Structure*. Japan, IOS Press, 314–326.

Furui, S. (1996): An overview of speaker recognition technology. In Lee, C.–Soong, F. K.–Kuldip, K. P. (szerk.) *Automatic Speech and Speaker Recognition*. Kluwer.

Gallwitz, F.–Niemann, H.–Nöth, E.–Warnke, W. (2002): Integrated recognition of words and prosodic phrase boundaries. 36. évf. *Speech Communication*, 81–95.

Gardner-Bonneau, D.–Blanchard, H. (2008): *Human Factors and Interactive Voice Response Systems*. 2. kiad. Springer.

Gáspár Á. (1971): Chordoscop, mint hallási fogyatékosok hangbeszédtanításának segédeszköze. In *A Gyógypedagógiai Tanárképző Főiskola évkönyve IV.* Budapest.

Gersho, A.–Gray, R. M. (1995): *Vector Quantization and Signal Compression*. 4. kiad. Norwell, MA, Kluwer.

Gibbon, F.–Hardcastle, W. (1998): Deviant Articulation in a Cleft Plate Child following Late Repair of the Hard Palate: A Description and Remediation Procedure using Electropalatography. *Clinical Linguistics and Phonetics*, 93–110.

Glotin, H.–Vergyri, D.–Neti, C.–Potamianos, G.–Luettin, J. (2001): Weighting schemes for audio-visual fusion in speech recognition. In *Proc. ICASSP*. 173–176.

Gobl, C.–Ní Chasaide, A. (2003): The role of voice quality in communicating emotion, mood and attitude. 40. évf. *Speech Communication*, 189–212.

Gocsál Á. (2000): A beszéd időviszonyai különböző életkorú személyeknél. In Gósy M. (szerk.) *Beszédkutatás'2000*. Budapest, MTA Nyelvtudományi Intézet, 39–50.

Gombocz Z. (1909): A magyar beszédhangok időtartamáról. Nyelvtudomány II., Budapest.

Gonsalo, E. L.–Olaszy, G.–Németh, G. (1993): Improvements of the Spanish version of the Multivox text-to-speech system. In *Proc. Eurospeech*. Berlin, 869–872.

Gordon, M.–Ladefoged, P. (2001): Phonation types: a cross-linguistic overview. 29. évf. *Journal of Phonetics*, 383–406.

Gordos, G.–Sándor, L. T. (1985): A limited vocabulary speech synthesiser terminal. In *Proc. of the Finnish-Hungarian symposium on information technology*. Helsinki, 3–10.

Gordos G.–Békési S.–Podoletz Gy.–Takács Gy. (1983): Nyelvfüggetlen beszédszintézis. Tanulmány, BME Híradástechnikai Elektronika Intézet.

Gordos G.–Takács Gy. (1983): *Digitális beszédfeldolgozás*. Budapest, Műszaki Könyvkiadó.

Gósy M. (1991): The perception of tempo. In Gósy M. (szerk.) *Temporal factors in speech*. Budapest, MTA Nyelvtudományi Intézet, 63–106.

Gósy M. (1997): Semleges magánhangzók a magyar beszédben. 121. évf. *Magyar Nyelvőr*, 9–19.

Gósy M. (1998): A beszédtervezés és a beszédkivitelezés paradoxona. 122. évf. *Magyar Nyelvőr*, 3–15.

Gósy M. (1999): A beszédprodukció tudatos változtatása: a beszélő személy utánzása. In Gósy M. (szerk.) *Beszédkutatás'1999*. Budapest, MTA Nyelvtudományi Intézet, 53–68.

Gósy M. (2000a): A beszédritmus elemzésének egy lehetséges megközelítése. 124. sz., *Magyar Nyelvőr*, 273–287.

Gósy M. (2000b): A beszédszünetek kettős funkciója. In Gósy M. (szerk.) *Beszédkutatás'2000*. Budapest, MTA Nyelvtudományi Intézet, 1–14.
Gósy M. (2002): Szükséges és szükségtelen hangátmenetek. In Gósy M. (szerk.) *Beszédkutatás'2002*. Budapest, MTA Nyelvtudományi Intézet, 20–32.
Gósy M. (2003): Virtuális mondatok a spontán beszédben. In Gósy M. (szerk.) *Beszédkutatás'2003*. Budapest, MTA Nyelvtudományi Intézet.
Gósy M. (2004a): A spontán magyar beszéd megakadásainak hallás alapú gyűjteménye. In Gósy M. (szerk.) *Beszédkutatás'2004*. Budapest, MTA Nyelvtudományi Intézet.
Gósy M. (2004b): *Fonetika, a beszéd tudománya*. Osiris.
Gósy M. (2005): *Pszicholingvisztika*. Budapest, Osiris.
Gósy M. (2006): A semleges magánhangzó nyelvi funkciói. In Gósy M. (szerk.) *Beszédkutatás'2006*. Budapest, MTA Nyelvtudományi Intézet, 8–22.
Gósy M. (2008a): Magyar spontánbeszéd-adatbázis – BEA. In Gósy M. (szerk.) *Beszédkutatás'2008*. Budapest, MTA Nyelvtudományi Intézet, 194–207.
Gósy M. (2008b): R hangok: kiejtés, hangzás, funkció. 132. évf. *Magyar Nyelvőr*.
Gósy M.–Beke A. (2010): Magánhangzó-időtartamok a spontán beszédben. 134. évf. 2. sz., *Magyar Nyelvőr*, 140–165.
Gósy M.–Horváth V. (2007): Fonetikai elemzések a spontán beszédben: alapok, kihívások. In Gósy M. (szerk.) *Beszédkutatás'2007*. Budapest, MTA Nyelvtudományi Intézet, 7–18.
Gósy, M.–Horváth, V. (2010): Changes in articulation accompanying functional changes in word usage. *Journal of the International Phonetic Association*, 135–142.
Gósy M.–Olaszy G. (1983): A gépi beszéd megértése. (Az Univoice magyar nyelvű, azonos idejű, számítógépes szövegszintetizáló rendszer percepciós vizsgálata). 85. évf. *Nyelvtudományi Közlemények*, 83–105.
Gósy M.–Olaszy G. (1991): A Multivox írás-beszéd átalakító beszédminőségének vizsgálata. In Bolla K. (szerk.) *Magyar Fonetikai Füzetek 23.* Budapest, MTA Nyelvtudományi Intézet, 62–73.
Gósy M.–Olaszy G.–Hirschberg J. (1984): *Eljárás szintetikus hangsorok előállítására hallásvizsgálatokhoz*. Magyar szabadalom. Lajstromszám: 193211.
Gósy, M.–Terken, J. (1994): Question marking in Hungarian: Timing and height of pitch peaks. 22. sz., *Journal of Phonetics*, 269–281.
Gronnum, N. (1992): *The groundworks of Danish intonation*. Copenhagen, Museum Tusculanum Press.
Györfi L.–Győri S.–Vajda I. (2000): *Információ- és kódelmélet*. Typotex, Budapest.
Halácsy P.–Kornai A.–Németh L.–Rung A.–Szakadát I.–Trón V. (2003): A Szószablya projekt. In *Proc. 1st Hungarian Computational Linguistics Conference*. Szegedi Tudományegyetem.

Halácsy, P.–Kornai, A.–Oravecz, Cs. (2007): HunPos – an open source trigram tagger. In *Proc. ACL*. 209–212.

Hamill, T.–Price, L. (2008): *The Hearing Sciences*. Plural Publishing Inc.

Hamon, C.–Mouline, E.–Charpentier, F. (1989): A diphone synthesis system based on time-domain prosodic modifications of speech. In *Proc. ICASSP*. 238–241.

Hanzo, L.–Somerville, F.–Woodard, J. (2001): *Voice Compression and Communications: Principles and Applications for Fixed and Wireless Channels*. New York, Wiley.

Hardcastle, J.–Hewlet, N. (1999): *Coarticulation: theory, data, and technics*. Cambridge University Press.

Hardcastle, W. J.–Gibbon, F. E.–Jones, W. (1999): Visual Display of Tongue-palate Contact: Electorpalatography in the Assessment and Remediation of Speech Disorders. 26. évf. *British Journal of Disorders Communication*, 41–74.

Harrington, J. (1988): Acoustic cues for automatic recognition of English consonants. In Jack, M.–Laver, J. (szerk.) *Aspects of speech technology*. Edinburgh University Press, 65–143.

Harris, L. (1953): A study of Building Blocks in Speech. 25. évf. *J. Acoust. Soc. Am.*, 962–969.

Hartmann, W. M. (1978): The Effect of Amplitude Envelope on the Pitch of Sinewave Tones. 63. sz., *J. Acoust. Soc. Am.*, 1105–1113.

Hedelin, P.–Huber, D. (1990): Pitch period determination of aperiodic speech signals. In *Proc. ICASSP*. 361–364.

Hegedűs L. (1930): Magyar hanglejtésminták grafikus ábrázolása. V. sz., *Collegium Hungaricum füzetek*.

Hegedűs L. (1958): Újabb vizsgálatok a magyar affrikáták köréből. *Nyelvtudományi Közlemények*, 4–22.

Hennecke, M.–Stork, D.–Prasad, K. (1996): Visionary speech: Looking ahead to practical speechreading systems. 331–349.

Henton, C.–Bladon, A. (1987): Creak as a sociophonetic marker. In Hyman, L. M.–Li, C. N. (szerk.) *Language, speech and mind*. London, Routledge, 3–29.

Hermansky, H. (1990): Perceptual linear predictive (PLP) analysis of speech. 87. évf. 4. sz., *J. Acoust. Soc. Am.*, 1738–1752.

Hermansky, H.–Morgan, N. (1994): RASTA processing of speech. 2. évf. 4. sz., *IEEE Transactions on Speech and Audio Processing*, 578–589.

Hess, W. (1983): *Pitch Determination of Speech Signals*. Springer.

Hirai, T.–Tenpaku, S. (2004): Using 5 ms segments in concatenative speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis*. Citeseer.

Hirano, M. (1981): *Clinical examination of voice*. Springer.

Hirst, D.–diCristo, A. (1998): *Intonation systems. A Survey of Twenty Languages*. Cambridge University Press.

Hollien, H.–Wendahl, R. W. (1968): Perceptual study of vocal fry. 43. évf. 3. sz., *J. Acoust. Soc. Am.*, 506–509.

Homer, D.–Ries, R.–Vatkins, S. (1939): A synthetic speaker. 227. évf. *J. Franklin Institute*, 739–764.

Horváth V. (2005): A magánhangzók nazalizációjáról. In Gósy M. (szerk.) *Beszédkutatás'2005*. Budapest, MTA Nyelvtudományi Intézet, 51–62.

Horváth V. (2008): A Hegedűs-archívum (1942–1962) feldolgozásának alapelvei. In *Nyelv, területiség, társadalom. A 14. élőnyelvi konferencia előadásai*. Budapest, Magyar Nyelvtudományi Társaság, 270–276.

Horváth V.–Gráczi T. E. (2010): Magánhangzórealizációk spontán beszédben. In Gósy M. (szerk.) *Beszédkutatás'2010*. Budapest, MTA Nyelvtudományi Intézet, 5–17.

Houtsma, A.–Rossing, T.–Wagenaars, W. (1988): Auditory Demonstrations, Philips Compact Disc 1126-061 and text.

Hozian, V.–Kacic, Z. (2003): Context-Independent Multilingual Emotion Recognition from Speech Signals. 6. évf. *International Journal of Speech Technology*, 311–320.

Hozian, V.–Kacic, Z. (2006): A Rule-Based Emotion-Dependent Feature Extraction Method for Emotion Analysis from Speech. 119. évf. 5. sz., *J. Acoust. Soc. Am.*, 3109–31206.

Hunt, A. J.–Black, A. W. (1996): Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICSLP*. 373–376.

Hunyadi, L. (1995a): Acoustic cues to sentential stress in Hungarian and their measurement. In Hunyadi, L.–Gósy, M.–Olaszy, G. (szerk.) *Studies in Applied Linguistics*. Debrecen, 29–48.

Hunyadi L. (1995b): Mondathangsúly a magyarban. In Gósy M. (szerk.) *Beszédkutatás'1995*. Budapest, MTA Nyelvtudományi Intézet, 35–45.

IEC (1999): MPEG-4 (1999): ISO//IEC 14496 standard, `http://www.iec.ch`.

Iida, A.–Campbell, N.–Higuchi, F.–Yasumura, M. (2003): A corpus-based speech synthesis system with emotion. 40. évf. 1-2. sz., *Speech Communication*, 161–187.

Illényi A.–Csányi K. (2001): *Mérnöki pszichoakusztika*. Budapesti Műszaki és Gazdaságtudományi Egyetem. Egyetemi jegyzet.

Imai, S. (1983): Cepstral analysis synthesis on the mel frequency scale. In *Proc. ICASSP*. 93–96.

ISO 226 (2003): Acoustics – Normal equal-loudness-level contours.

ISO 532 (1975): Method for calculating loudness level.

ISO-8859-1 (1998): Character Encoding: http://www.ic.unicamp.br/ stolfi/EXPORT/www/ISO-8859-1-Encoding.html.

ITU-R BS. (1998): Draft New Recommendation, Method for Objective Measurements of Perceived Audio Quality, Revision 1 to Document 10/20-E.

ITU-T F.902 (1995): Recommendation F.902. Interactive Services Design Guidelines.

ITU-T P.800 (1996): Recommendation P.800, Methods for Subjective Determination of Transmission Quality.

ITU-T P.830 (1996): Recommendation P.830, Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs.

ITU-T P.861 (1996): Recommendation P.861, Objective Quality Measurement of Telephone Band (300-3400 Hz) Speech Codecs.

ITU-T P.862 (2000): Recommendation P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.

ITU-T Q.23 (1988): Recommendation Q.23. Technical Features of Push-Button Telepohne Sets, http://www.itu.int/rec/T-REC-Q.23-198811-I/en.

Iyengar, G.–Potamianos, C., G.and Neti–Faruquie, T.–Verma, A. (2001): Robust detection of visual ROI for automatic speechreading. In *Proc. IEEE Multimedia Signal Processing*. 79–84.

Jászó A. (1991): *A magyar nyelv könyve*. Trezor Kiadó.

Jelinek, F. (1976): Continuous Speech Recognition by Statistical Methods. In *Proc. IEEE ICASSP*, vol. 4. 532–556.

Jelinek, F.–Mercer, R. L. (1980): Interpolated estimation of Markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice*. Amsterdam, The Netherlands: North-Holland, May.

Jin, Q. (2007): *Robust Speaker Recognition*. PhD-dissertation (Carnegie Mellon University Pittsburgh).

John, L. (1981): *Language and linguistics: an introduction*. Cambridge University Press, Cambridge, UK.

Johnstone, B. M.–Boyle, A. J. F. (1967): Basilar Membrane Vibration Examined with Mössbauer Technique. 158. évf. *Science*, 389–393.

Kaenel, M. (1998): Objective End-to-end Speech Quality Measurements. 1. évf. 15. sz., *QSDG Magazine*, 16–20.

Kálmán L.–Nádasdy Á. (1994): A hangsúly. In Kiefer F. (szerk.) *Strukturális magyar nyelvtan 2. Fonológia*. Akadémiai Kiadó.

Karttunen, L. (1998): The proper treatment of optimality in computational phonology. In *Proc. International Workshop on Finite State Methods in Natural Language Processing*. Association for Computational Linguistics, 1–12.

Kassai I. (1998): *Fonetika*. Budapest, Nemzeti Tankönyvkiadó.

Katz, S. M. (1987): Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. 35. évf. 3. sz., *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 400–401.

Kempelen, W. (1791): *Mechanismus der Menschlichen Sprache*. Wien.

Kent, R. D.–Read, C. (1992): *The Acoustic Analysis of Speech*. San Diego, California, Singular Publishing Group.

Khanna, S. M.–Leonard, D. G. B. (1982): Basilar Membrane Tuning in the Cat Cochlea. 215. évf. *Science*, 305–308.

Kiang, N. Y. S.–Moxon, E. C. (1974): Tails of Tuning Curves of Auditory-nerve Fibers. 55. évf. *J. Acoust. Soc. Am.*, 6201–6206.

Kingsbury, P.–Strassel, S.–McLemore, C.–MacIntyre, R. (1997): CALLHOME American English Lexicon (PRONLEX). *Linguistic Data Consortium, Philadelphia*.

Király J. (1989): A PC-TALKER beszédszintetizátor és digitális hangrögzítő-visszajátszó rendszer. 6. évf. 12. sz., *Magyar Elektronika*.

Kiss, G.–Arató, A.–Lukács, J.–Surlyán, J.–Vaspöri, T. (1987): BraiLab, a Full Hungarian Text-to-Speech Microcomputer for the Blind. In *Proc. World Conference in Phonetics*. 116–131.

Kiss G.–Olaszy G. (1984): A Hungarovox magyar nyelvű, szótár nélküli, valós idejű párbeszédes beszédszintetizáló rendszer. 2. sz., *Információ Elektronika*, 98–111.

Kiss, Z.–Bárkányi, Zs. (2006): A Phonetically Based Approach to the Phonology of [v] in Hungarian. 53. évf. 2. sz., *Acta Linguistica Hungarica*, 175–206.

Klatt, D. (1976): Linguistic uses of segmental duration in English. *J. Acoust. Soc. Am.*, 1208–1221.

Kohavi, R. (1995): A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proc. International Joint Conference on Artificial Intelligence*. 1137–1143.

Kondoz, M. (2004): *Digital Speech: Coding for Low Bit Rate Communication Systems*. England, Wiley.

Kontra M. (1988): *Beszélt nyelvi tanulmányok*. Linguistica, Series A, 1 sorozat. MTA Nyelvtudományi Intézet.

Kostoulas, T.–Ganchev, T.–Fakotakis, N. (2007): Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data. In *Proc. COST Action 2102 International Conference*. 235–242.

Koutny, I. (2008): *Natural Language Processing for Hungarian Speech Synthesis*. Poznan, Poland, Wydawnictwo Naukowe UAM.

Koutny, I.–Olaszy, G.–Olaszi, P. (2000): Prosody prediction from text in Hungarian and its realisation in TTS conversion. 3–4. évf. *International Journal of Speech Technology*, 187–200.

Kovács M. (2000): Beszédhangok kontextusfüggő időviszonyai. In Gósy M. (szerk.) *Beszédkutatás'2000*. Budapest, MTA Nyelvtudományi Intézet, 15–24.

Kovács M. (2002): Az affrikáták időszerkezetéről. In Hunyadi L. (szerk.) *Kísérleti Fonetika Laboratóriumi Fonológia*. Debreceni Egyetem Kossuth Egyetemi Kiadója, 39–54.

Kovács-Vass E. (1974): *Logopédiai jegyzet*. Budapest, Tankönyvkiadó.

Kovács-Vass, E. (1983): *Experimental Phonetic Research of Oral Sigmatisms*. PhD-dissertation (Prague).

Laczkó M. (1993): A tempó és a szünet viszonya a hangos olvasásban. In Gósy M.–Siptár P. (szerk.) *Beszédkutatás'1993*. Budapest, MTA Nyelvtudományi Intézet, 185–194.

Lambert, E.–Chesnet, D. (2001): Novlex: une base de données lexicales pour les élèves de primaire. 101. évf. 2. sz., *L'Année Psychologique*, 277–288.

Lamel, L.–DeMori, R. (1995): Speech recognition of European languages. In *Proc. IEEE Automatic Speech Recognition Workshop*. 51–54.

Lavagetto, F.–Lavagetto, P. (1996): Speechreading by Humans and Machines. In Stork, D.–Hennecke, M. (szerk.) *Time Delay Neural Networks for Articulatory Estimation from Speech*. Berlin, Springer, 437–444.

Laver, J. (1980): *The phonetic description of voice quality*. Cambridge University Press.

Laver, J. (1994): *Principles of Phonetics*. Cambridge University Press.

Laziczius Gy. (1944): *Fonétika*. Budapest, Tankönyvkiadó.

Lee, M.–van Santen, J.–Möbius, B.–Olive, J. (1999): Formant tracking using segmental phonemic information. In *Proc. Eurospeech*, vol. 6. 2789–2792.

Lehiste, I. (1970): *Suprasegmentals*. Cambridge, M.I.T. Press.

Lehtonen, J. (1970): Aspects of quantity in standard Finnish. In *Studia Phonologica Jyväskyläensia VI.* Jyväskylä.

Levinson, S.–Rabiner, L.–Sondhi, M. (1983): An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. 62. évf. *Bell Sys. Tech. J.*, 1035–1074.

Lewis, M. P. (2009): *Ethnologue: Languages of the World*. 16. kiad. Dallas, SIL International. http://www.ethnologue.com/.

Liebermann, P.–Blumstein, S. (1988): *Speech Psychology, Speech Perception and Acoustic Phonetics*. Cambridge University Press.

Liljencrants, J. (1967): The OVE III speech synthesizer. 2. évf. 3. sz., *Quarterly Progress and Status Report*. Dept. for Speech, Music and Hearing, KTH Stockholm.

Lindblom, B. (1990): Explaining phonetic variation: A sketch of the H&H theory. 55. évf. *Speech Production and Speech Modelling*, 403–439.

Lüngen, H.–Ehlebracht, K.–Gibbon, D.–Simoes, A. (1998): Bielefelder Lexikon und Morphologie in Verbmobil Phase II. 233. évf. *Verbmobil Report*.

Lyon, R. (2000): *Designing for Product Sound Quality*. Marcel Dekker.

MacQueen, J. B. (1967): Some Methods for classification and Analysis of Multivariate Observations. In *Proc. Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley, University of California Press, 281–297.

Magdics K. (1965): A magyar beszédhangok akusztikai szerkezete. In *Nyelvtudományi Értekezések*. 49. köt. Akadémiai Kiadó.

Magdics K. (1966): A magyar beszédhangok időtartama. *Nyelvtudományi közlemények*, vol. 68, 125–139.

MaGee, S. (2005): How to Create Your Company's Identity. www.EdwardLowe.org.

Maia, R.–Tomoki, T.–Heiga, Z.–Yoshihiko, N.–Keiichi, T. (2007): A trainable excitation model for HMM-based speech synthesis. In *Proc. Interspeech*. 1909–1912.

Majerník, V.–Kaluzny, J. (1979): On the Auditory Uncertainty Relations. *Acustica*, 132–146.

Markel, J. E.–Gray, A. H. (1976): *Linear Prediction of Speech*. Springer.

Markó A. (2005): *A spontán beszéd néhány szupraszegmentális jellegzetessége*. PhD-értekezés (ELTE).

Markó A. (2006): A hümmögés mint beszédaktus. In Mártonfi A.–Papp K.–Slíz M. (szerk.) *101 írás Pusztai Ferenc tiszteletére*. Budapest, Argumentum Kiadó, 604–612.

Markó A. (2007): Kérdő funkciójú hanglejtésformák a spontán beszédben. In Gósy M. (szerk.) *Beszédkutatás'2007*. Budapest, MTA Nyelvtudományi Intézet, 59–74.

Markó A. (2009): Stigmatizált hanglejtésforma a spontán beszédben. In Gósy M. (szerk.) *Beszédkutatás'2009*. Budapest, MTA Nyelvtudományi Intézet, 88–106.

Massaro, D. W. (1998b): *Perceiving Talking Faces: From Speech Perception to a Behavioural Principle*. Cambridge, MA, MIT Press.

Massaro, D.–Stork, D. (1998a): Speech recognition and sensory integration. 86. évf. 3. sz., *American Scientist*, 236–244.

Masuko, T.–Keiichi, T.–Takao, K.–Satoshi, I. (1997): Voice characteristics conversion for HMM-based speech synthesis system. In *Proc. ICASSP*. 1611–1614.

Matthews, I.–Potamianos, G.–Neti, C.–Luettin, J. (2001): A comparison of model and transform-based visual features for audio-visual LVCSR. In *Proc. IEEE Multimedia Expo*. Tokyo, 825–858.

Mátyás J. (2003): Vizuális beszédszintézis. Diplomaterv, Miskolci Egyetem.

McDermott, E. (1997): *Discriminative Training for Speech Recognition*. PhD-dissertation (Waseda Japan).

McGrath, M.–Summerfield, Q. (1985): Intermodal Timing Relations and Audio-visual Speech Recogniton. 77. évf. 2. sz., *J. Acoust. Soc. Am.*, 678–685.

MEA (1982): MEA 8000 Applications. Laboratory report. Philips.

Menyhárt K. (1998): Nyelvi meghatározottság a beszédszünetek észlelésében. In Gósy M. (szerk.) *Beszédkutatás'1998*. Budapest, MTA Nyelvtudományi Intézet, 47–57.

Menyhárt K. (2006): Koartikulációs folyamatok két magánhangzó kapcsolatában. In Gósy M. (szerk.) *Beszédkutatás'2006*. Budapest, MTA Nyelvtudományi Intézet, 44–56.

MERL (2008): Evaluation of speech recognition engines. `http://www.merl.com/projects/Speechengineeval/`, November 18.

Merriam–Webster (2003): *Merriam-Webster's collegiate dictionary*. Merriam-Webster.

Mihajlik P. (2010): *Spontán magyar nyelvű beszéd gépi felismerése nyelvspecifikus szabályok nélkül*. PhD-értekezés (Budapesti Műszaki és Gazdaságtudományi Egyetem).

Mihajlik, P.–Révész, T.–Tatai, P. (2002): Phonetic Transcription in Automatic Speech Recognition. *Acta Linguistica Hungarica*, vol. 49, no. 3–4, 407–425.

Mihajlik P.–Fegyó T.–Tatai P. (2006): Új eljárás a gépi beszédfelismerés környezetfüggő beszédhangmodelljeinek kialakítására. In Gósy M. (szerk.) *Beszédkutatás'2006*. Budapest, MTA Nyelvtudományi Intézet, 218–230.

Mihalcea, R.–Nastase, V. (2002): Letter Level Learning for Language Independent Diacrtitics Restoration. In *Proc. Computational Linguistics*. 1–7.

Mitton, R.–Street, M. (1992): A description of a computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English. `http://ota.ahds.ac.uk/`.

Möbius, B. (1995): Components of a quantitative model of German Intonation. In *Proc. International Congress of Phonetic Sciences*. 108–115.

Mohri, M.–Pereira, F.–Riley, M. (2002): Weighted finite-state transducers in speech recognition. 16. évf. 1. sz., *Computer Speech and Language*, 69–88.

Mohri, M.–Pereira, F.–Riley, M.–Allauzen, C. (1997): AT&T FSM finite state machine library. *AT&T Labs-Research*.

Molnár J. (1986): *A magyar beszédhangok atlasza*. Budapest, Tankönyvkiadó.

MPRSZ (2000): PR meghatározások. http://www.mprsz.hu/szolgaltatas/szotar/, Magyar Public Relations Szövetség.

MTA (1985): *A magyar helyesírás szabályai*. 11. kiad. Akadémiai Kiadó. `http://mek.oszk.hu/01500/01547/index.phtml`. (A letöltés ideje: 2008. július 3.).

Nádasdy Á. (2006): Nyelv és írás. In Kiefer F. (szerk.) *Magyar nyelv*. Budapest, Akadémiai Kiadó, 907–931.

Nagy B. (2008): Huhyphn: magyar elválasztásiminta-gyűjtemény, `http://www.tipogral.hu/`.

Nakamura, M.–Sawada, H. (2006): Talking Robot and the Analysis of Autonomous Voice Acquisition. In *Proc. IROS*. 4684–4689.

Navas, E.–Hernáez, I.–Luengo, I. (2006): An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS. 14. évf. 4. sz., *IEEE Transaktions on Audio, Speech and Language Processing*, 1117–1127.

Németh, G. (1998): From Near-Nil to Everyday Life: Speech Technology Based Telecommunications Services in Hungary. In *Proc. Workshop on Interactive Voice Technology for Telecommunications Applications*. 191–196.

Németh, G.–Kiss, G.–Zainkó, Cs.–Olaszy, G.–Tóth, B. (2008): Speech Generation in Mobile Phones. In Gardner-Bonneau, D.–Blanchard, H. (szerk.) *Human Factors and Interactive Voice Response Systems*. Springer, 163–191.

Németh G.–Olaszy G.–Bartalis M.–Kiss G.–Zainkó Cs.–Mihajlik P. (2007a): Speech based Drug Information System for Aged and Visually Impaired Persons. In *Proc. Interspeech*. 2533–2536.

Németh, G.–Zainkó, Cs.–Bartalis, M.–Olaszy, G.–Kiss, G. (2009): Human Voice or Prompt Generation? Can they Co-exist in an Application? In *Proc. Interspeech*. Brighton UK, 620–623.

Németh, G.–Zainkó, Cs.–Fekete, L.–Olaszy, G.–Endrédi, G.–Olaszi, P.–Kiss, G.–Kiss, P. (2000): The design, implementation and operation of a Hungarian e-mail reader. 3–4. évf. *International Journal of Speech Technology*, 216–228.

Németh, G.–Zainkó, Cs.–Kiss, G.–Fék, M.–Gordos, G.–Olaszy, G. (2003): Language processing for name and address reading in Hungarian. In *Proc. International Conference on Natural Language Processing and Knowledge Engineering, 2003*. 238–243.

Németh, G.–Zainkó, Cs.–Kiss, G.–Olaszy, G.–Fekete, L.–Tóth, D. (2007b): Replacing a Human Agent by an Automatic Reverse Directory Service. In *Advances in Information Systems Development*. Springer, 321–328.

Németh, G.–Zainkó, Cs. (2002): Multilingual statistical text analysis, Zipf's law and Hungarian speech generation. 49. évf. 3–4. sz., *Acta Linguistica Hungarica*, 385–405.

Németh G. (2006): Az akusztikai arculat szerepe az infokommunikációs szolgáltatók megítélésében. 8. sz., *Híradástechnika*, 17–21.

Németh, G.–Fék, M.–Csapó, T. G. (2007c): Increasing Prosodic Variability of Text-To-Speech Synthesizers. In *Proc. Interspeech*. Antwerp, Belgium, 474–477.

Németh G.–Olaszy G.–Fék M. (2006): Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei. In Gósy M. (szerk.) *Beszédkutatás'2006*. Budapest, MTA Nyelvtudományi Intézet, 183–196.

New, B.–Pallier, C.–Ferrand, L.–Matos, R. (2001): Une base de données lexicales du français contemporain sur internet: LEXIQUE. 101. évf. *L'Année Psychologique*, 447–462.

Nikléczy, P.–Olaszy, G. (2004): Kempelen's speaking machine from 1791: possibilities and limitations. (Recovering a 200 year-old technology). 62. sz., *Grazer linguistische Studien*, 111–120.

Nikléczy P. (2001): A műszeres személyazonosítás lehetőségei rövid időtartamú beszédminták alapján. In Gósy M. (szerk.) *Beszédkutatás'2001*. Budapest, MTA Nyelvtudományi Intézet, 154–171.

Nikléczy P.–Olaszy G. (2002): Kempelen beszélőgépének rekonstruálása. In Gósy M. (szerk.) *Beszédkutatás'2002*. Budapest, MTA Nyelvtudományi Intézet, 5–17.

Nolan, F. (1983): *The Phonetic Bases of Speaker Recognition*. Cambridge University Press.

Noll, M. (1969): Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In *Proc. Symposium on Computer Processing Communications*. Polytechnic Institute of Brooklyn, 779–797.

Nouza, J. (1999): Computer-aided spoken-language training with enhanced visual and auditory feedback. In *Proc. Eurospeech*. 183–186.

Novák A.–M. Pintér T. (2006): Milyen a még jobb Humor. In Alexin Z.–Csendes D. (szerk.) *IV. Magyar Számítógépes Nyelvészeti Konferencia*. 60–69.

Odell, J. J. (1995): *The use of context in large vocabulary speech recognition*. PhD-dissertation (Cambridge University).

Olaszi, P. (2000): Number elements in Hungarian. *International Journal of Speech Technology*, 546–462.

Olaszi P. (2002): *Magyar nyelvű szöveg-beszéd átalakítás: nyelvi modellek, algoritmusok és megvalósításuk*. PhD-értekezés (Budapesti Műszaki és Gazdaságtudományi Egyetem).

Olaszy G. (1978): A zönge szerepe az egyéni hangszínezet kialakításában. In Bolla K. (szerk.) *Magyar Fonetikai Füzetek 4*. Budapest, MTA Nyelvtudományi Intézet, 137–147.

Olaszy, G. (1984): A phonetically based data and rule system for the real time text-to-speech synthesis of Hungarian. In *Proc. International Congress of Phonetic Sciences*. 243–246.

Olaszy G. (1985): A magyar beszéd leggyakoribb hangsorépítő elemeinek szerkezete és szintézise. In *Nyelvtudományi értekezések 121.* 121 sorozat. Budapest, Akadémiai Kiadó.

Olaszy G. (1989a): *Elektronikus beszédelőállítás, A magyar beszéd akusztikája és formánsszintézise*. Műszaki Kiadó.

Olaszy, G. (1989b): MULTIVOX-A flexible text-to-speech system for Hungarian, Finnish, German, Esperanto, Italian and other languages for IBM-PC. In *Proc. Europspeech*, vol. 2. Paris, 525–529.

Olaszy, G. (1991a): Adaptation of the Multivox text-to-speech system to Italian. In *Proc. Eurospeech*. Genova, 1247–1250.

Olaszy, G. (1991b): The inherent time structure of speech sounds. In Gósy, M. (szerk.) *Temporal Factors in Speech*. Budapest, MTA Nyelvtudományi Intézet, 153–174.

Olaszy G. (1994a): *PC-Robot, programozható beszéd és ének szintetizátor*. Budapest, Nikol Elektronika.

Olaszy, G. (1994b): Sound duration measurements in declarative sentences. 42. évf. *Acta Linguistica Hungarica*, 51–62.

Olaszy G. (1995a): A kérés, a figyelmeztetés és a kérdés prozódiája a kijelentő mondatok tükrében. In Gósy M. (szerk.) *Beszédkutatás'1995*. Budapest, MTA Nyelvtudományi Intézet, 46–61.

Olaszy G. (1995b): Számok kiejtésének fonetikai vizsgálata. In Gósy M. (szerk.) *Beszédkutatás'1995*. Budapest, MTA Nyelvtudományi Intézet, 72–78.

Olaszy G. (1996): Szabályrendszer prozódiai elemek gépi megvalósításához. In Gósy M. (szerk.) *Beszédkutatás'1996*. Budapest, MTA Nyelvtudományi Intézet, 72–78.

Olaszy G. (2001a): *A beszéd akusztikai-fonetikai elemzése és modellezése, különös tekintettel a korszerű beszédépítés követelményeire*. MTA doktori értekezés (Magyar Tudományos Akadémia).

Olaszy G. (2001b): Prozodémák fonetikai reprezentációja. In Gósy M. (szerk.) *Beszédkutatás'2001*. Budapest, MTA Nyelvtudományi Intézet, 28–45.

Olaszy, G. (2002): The most important prosody patterns of Hungarian. 49. évf. 3–4. sz., *Acta Linguistica Hungarica*, 277–306.

Olaszy G. (2003): Az artikuláció akusztikai vetülete- a hangsebészet elmélete és gyakorlata. In Hunyadi L. (szerk.) *Kísérleti fonetika, laboratóriumi fonológia*. Debrecen, Debreceni Egyetem Kossuth Egyetemi Kiadója, 241–254.

Olaszy G. (2005): Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella- és a reklámok felolvasásában. In Gósy M.–Markó A. (szerk.) *Beszédkutatás'2005*. Budapest, MTA Nyelvtudományi Intézet, 21–50.

Olaszy G. (2006a): A korpusz alapú beszédszintézis nyelvi, fonetikai kérdései. *Híradástechnika*, 43–50.

Olaszy G. (2006b): Hangidőtartamok és időszerkezeti elemek a magyar beszédben. In *Nyelvtudományi Értekezések*. Akadémiai Kiadó.

Olaszy G. (2007a): A koartikulációs néma fázis jelensége. In Gósy M. (szerk.) *Beszédkutatás'2007*. Budapest, MTA Nyelvtudományi Intézet, 47–58.

Olaszy G. (2007b): *Mássalhangzó-kapcsolódások a magyar beszédben*. Tinta Könyvkiadó.

Olaszy G. (2010): Az [i]+V és V+[i] hangkapcsolódások akusztikai elemzése a hiátustöltés magyarázásához. In Gósy M. (szerk.) *Beszédkutatás'2010*. Budapest, MTA Nyelvtudományi Intézet, 73–83.

Olaszy G.–Abari K. (2005): Adatbázisok és számítógépprogramok a magyar beszéd időszerkezeti vizsgálatához. *Alkalmazott Nyelvtudomány*, vol. V, no. 1–2, 41–62.

Olaszy G.–Bartalis M. (2008): Jelfeldolgozási algoritmusok kombinációja a gépi hanghatárjelölés javítására. In Gósy M. (szerk.) *Beszédkutatás'2008*. Budapest, 208–220.

Olaszy, G.–Gordos, G. (1987): Automatic text-to-speech system applied in a reading machine. In *Proc. Eurospeech*. Edinburgh, 25–29.

Olaszy, G.–Gordos, G.–Németh, G. (1992): The MULTIVOX multilingual text-to-speech converter. In Bailly, G.–Benoit, C.–Sawallis, T. (szerk.) *Talking machines: Theories, Models and Applications*. Elsevier, 385–411.

Olaszy G.–Haraszti Cs. (2007): Lakossági gyógyszerinformációs rendszer beszédmodulokkal. In *Magyar Tudomány*. 3. köt. Budapest, Magyar Tudományos Akadémia, 78–81.

Olaszy G.–Kiss G. (1982): *Eljárás és berendezés szintetizátor(ok) vezérlésére, szótár nélküli szintetizált beszéd a vezérléssel quasi azonos időben történő előállítására.* Magyar szabadalom. Lajstromszám: 185527.

Olaszy G.–Kiss G.–Németh G.–Olaszi P. (2000a): Profivox: a legkorszerűbb hazai beszédszintetizátor. In *Beszédkutatás'2000*. Budapest, MTA Nyelvtudományi Intézet, 167–179.

Olaszy, G.–Koutny, I. (2001): Intonation of Hungarian questions and their prediction from text. In Puppel, S.–Grazina, D. (szerk.) *Prosody 2000*. Poznan, 179–196.

Olaszy, G.–Koutny, I.–Czap, L. (1988): Automatic synthesis of Esperanto. In *Proc. Acoustics*. Budapest, 132–137.

Olaszy G.–Németh G. (1996): *Eljárás emberi beszédből kivágott beszédelemek halmazának és összekapcsolásuk szabályainak meghatározására gép által magyar nyelven bemondandó tetszőleges számok, telefonszámok, pénzösszegek, számlaszámok, mérési eredmények számadatainak automatikus gépi összeállításához és felolvasásához*. Magyar szabadalom. Lajstromszám: P 9601427.

Olaszy, G.–Németh, G. (1993): Voxaid: an interactive speaking comunication aid software for the speech impaired. In *Proc. Eurospeech*. Berlin, 1821–1824.

Olaszy, G.–Németh, G. (1999): IVR for banking and residential telephone subscribers using stored messages combined with a new number-to-speech synthesis method. In Gardner-Bonneau, D. (szerk.) *Human Factors and Voice Interactive System*. Kluwer, 237–256.

Olaszy, G.–Németh, G.–Gordos, G.–Koutny, I. (1989): Approach to a multilingual text-to-speech synthesis. In *Proc. Joint Seminar on Speech Processing*. Vienna, 14–17.

Olaszy, G.–Németh, G.–Kiss, G. (2001): Hungarian audiovisual prosody composer and TTS development tool. In Puppel, S.–Grazina, D. (szerk.) *Prosody 2000*. Poznan, 167–178.

Olaszy, G.–Németh, G.–Olaszi, P.–Kiss, G.–Zainkó, Cs.–Gordos, G. (2000b): Profivox – a Hungarian TTS system for telecommunications applications. 3–4. évf. *International Journal of Speech Technology*, 201–215.

Olaszy G.–Olaszi P. (1998): Hangidőtartamok mesterséges változtatása periódusok kivágásával, megismétlésével. In Gósy M. (szerk.) *Beszédkutatás'1998*. Budapest, MTA Nyelvtudományi Intézet, 151–162.

Olaszy G.–Podoletz Gy.–Fisher J.–Poppe A. (1986): LPC-elven működő text-to-speech beszédszintetizáló rendszer fejlesztése. 21. évf. 5. sz., *Információ – Elektronika*, 247–255.

Olaszy G.–Rácz Z.–Bartalis M. (2009): Formánsmérések automatizálása, formánsadatbázisok létrehozása. In Gósy M. (szerk.) *Beszédkutatás'2009*. Budapest, MTA Nyelvtudományi Intézet, 134–147.

O'Shaughnessy, D. (1981): A study of French vowel and consonant durations. *Journal of Phonetics*, vol. 9, 385–406.

O'Shaughnessy, D. (1987): *Speech Communication: Human and Machine. Reading.* Massachusetts, Addison-Wesley Publishing Company. ISBN 0-201-16520-1.

Packard, J. (2000): *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

Paczolay D.–Kocsor A.–Sejtes G.–Hégely G. (2004): A Beszédmester csomag bemutatása: informatikai és nyelvi aspektusok. IV. évf. 1. sz., *Alkalmazott Nyelvtudomány*, 57–79.

Paczolay, D.–Kocsor, A.–Tóth, L. (2003): Real-Time Vocal Tract Length Normalisation in a Phonological Awareness Teaching System. In Matuosek, V.–Mautner, P.–Moucek, R.–Tauser, K. (szerk.) *Text, Speech and Dialogue*. Springer, 309–314.

Padmanadham, M.–Bahl, L. R.–Nahamoo, D.–Picheny, M. A. (1998): Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems. 6. évf. (1998. January) 1. sz., *IEEE Transactions on Speech and Audio Processing*, 701–704.

Pasdeloup, V. (1992): A prosodic model for French text-to-speech synthesis. In G. Bailly, C.–Benoit, T.–Sawallis (szerk.) *Talking Machines: Theories, Models and Designs*. Elsevier, 335–347.

Petajan, E. (1984): Automatic lipreading to enhance speech recognition. In *Proc. Global Telecommunications Conference*. Atlanta, GA, 265–272.

Peterson, G.–Wang, W.–Siversten, E. (1958): Segmentation technics in speech synthesis. 30. évf. *J. Acoust. Soc. Am.*, 739–742.

Pfister, B.–Romsdorfer, H. (2003): Mixed-lingual text analysis for polyglot TTS synthesis. In *Proc. Eurospeech*. Geneva, 2037–2040.

Pierrehumbert, J.–Talkin, D. (1992): Lenition of /h/ and glottal stop. In Docherty, D.–Ladd, D. R. (szerk.) *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge University Press, 90–117.

Plomp, R. (1976): *Aspects of Tone Sensation. London*. Academic Press.

Pollak, P.–Cernocky, J.–Boudy, J.–Choukri, K.–van den Heuvel, H.–Vicsi, K.–Virag, A.–Siemund, A.–Majewski, R.–Sadowksi, W.–Staroniewicz, J.–Tropf, P.–Kochanina, H.–Ostrouchov, A.–Rusko, M.–Trnka, M. (2000): SpeechDat(E) – Eastern European Telephone Speech Databases. In *Proc. LREC Satelite workshop XLDB – Very large Telephone Speech Databases*. Athens, 20–25.

Potamianos, G.–Neti, C. (2000): Stream confidence estimation for audio-visual speech recognition. III. évf. (2000. October), *Proc. ICSLP*, 746–749.

Potamianos, G.–Neti, C.–Luettin, J.–Matthews, I. (2004): Audio-Visual Automatic Speech Recognition: An Overview. In Bailly, G.–Vatikiotis-Bateson, E.–Perrier, P. (szerk.) *Issues in Visual and Audio-Visual Speech Processing*. MIT Press.

Povey, D.–Woodland, P. (2002): Minimum phone error and I-smoothing for improved discriminative training. In *Proc. IEEE ICASSP*, vol. 1. 105–108.

Přibilová, A.–Přibil, J. (2009): Spectrum modification for emotional speech synthesis. In Esposito, A.–Hussain, A.–Marinaro, M.–Martone, R. (szerk.) *Multimodal Signals: Cognitive and Algorithmic Issues*. Springer, 232–241.

Prószéky G. (1985): *Magyar szövegek számítógépes morfológiai elemzése. (A Nagyszótár számára rögzített folyamatos szövegek szövegszavainak tő- és toldalékmorfémákra való bontását megvalósító automata terve)*. Budapest, MTA Nyelvtudományi Intézet.

Prószéky, G.–Tihanyi, L. (1996): Humor: a morphological system for corpus analysis. In *Proc. of the first TELRI Seminar in Tihany*. 149–158.

Prószéky G.–Tihanyi L. (2009): Webforditas.hu: egy internetes nyelvtechnológiai szolgáltatás tanulságai. In Veronika V. (szerk.) *VI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, 19–23.

Rabiner, L. (1989): A tutorial on hidden Markov models and selected applications in speech recognition. In *Proc. IEEE*. 257–286.

Rabiner, L. R.–Schafer, R. W. (1978): *Digital Processing of Speech Signals*. Englewood Cliffs, NJ., Prentice-Hall.

Rabiner, L. (1968): Digital-Formant Synthesizer for Speech-Synthesis Studies. 43. évf. *J. Acoust. Soc. Am.*, 822–828.

Rabiner, L.–Juang, B. (1993): *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, Prentice-Hall.

Raux, A.–Black, A. (2003): A Unit Selection Approach to F_0 Modeling and its Application to Emphasis. In *Proc. ASRU*. 700–705.

Redi, L.–Shattuck-Hufnagel, S. (2001): Variation in the realization of glottalization in normal speakers. 29. évf. 4. sz., *Journal of Phonetics*, 407–429.

Rehor, K. (2007): World of VoiceXML, `http://www.kenrehor.com/voicexml/`.

Rhode, W. S.–Robles, L. (1974): Evidence from Mössbauer Experiments for Nonlinear Vibration in the Cochlea. 55. évf. *J. Acoust. Soc. Am*, 588–594.

Riley, M. D. (1992): Statistical tree based modeling of phonetic segment durations. In Bailly, G.–Benoit, C.–Swallis, T. (szerk.) *Talking machines: Theories, Models and Applications*. Elsevier, 265–274.

Roark, B.–Sproat, R. W. (2007): *Computational approaches to syntax and morphology*. Oxford University Press.

Robinson, T. (1997): BEEP – British English Example Pronunciations Dictionary.

Roche, E.–Schabes, Y. (1997): *Finite-state language processing*. The MIT Press.

Roederer, J. G. (1975): *Introduction to the Physics and Psychophysics of Music*. 2. kiad. New York, Springer.

Rossing, T.–Houtsma, J. (1986): Effects of signal envelope on the pitch of short sinusoidal tones. 79. évf. *J. Acoust. Soc. Am.*, 1926–1933.

Rossing, T. D. (1990): *The Science of Sound*. Illinois, Addison-Wesley Publishing Company.

Rutten, P.–Fackrell, J. (2003): The application of interactive speech unit selection in TTS systems. In *Proc. Interspeech*. 235–238.

Saenger, P. H. (2000): *Space between words: the origins of silent reading*. Stanford University Press, USA.

Sagisaka, Y.–Kaiki, N.–Iwahashi, N.–Mimura, K. (1992): ATR v-Talk speech synthesis system. In *Proc. ICSLP*. Banff, Canada, 483–486.

Sakoe, H.–Chiba, S. (1978): DynamicProgramming Algorithm Optimalization for Spoken Word Recognition. 26. évf. 1. sz., *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 43–49.

Scharf, B. (1970): Critical Bands. In Tobias, J. (szerk.) *Foundations of Modern Auditory Theory*. New York, Academic Press.

Scherer, K. R. (2003): Vocal communication of emotion: A review of research paradigms. 40. évf. 1-2. sz., *Speech communication*, 227–256.

Schramm, H. (2006): *Modeling Spontaneous Speech Variability for LargeVocabulary Continuous Speech Recognition*. PhD-dissertation (RWTH Aachen University).

Schweitzer, A.–Braunschweiler, N.–Klankert, T.–Möbius, B.–Sauberlich, B. (2003): Restricted Unlimited Domain Synthesis. In *Proc. Eurospeech*. Geneve, 1321–1324.

Sebe, N.–Cohen, I.–Huang, T. (2005): Multimodal Emotion Recognition. In Chen, C.–Wang, P. (szerk.) *Handbook of Pattern Recognition and Computer Vision*. World Scientific.

Sejnowski, T.–Rosenberg, C. (1993): NetTalk Corpus. Johns Hopkins University Cognitive Science Center Baltimore.

Seppänen, T.–Väyrynen, E.–Tovanen, J. (2003): Prosody-based classification of emotions in spoken Finnish. In *Proc. Eurospeech*. Geneva, 717–720.

Shannon, C. E. (1948): A Mathematical Theory of Communication. 27. évf. *Bell System Technical Journal*, 379–423., 623–656.

Sharma, C.–Kunins, J. (2002): *VoiceXML: Strategies and Techniques for Effective Voice Application Development with VoiceXML 2.0*. Wiley.

Shen, L.–Satta, G.–Joshi, A. (2007): Guided learning for bidirectional sequence classification. In *Proc. Annual Meeting-Association for Computational Linguistics*. 760–772.

Shinoda, K.–Watanabe, T. (2000): MDL-based context-dependent subword modeling for speech recognition. 21. évf. 2. sz., *J. Acoust. Soc. Jpn. (E)*, 79–86.

Siemund, R.–Höge, H.–Kunzmann, S.–Marasek, K. (2000): SPEECON, Speech Data for Consumer Devices. In *Proc. LREC*. Athens, 883–886.

Silberztein, M. (1993): *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris, Masson.

Silver, R. (2001): *Art as language*. Philadelphia, PA, USA, Psychology Press.

Siptár, P. (1996): A Janus faced Hungarian consonant. In *The Even Yearbook*. vol. 2. ELTE, 83–96.

Siptár P. (2002a): Hiátus. In Hunyadi L. (szerk.) *Kísérleti fonetika, laboratóriumi fonológia*. Debrecen, Debreceni Egyetem Kossuth Egyetemi Kiadó, 85–99.

Siptár P. (2002b): Optimális hiátustöltés. In Gósy M. (szerk.) *Beszédkutatás'1995*. Budapest, MTA Nyelvtudományi Intézet, 70–83.

Siptár P. (2006a): A fonéma tündöklése és... . 102. évf. *Magyar Nyelv*, 408–420.

Siptár P. (2006b): Affrikáta vagy hangkapcsolat? In Mártonfi A.–Papp K.–Slíz M. (szerk.) *101 írás Pusztai Ferenc tiszteletére*. Budapest, Argumentum, 406–410.

Siptár P. (2006c): Hangtan. In Ferenc K. (szerk.) *Magyar nyelv*. Budapest, Akadémiai Kiadó.

Siptár P.–Szentgyörgyi Sz. (2004): A magyar H-féle hangok optimális elemzése. 101. évf. *Nyelvtudományi Közlemények*, 57–90.

Slifka, J. (2006): Some physiological correlates to regular and irregular phonation at the end of an utterance. 20. évf. *Journal of Voice*, 171–186.

Soonklang, T.–Damper, R.–Marchand, Y. (2008): Multilingual pronunciation by analogy. 14. évf. 04. sz., *Natural Language Engineering*, 527–546.

Sproat, R. W. (1997): *Multilingual Text-to-Speech Synthesis: the Bell Labs Approach*. Kluwer.

Sproat, R.–Black, A.–Chen, S.–Kumar, S.–Ostendorf, M.–Richards, C. (2001): Normalization of non-standard words. 15. évf. 3. sz., *Computer Speech & Language*, 287–333.

Stern, R.–Liu, F.–Ohshima, Y.–Sullivan, T.–Acero, A. (1992): Multiple Approaches to Robust Speech Recognition. In *Proc. DARPA Speech and Natural Language Workshop*. New York.

Stevens, K. N. (1972): Evidence from Articulatory-Acoustic Data. In David, E.–Denes, P. (szerk.) *Human Communication: A Unified View*. New York, McGraw-Hill, 51–58.

Stevens, K. N. (1992): Speech synthesis methods: Homage to Denis Klatt. In Bailly, G.–Benoit, C. (szerk.) *Talking Machines*. Elsevier.

Stevens, K. N. (1998): *Acoustic Phonetics*. Cambridge, The MIT Press.

Stevens, K. N. (2000): *Acoustic Phonetics*. MIT Press.

Sumby, W.–Pollack, I. (1954): Visual contribution to speech intelligibility in noise. 26. évf. 2. sz., *J. Acoust. Soc. Am.*, 212–215.

Summerfield, A. (1987): Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B.–Campbell, R. (szerk.) *Hearing by Eye: The*

Psychology of Lip-Reading. London, United Kingdom, Lawrence Erlbaum Associates, 3–51.

Szaszák Gy. (2008): *A szupraszegmentális jellemzők szerepe és felhasználása a gépi beszédfelismerésben*. PhD-értekezés (Budapesti Műszaki és Gazdaságtudományi Egyetem).

Székely É. (2009): *E-Mese: A Hungarian voice for the Bonn Open Synthesis System*. PhD-dissertation (Utrecht University).

Szende T. (1995): *A beszéd hangszerelése (Idő, hangmagasság, hangerő és határjelzés a közlésben)*. Budapest, MTA Nyelvtudományi Intézet.

Szende T. (1975): Magánhangzóközi affrikátáink természetéről. 71. évf. *Magyar Nyelv*, 432–438.

Szende T. (1976): *A beszédfolyamat alaptényezői*. Budapest, Akadémiai Kiadó.

Szőke V. (2003): *Távközlési vállalatok arculatának új dimenziója*. Diplomaterv (Budapesti Corvinus Egyetem). Budapest.

Tachibana, M.–Junichi, Y.–Masuko, T.–Takao, K. (2005): Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. E88-D. évf. 11. sz., *IEICE Transactions Inf. & Systems*, 2484–2491.

Takács Gy. (1989): A budapesti hétszámjegyes átállás tájékoztató bemondásai. 10. sz., *Posta és Távközlési Szaklap*, 5–6.

Tamm A.–Olaszy G. (2005): Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához. In Alexin Z.–Csendes D. (szerk.) *III. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, 383–393.

Tamura, M.–Masuko, T.–Keiichi, T.–Takao, K. (2001): Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *Proc. ICASSP*. 805–808.

Tarján B.–Mihajlik P. (2010): On Morph-based LVCSR Improvements. In *Proc. SLTU*. 10–16.

Tarnóczy T. (1984): *Hangnyomás, hangosság, zajosság*. Budapest, Akadémiai Kiadó.

Tarnóczy T. (1941): *A magyar magánhangzók akusztikai szerkezete*. Budapest, Kir. Magy. Pázmány Péter Tudományegyetem általános Nyelvészeti és Fonetikai Intézete.

Tarnóczy T. (1974): A magánhangzók vizsgálatának akusztikai problémái. 10. évf. *általános Nyelvészeti Tudományok*, 181–196.

Tarnóczy T. (1982): *Zenei akusztika*. Budapest, Zeneműkiadó.

Tasaki, I. (1954): Nerve Impulses in Individual Auditory Nerve Fibres of Guinea Pigs. 17. évf. *Journal of Neurophysiology*, 97–101.

Tátrai Sz. (2009): A megnyilatkozás fogalmának perspektivikus értelmezése felé. In Keszler B.–Tátrai Sz. (szerk.) *Diskurzus a grammatikában & grammatika a diskurzusban*. Budapest, Tinta Könyvkiadó.

Taylor, P. (1998): The TILT Intonation model. In *Proc. ICSLP*. Sydney, ISCA, 1234–1237.

Taylor, P. (2009): *Text-to-Speech Synthesis*. Cambridge University Press.

Teleki Cs.–Vicsi K. (2006): Többnyelvű európai híranyag-adatbázis gyűjtése és feldolgozási módszereinek kutatása multimédiás műsorok automatikus feldolgozásához. 61. évf. 8. sz., *Híradástechnika*, 3–10.

Terhardt, E. (1979): Calculating Virtual Pitch. 1. sz., *Hearing Research*, 155–182.

Terhardt, E.–Fastl, H. (1971): Zum Einfluss von Störtönen und Störgeräuschen auf die Tonhöhe von Sinustönen. 25. sz., *Acustica*, 53–61.

Thiede, T.–Treurniet, W.–Bitto, R.–Schmidmer, C.–Sporer, T.–Beerends, J.–Colomes, C.–Keyhl, M.–Stoll, G.–Brandenburg, K.–Feiten, B. (2000): PEAQ–the ITU standard for objective measurement of perceived audio quality. 48. évf. *Journal of the Audio Engineering Society*, 3–29.

Thorndike, E. L.–Lorge, I. (1944): The Teachers' Word Book of 30,000 Words. *Linguistic Data Consortium, Philadelphia*.

Titterington, D.–Smith, A.–Makov, U. (1985): *Statistical Analysis of Finite Mixture Distributions*. Wiley.

Titze, I. R. (1995): Definitions and nomenclature related to voice quality. In Fujimura, O.–Hirano, M. (szerk.) *Vocal Fold Physiology – Voice Quality Control*. San Diego, Singular, 335–342.

Tótfalusi I. (2006): *Kiejtési szótár – Idegen nevek, szavak, kifejezések és szólások helyes kiejtése*. Tinta Könyvkiadó, Budapest.

Tóth, B.–Németh, G. (2006): VoxAid 2006: Telephone Communication for Hearing and/or Vocally Impaired People. In *Proc. International Conference on Computer Helping People with Special Needs*. Linz, Austria, Springer, 651–658.

Tóth, B.–Németh, G. (2007): Speech Enabled GPS Based Navigation System for Blind People on Symbian Based Mobile devices in Hungarian. In *Proc. Regional Conference on Embedded and Ambient Systems*. Budapest, 69–74.

Tóth, B.–Németh, G.–Kiss, G. (2004): Mobile Devices Converted into a Speaking Communication Aid. In *Proc. International Conference on Computer Helping People with Special Needs*. Paris, France, Springer, 1016–1023.

Tóth, S. L.–Sztahó, D.–Vicsi, K. (2007): Speech Emotion Perception by Human and Machine. In *Proc. COST Action 2102 International Conference*. Patras, Greece, Springer, 213–224.

Tóthfalusi I. (2006): *Kiejtési szótár*. Tinta Kiadó.

Toutanova, K.–Klein, D.–Manning, C.–Singer, Y. (2003): Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. Computational Linguistics on Human Language Technology*. Association for Computational Linguistics, 173–180.

Trón, V.–Kornai, A.–Gyepesi, G.–Németh, L.–Halácsy, P.–Varga, D. (2005): Hunmorph: open source word analysis. In *Proc. Workshop on Software*. Association for Computational Linguistics, 77–85.

Tucker, R. (1992): Voice Activity Detection Using a Periodicity Measure. In *Proc. Int. Electrical Engineering*. 377–380.

Umeda, N. (1975): Vowel duration in American English. 58. évf. *J. Acoust. Soc. Am.*, 440–445.

Umeda, N. (1977): Consonant duration in American English. 61. évf. *J. Acoust. Soc. Am.*, 846–858.

van Bezooijen, R.–Pols, L. W. (1990): Evaluating text-to-speech systems: Some methodological aspects. 9. évf. 4. sz., *Speech Communication*, 263–270. http://dx.doi.org/10.1016/0167-6393(90)90002-Q.

van Eynde, F.–Gibbon, D. (2000): *Lexicon Development for speech and language processing*. Kluwer, 115–139.

van Santen, J. (1992): Contextual effects on Vowel Duration. 11. évf. *Speech Communication*, 513–546.

van Santen, J. (1993): Perceptual experiments for diagnostic testing of text-to-speech systems. 7. évf. *Computer Speech and Language*, 49–100.

van Santen, J.–Buchbaum, A. L. (1997): Methods for optimal text selection. In *Proc. Eurospeech*, vol. 2. 553–556.

van Santen, J.–Kain, A.–Klabbers, E.–Mishra, T. (2005): Synthesis of Prosody using Multilevel Unit Sequences. 46. évf. 3–4. sz., *Speech Communication*, 365–375.

van Santen, J.–Olive, J. (1990): The analysis of contextual effects on segmental duration. 4. évf. *Computer Speech and Language*, 359–390.

Vandecatseye, A.–Martens, J.–Neto, J.–Meinedo, H.–Garcia Mateo, C.–Dieguez, J.–Mihelic, F.–Zibert, J.–Nouza, J.–David, P.–Pleva, M.–Cizmar, A.–Papageorgiou, H.–Alexandris, C. (2004): The COST 278 pan-European. Broadcast News Database. In *Proc. LREC*. Lisbon, 873–876.

Váradi T. (2003): A Budapesti Szociolingvisztikai Interjú. In Kiefer F.–Siptár P. (szerk.) *A magyar nyelv kézikönyve*. Budapest, Akadémiai Kiadó, 339–359.

Váradi V. (2008): A virtuális mondatok műfaji meghatározottsága. In Gósy M. (szerk.) *Beszédkutatás'2008*. Budapest, MTA Nyelvtudományi Intézet, 134–147.

Varga, L. (2002): *Intonation and Stress: Evidence from Hungarian*. New York, Palgrave Macmillan.

Varga L. (1993): A magyar beszéddallamok fonológiai, szemantikai és szintaktikai vonatkozásai. In *Nyelvtudományi értekezések*. 135. köt. Akadémiai Kiadó.

Varga L. (1994): A hanglejtés. In Kiefer F. (szerk.) *Strukturális magyar nyelvtan 2. Fonológia*. Akadémiai Kiadó, 468–546.

Varga L. (2000): A magyar mellékhangsúly fonológiai státusáról. 124. évf. *Magyar Nyelvőr*, 91–168.

Váry Á. (2001): Egy multimédiás többnyelvű beszédfejlesztő rendszer használatának pedagógiai tapasztalatai. 9. évf. 1–2. sz., *NyelvInfo*, 34–37.

Vary, P.–Martin, R. (2006): *Digital Speech Transmission*. Wiley.

Veilleux, N. M.–Ostendorf, M. (1993): Prosody/parse scoring and its application in ATIS. In *Proc. ARPA Human Language Technology Workshop*. 335–400.

Velkei S.–Vicsi K. (2004): Beszédfelismerő modellépítési kísérletek akusztikai, fonetikai szinten kórházi leletező, beszédfelismerő kifejlesztése céljából. In *II. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, 307–314.

Vicsi, K. (1996): SAMPA computer readable phonetic alphabet, Hungarian, `http://www.phon.ucl.ac.uk/home/sampa/hungarian.htm`.

Vicsi, K. (2004): An Overview of Speechtraining Methods Based on Multi Modal Feedback. In Pierre, D.–Georg, M. (szerk.) *Dynamics of speech production and perception*. Amsterdam, IOS Press, 1–10.

Vicsi K. (2005): *Számítógép alapú kiejtésoktatás és beszédfejlesztés műszaki tudományos alaprendszerének elméleti és gyakorlati kidolgozása*. MTA doktori értekezés (Magyar Tudományos Akadémia).

Vicsi K.–Kocsor A.–Tóth S. L.–Szaszák G.–Teleki Cs.–Bánhalmi A.–Paczolay D. (2005): A magyar referencia adatbázis és alkalmazása orvosi diktáló rendszerek kifejlesztéséhez. In *Proc. III. Számítógépes Nyelvészeti Konferencia*. Szeged, 435–438.

Vicsi, K.–Roach, P.–Öster, A.–Kacic, Z.–Barczikay, P.–Sinka, I. (1999): SPECO – A Multimedia Multilingual Teaching and Training System for Speech handicapped Children. In *Proc. Eurospeech*. Budapest, Hungary, 859–862.

Vicsi, K.–Szaszák, Gy. (2008): Using Prosody for the Improvement of ASR.: Sentence Modality Recognition. In *Proc. Interspeech*. Brisbane, Ausztrália, 2877–2880.

Vicsi K.–Sztahó D. (2009): Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban. In Tanács A.–Szauter D.–Vincze V. (szerk.) *VI. Magyar Számítógépes Nyelvészeti Konferencia*. 217–225.

Vicsi K.–Tóth L.–Kocsor A.–Gordos G.–Csirik J. (2002): MTBA – Magyar nyelvű telefonbeszéd adatbázis. LVII. évf. 8. sz., *Híradástechnika*, 35–39.

Vicsi, K.–Vig, A. (1995): Text independent neural network/rule based hybrid, Continuous Speech Recognition System. In *Proc. Eurospeech*, vol. 3. 2201–2204.

Vicsi K.–Vig A. (1998a): Az első magyar nyelvű beszdéadatbázis. In Gósy M. (szerk.) *Beszédkutatás'1998*. Budapest, MTA Nyelvtudományi Intézet, 163–177.

Vicsi, K.–Víg, A. (1998b): LIAS: Language Independent Automatic Segmentation Technique Using Sampa Labeling of Phonemes. In *Proc. LREC*. Spanyolország, Granada, 1317–1323.

Vicsi, K.–Sztahó, D. (2010): Problems of the automatic emotion recognitions in spontaneous speech; an example for the recognition in a dispatcher center. EURASIP Journal on Advances in Signal Processing, Special Issue on Emotion and Mental State Recognition from Speech. Under issue.

VoiceXML (2004): Voice Extensible Markup Language (VoiceXML) Version 2.0, http://www.w3.org/TR/voicexml20.

VoiceXML (2007): VoiceXML Development Guide, Version 2.1, http://www.vxml.org.

Wallace, J. (1998): Applications of Speech Recognition in the Primary School Classroom. In *Proc. ESCA-STILL*. Marholmen, Sweden, 21–24.

Weide, R. (1998): The Carnegie Mellon Pronouncing Dictionary cmudict. 0.6. Carnegie Mellon University: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Weiss, S.–Stewart, R. W.–Davis, G. M. (2002): Noise and Digital Signal Processing. In Davis, G. M. (szerk.) *Noise reduction in ASR Applications*. Boca Raton, Florida, USA, CRC Press, 3–46.

Wells, J. (1997): SAMPA computer readable phonetic alphabet. In Gibbon, D.–Moore, R.–Winski, R. (szerk.) *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York, Mouton de Gruyter.

Whitworth, B.–Zaic, M. (2003): The WOSP Model: Balanced Information System Design and Evaluation. 12. évf. *Communications of the Association for Information Systems*, 258–282.

Winckel, F. (1967): *Sound and Sensation*. Dover Publ. Inc.

Yao, X.–Bhutada, P.–Georgila, K.–Sagae, K.–Artstein, R.–Traum, D. (2010): Practical Evaluation of Speech Recognizers for Virtual Human Dialogue Systems. In *Proc. LREC*. Valletta, Malta.

Yoshimura, T.–Keiichi, T.–Masuko, T.–Takao, K.–Tadashi, K. (1997): Speaker interpolation in HMM-based speech synthesis system. In *Proc. Eurospeech*. 2523–2526.

Young, S. J. (2006): *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

Zainkó, Cs. (2008): Magyar nyelvű, kötött témájú korpusz-alapú beszédszintézis és a kötetlenség felé vezető út vizsgálata. LXIII. évf. 5. sz., *Híradástechnika*, 12–17.

Zainkó Cs. (2009): A magyar nyelv betűstatisztikája beszédfeldolgozási szempontok figyelembevételével. In *VI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, 238–245.

Zainkó, Cs.–Csapó, T.–Németh, G. (2010): Special Speech Synthesis for Social Network Websites. In *Text, Speech and Dialogue*. Springer, 367–373.

Zainkó, Cs.–Fék, M.–Németh, G. (2008): Expressive Speech Synthesis Using Emotion-Specific Speech Inventories. In Esposito, A.–Bourbakis, N.–Avouris, N.–Hatzilygeroudis, I. (szerk.) *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Lecture Notes in Computer Science sorozat, vol. 5042. 225–234.

Zainkó Cs.–Németh G. (2002): Az automatikus SMS-felolvasás problémái. In Gósy M. (szerk.) *Beszédkutatás'2002*. Budapest, MTA Nyelvtudományi Intézet, 197–211.

Zainkó Cs.–Németh G.–Olaszy G.–Gordos G. (2000): *Eljárás adott nyelven ékezetes betűk használata nélkül készített szövegek ékezetes betűinek visszaállítására*. Magyar szabadalom. Lajstromszám: 226740 P 00 03443.

Zibert, J.–Mihelic, F.–Martens, J.–Neto, J.–Meinedo, H.–Docio, L.–Mateo, C.–David, P.–Nouza, J.–Pleva, M.–Cizmar, A.–Zgank, A.–Kacic, Z.–Teleki, Cs.–Vicsi, K. (2005): The COST 278 Broadcast News segmentation and speaker clustering evaluation–overview, methodology, systems, results.

Zwicker, E. (1982): *Psychoakustik*. Springer.

Zwicker, E.–Fastl, H. (1990): *Psychoacoustics: Facts and Models*. Berlin, Springer.

Zwicker, E.–Flottorp, G.–Stevens, S. (1957): Critical Bandwidth in Loudness Summation. 29. évf. *J. Acoust. Soc. Am.*, 548–557.

Zwicker, E.–Terhardt, E. (1980): Analytical expressions for band rate and critical bandwidth as a function of frequency. 68. évf. *J. Acoust. Soc. Am.*, 1523–1525.

Zwislocki, J. J. (1969): Temporal Summation of Loudness: An Analysis. 46. sz., *J. Acoust. Soc. Am.*, 431–441.

FÜGGELÉK

F. függelék
Hangkapcsolatok

Olaszy Gábor

F.1. CC hangkapcsolatok

A CC hangkapcsolatok gyakorisági sorrendje 2 millió szóból álló szövegkorpuszból számolva, a leggyakoribbtól a legritkábbakig. A hangokat a betűjelükkel jelöltük.

Sorszám	C	C	Előfordulás	Sorszám	C	C	Előfordulás	Sorszám	C	C	Előfordulás
1	n	t	592843	2	l	t	509683	3	sz	t	383109
4	r	t	357687	5	n	k	300520	6	n	d	298066
7	s	t	274251	8	m	b	243536	9	t	v	192907
10	t	k	171934	11	l	m	162533	12	t	h	157977
13	r	g	152558	14	k	t	152499	15	l	k	147048
16	l	n	144375	17	t	m	139042	18	r	m	138949
19	t	l	130809	20	s	m	128116	21	g	b	124581
22	s	k	123514	23	l	v	115498	24	t	sz	114115
25	k	sz	113373	26	r	d	111299	27	t	s	103506
28	r	k	102141	29	l	d	100387	30	t	n	97530
31	t	r	95108	32	d	b	94597	33	j	t	92052
34	n	c	90927	35	k	m	86286	36	k	h	82410
37	k	r	81593	38	t	f	74208	39	r	s	73966
40	k	f	69549	41	m	l	67573	42	l	h	66581
43	r	v	65390	44	r	n	64852	45	n	cs	63406
46	k	s	61909	47	gy	m	61558	48	k	v	60739
49	n	h	60361	50	s	v	59633	51	sz	k	59388
52	n	s	59370	53	l	s	58841	54	m	t	58426
55	k	n	58054	56	l	g	55858	57	s	h	55133
58	g	n	53427	59	s	f	53152	60	m	p	52309
61	k	l	51997	62	g	r	51466	63	s	n	51191
64	n	l	50748	65	r	c	50542	66	z	d	50521
67	z	n	50179	68	d	n	49627	69	l	f	48628
70	d	v	47641	71	ty	sz	47202	72	ty	k	46992
73	g	v	46672	74	s	r	46507	75	r	sz	46409
76	l	c	46335	77	s	l	46119	78	n	sz	45372
79	sz	h	44721	80	m	v	44405	81	p	r	43426

Sorszám	C	C	Előfordulás	Sorszám	C	C	Előfordulás	Sorszám	C	C	Előfordulás
82	r	b	42644	83	r	j	42015	84	ny	gy	41479
85	gy	v	39623	86	l	sz	39491	87	r	z	39188
88	r	h	38861	89	g	l	37735	90	l	b	37620
91	m	k	37152	92	s	sz	36976	93	g	m	35934
94	g	j	34842	95	r	l	34616	96	t	p	34579
97	s	p	33601	98	j	d	33294	99	ty	t	32936
100	d	r	32859	101	z	b	31918	102	m	sz	31653
103	l	r	30723	104	r	f	30702	105	p	t	30015
106	ty	h	29428	107	r	ny	29172	108	z	m	27964
109	h	t	27962	110	b	r	27944	111	b	l	26752
112	m	r	26686	113	m	h	26319	114	p	k	26182
115	k	p	26080	116	k	cs	26065	117	n	r	25878
118	z	v	25860	119	c	sz	25437	120	gy	n	25436
121	p	s	24889	122	ny	t	24642	123	gy	b	24071
124	l	cs	23871	125	s	ty	23654	126	ty	f	23230
127	d	m	22978	128	m	n	22531	129	m	f	22366
130	g	z	22294	131	ty	l	21939	132	p	l	21607
133	s	cs	21254	134	n	z	20848	135	sz	n	20597
136	gy	r	20570	137	sz	l	20450	138	gy	l	20291
139	zs	d	20184	140	t	j	20158	141	r	cs	19840
142	j	n	19691	143	ny	v	17472	144	cs	k	17382
145	j	r	17238	146	z	g	17163	147	p	j	17108
148	g	d	16847	149	ty	s	16458	150	p	sz	16327
151	r	ty	16307	152	k	j	16300	153	n	j	16157
154	ty	p	15947	155	f	r	15441	156	l	p	15295
157	j	sz	15195	158	z	l	15072	159	m	s	14973
160	zs	g	14957	161	c	h	14834	162	c	v	14398
163	j	b	14338	164	j	k	13964	165	j	l	13629
166	p	h	13594	167	d	g	13497	168	m	cs	13374
169	r	p	13320	170	m	j	13237	171	m	z	13162
172	s	j	13126	173	ny	k	12680	174	c	k	12671
175	p	cs	12662	176	sz	f	12513	177	r	gy	12446
178	b	z	12223	179	b	m	11799	180	ny	h	11668
181	t	ny	11438	182	ny	b	11283	183	j	z	11233
184	d	l	11172	185	sz	p	10860	186	sz	m	10859
187	ny	sz	10310	188	l	gy	9699	189	z	r	9376
190	p	f	9272	191	sz	r	9078	192	b	d	8999
193	sz	v	8792	194	zs	gy	8748	195	ny	ty	8630
196	j	h	8496	197	ny	l	8362	198	r	zs	8246
199	l	z	7975	200	j	s	7915	201	j	m	7894
202	p	v	7883	203	k	c	7795	204	s	ny	7714
205	p	n	7610	206	gy	d	7408	207	ny	s	7407
208	k	ny	7372	209	v	b	7167	210	b	j	7132
211	g	gy	6676	212	s	c	6625	213	gy	z	6614
214	gy	j	6611	215	g	ny	6543	216	j	f	6451
217	b	v	6413	218	j	v	6338	219	c	l	6207
220	cs	m	6000	221	ny	n	5971	222	ny	m	5930
223	j	g	5853	224	l	ny	5735	225	sz	cs	5459
226	ny	f	5459	227	ny	r	5320	228	v	j	5118
229	b	n	5093	230	f	t	4938	231	p	m	4903

Sorszám	C	C	Előfordulás	Sorszám	C	C	Előfordulás	Sorszám	C	C	Előfordulás
232	ny	z	4832	233	z	ny	4666	234	m	d	4604
235	gy	g	4580	236	dz	b	4451	237	f	l	4371
238	m	c	4354	239	f	j	4285	240	c	r	3772
241	m	g	3705	242	ny	p	3672	243	sz	ty	3648
244	c	t	3523	245	v	r	3514	246	cs	n	3484
247	sz	ny	3145	248	m	ny	3056	249	cs	t	3036
250	gy	ny	2951	251	z	gy	2894	252	v	n	2791
253	l	ty	2773	254	m	gy	2748	255	j	cs	2706
256	c	m	2487	257	z	j	2411	258	h	n	2383
259	j	p	2295	260	v	m	2230	261	ny	cs	2186
262	g	zs	2040	263	b	ny	2034	264	ny	d	1975
265	v	l	1948	266	p	c	1902	267	c	n	1897
268	cs	r	1847	269	d	j	1798	270	v	d	1656
271	cs	sz	1654	272	cs	h	1621	273	cs	v	1609
274	n	zs	1545	275	c	f	1535	276	j	c	1529
277	sz	j	1524	278	h	r	1514	279	h	l	1491
280	j	ty	1481	281	cs	f	1399	282	sz	c	1388
283	b	g	1289	284	ny	j	1210	285	l	zs	1101
286	cs	l	1084	287	zs	l	1083	288	h	s	1053
289	ny	g	1032	290	j	ny	1022	291	c	ny	1014
292	d	ny	996	293	c	p	983	294	m	zs	981
295	b	gy	974	296	gy	zs	918	297	c	s	839
298	j	gy	827	299	p	ny	798	300	c	j	794
301	cs	p	785	302	b	zs	701	303	cs	j	699
304	ny	c	610	305	f	k	605	306	f	n	603
307	dz	d	589	308	h	m	588	309	j	zs	534
310	dzs	g	422	311	zs	v	411	312	f	s	397
313	v	g	386	314	ty	r	373	315	zs	m	372
316	h	k	356	317	f	m	331	318	dzs	d	306
319	v	ny	299	320	f	h	296	321	ny	zs	285
322	h	j	265	323	f	v	264	324	n	g	261
325	zs	n	253	326	dz	gy	253	327	c	cs	252
328	v	gy	239	329	f	p	224	330	h	sz	222
331	ty	m	196	332	cs	ny	188	333	zs	r	174
334	f	sz	171	335	cs	c	170	336	h	f	157
337	ty	v	156	338	v	z	150	339	h	p	134
340	k	ty	103	341	h	cs	103	342	ty	n	86
343	p	ty	81	344	h	ny	76	345	f	ty	75
346	zs	ny	56	347	f	cs	54	348	h	c	50
349	v	zs	48	350	m	ty	43	351	f	c	38
352	zs	j	32	353	f	ny	31	354	h	ty	27
355	ty	ny	22	356	c	ty	8	357	zs	b	1
358	dz	g	1	359	dz	n	1	360	dz	v	1
361	dz	zs	1	362	dz	l	1	363	dzs	b	1
364	dzs	gy	1	365	dzs	m	1	366	dzs	n	1
367	dzs	ny	1	368	dzs	j	1	369	dzs	v	1
370	dzs	l	1	371	dzs	r	1	372	cs	ty	1
373	cs	s	1								

F.2. CCC hangkapcsolatok

A CCC hangkapcsolatok gyakorisági sorrendje 2 millió szóból álló szövegkorpuszból számolva, a leggyakoribbtól a legritkábbakig. A hangokat a betűjelükkel jelöltük.

sorszám	C	C	C	előford.	sorszám	C	C	C	előford.	sorszám	C	C	C	előford.
1	l	t	s	5583	2	s	t	r	3708	3	n	t	r	3389
4	n	k	r	3110	5	sz	t	r	2889	6	n	t	h	2191
7	n	k	n	2101	8	m	p	l	1949	9	r	t	h	1801
10	sz	t	h	1606	11	n	d	r	1520	12	r	c	h	1389
13	r	t	n	1362	14	r	t	r	1242	15	n	s	t	1221
16	n	g	j	1168	17	s	t	v	1149	18	n	k	t	1087
19	r	t	k	1082	20	r	t	f	1066	21	n	t	k	1010
22	l	t	h	983	23	n	t	v	968	24	n	k	h	897
25	j	t	h	888	26	n	c	h	886	27	n	g	l	884
28	r	s	t	880	29	sz	t	v	874	30	n	g	r	868
31	n	g	v	850	32	r	t	sz	826	33	m	p	r	758
34	n	k	c	753	35	r	t	m	753	36	r	t	v	750
37	k	t	r	743	38	f	t	v	732	39	n	sz	k	681
40	s	p	r	672	41	n	t	m	643	42	n	cs	n	637
43	l	t	r	630	44	l	d	r	629	45	n	t	l	625
46	r	p	r	607	47	r	s	k	596	48	l	t	n	592
49	n	g	b	575	50	m	b	r	558	51	n	g	z	539
52	n	d	b	528	53	n	s	k	483	54	n	c	t	471
55	n	k	j	467	56	s	t	m	463	57	m	b	j	459
58	r	c	k	459	59	l	p	r	454	60	r	t	l	435
61	n	c	sz	432	62	r	d	j	430	63	n	c	k	429
64	n	t	sz	427	65	n	g	n	419	66	n	t	n	411
67	n	c	v	405	68	r	s	f	389	69	s	t	h	382
70	l	d	b	376	71	l	m	j	374	72	n	d	v	358
73	r	t	p	353	74	n	s	p	351	75	l	d	m	343
76	l	d	v	341	77	r	s	p	337	78	m	b	l	327
79	n	c	r	316	80	l	c	v	314	81	j	t	v	309
82	r	s	h	306	83	n	z	b	302	84	l	d	g	300
85	sz	t	m	298	86	l	t	v	296	87	n	cs	t	286
88	n	d	l	285	89	l	t	k	283	90	ny	v	b	283
91	sz	t	k	275	92	s	t	k	273	93	r	d	r	272
94	sz	p	r	270	95	sz	k	r	262	96	n	c	m	254
97	r	d	v	251	98	t	p	r	251	99	s	t	f	246
100	n	k	sz	243	101	r	p	h	242	102	l	m	sz	239
103	s	t	l	237	104	s	t	n	237	105	n	g	m	234
106	l	m	k	232	107	n	t	p	230	108	n	t	f	225
109	l	m	t	223	110	sz	t	n	222	111	l	c	sz	221
112	l	k	l	213	113	r	s	m	211	114	l	cs	t	210
115	l	s	t	210	116	n	sz	f	209	117	r	d	n	207
118	r	s	l	206	119	n	k	v	204	120	r	s	r	201
121	n	k	f	198	122	r	d	b	197	123	r	s	v	197
124	n	c	f	196	125	r	s	j	193	126	sz	t	l	192
127	l	m	f	190	128	n	c	n	189	129	l	cs	f	185
130	n	sz	t	185	131	l	m	r	181	132	n	s	n	181

sorszám	C	C	C	előford.	sorszám	C	C	C	előford.	sorszám	C	C	C	előford.
133	j	t	k	178	134	r	k	j	176	135	m	p	h	175
136	r	c	m	173	137	r	m	k	168	138	s	p	l	166
139	n	c	l	165	140	r	c	r	163	141	s	k	r	162
142	m	p	t	160	143	l	m	v	158	144	ny	p	r	156
145	l	m	b	155	146	n	sz	p	152	147	j	z	r	150
148	r	g	b	150	149	r	b	l	146	150	r	m	j	144
151	r	m	t	144	152	r	k	l	143	153	k	p	r	142
154	r	k	r	141	155	n	cs	k	137	156	l	cs	k	133
157	r	k	t	132	158	s	t	p	130	159	c	k	h	129
160	n	k	s	127	161	n	z	r	127	162	r	sz	t	127
163	r	v	r	126	164	l	cs	sz	125	165	l	f	r	125
166	r	k	h	124	167	sz	k	v	124	168	l	t	sz	122
169	m	s	t	121	170	l	s	p	117	171	k	t	l	116
172	n	s	h	115	173	n	s	r	112	174	n	t	ny	112
175	r	v	b	111	176	r	d	m	109	177	l	v	b	108
178	l	b	r	106	179	j	t	sz	105	180	j	t	m	103
181	ny	v	r	103	182	l	t	l	102	183	n	d	m	99
184	n	g	d	93	185	l	m	gy	92	186	l	v	r	91
187	l	m	h	90	188	n	c	p	90	189	p	s	t	90
190	g	d	r	88	191	j	t	r	88	192	k	s	t	85
193	n	cs	r	85	194	s	k	l	85	195	r	n	t	84
196	r	g	n	83	197	r	zs	v	83	198	l	cs	p	82
199	l	g	r	81	200	l	m	s	81	201	l	t	f	81
202	n	z	j	81	203	r	p	l	79	204	r	c	n	77
205	l	t	m	76	206	sz	k	j	76	207	m	t	h	75
208	n	k	cs	75	209	n	s	l	72	210	n	k	p	71
211	r	m	b	70	212	k	s	p	69	213	l	c	m	69
214	r	d	l	69	215	ny	v	n	67	216	r	c	sz	67
217	t	k	r	67	218	l	gy	b	66	219	n	z	l	66
220	n	sz	c	65	221	n	s	f	64	222	m	k	r	62
223	l	t	p	61	224	r	d	ny	61	225	r	m	f	61
226	p	f	r	60	227	p	t	r	60	228	t	f	r	60
229	k	sz	t	59	230	l	c	r	59	231	m	p	s	59
232	r	gy	r	58	233	r	v	n	58	234	r	zs	r	58
235	z	d	r	58	236	n	k	m	57	237	r	m	sz	57
238	l	m	d	55	239	p	sz	t	55	240	r	f	r	55
241	m	f	l	54	242	n	s	v	54	243	k	t	s	53
244	ny	k	r	52	245	r	m	p	52	246	c	k	j	51
247	j	p	r	51	248	l	cs	v	51	249	l	c	t	50
250	ny	d	r	50	251	r	cs	m	50	252	m	d	r	49
253	r	gy	b	49	254	r	v	d	49	255	l	p	ny	48
256	r	m	r	48	257	l	m	p	47	258	r	ny	s	47
259	n	s	m	46	260	ny	s	t	46	261	j	t	n	44
262	l	s	r	44	263	m	t	r	44	264	k	t	m	43
265	k	sz	k	42	266	r	c	j	42	267	r	k	v	42
268	l	v	n	41	269	r	c	p	41	270	j	sz	k	40
271	l	p	sz	40	272	ny	t	r	40	273	n	c	ny	38
274	s	f	r	38	275	k	t	f	37	276	l	m	l	37
277	ny	v	l	37	278	r	v	g	37	279	r	v	m	37
280	t	p	l	37	281	d	v	r	36	282	r	k	f	36

sorszám	C	C	C	előford.	sorszám	C	C	C	előford.	sorszám	C	C	C	előford.
283	n	sz	n	35	284	k	p	sz	33	285	m	sz	t	33
286	m	p	sz	32	287	p	sz	l	32	288	r	ny	b	32
289	r	ny	r	32	290	r	v	l	32	291	j	k	v	31
292	j	z	b	31	293	l	cs	n	31	294	z	d	v	31
295	k	t	k	30	296	k	sz	n	29	297	n	sz	m	29
298	ny	p	l	29	299	r	m	cs	29	300	f	t	h	28
301	n	sz	v	28	302	r	zs	b	28	303	r	zs	g	28
304	k	f	r	27	305	r	z	b	27	306	m	s	p	26
307	n	d	g	26	308	sz	k	h	26	309	k	sz	l	25
310	k	t	n	25	311	m	p	k	25	312	r	m	v	25
313	f	t	n	24	314	f	t	r	24	315	l	f	p	24
316	n	cs	ny	24	317	t	k	l	24	318	m	f	r	23
319	m	g	r	23	320	j	l	r	22	321	k	p	l	22
322	k	sz	v	22	323	r	ny	h	22	324	l	gy	r	21
325	n	s	cs	21	326	n	s	j	21	327	cs	p	r	20
328	k	t	v	20	329	l	f	t	20	330	r	d	g	20
331	r	n	r	20	332	j	l	t	19	333	l	v	l	19
334	r	g	j	19	335	g	b	r	18	336	l	k	v	18
337	p	s	p	18	338	r	ny	n	18	339	s	p	sz	18
340	j	l	n	17	341	ny	f	r	17	342	d	v	b	16
343	j	sz	t	16	344	r	ty	r	16	345	k	sz	p	15
346	r	k	p	14	347	f	sz	t	13	348	ny	v	d	13
349	r	f	t	13	350	r	ny	p	13	351	sz	p	l	13
352	f	t	k	12	353	j	l	b	12	354	j	z	n	12
355	k	sz	r	12	356	m	sz	l	12	357	n	k	ny	12
358	p	sz	r	12	359	r	p	c	12	360	d	g	r	11
361	j	k	l	11	362	j	n	t	11	363	n	cs	p	11
364	p	k	r	11	365	k	sz	h	10	366	r	ty	k	10
367	r	p	sz	9	368	t	p	sz	8	369	d	v	n	7
370	t	k	v	7	371	l	f	k	6	372	p	sz	n	6
373	d	b	l	5	374	h	t	r	5	375	ny	g	r	5
376	p	sz	h	5	377	z	d	b	5	378	f	k	l	4
379	j	n	b	4	380	j	n	h	4	381	j	n	r	4
382	l	cs	c	4	383	l	f	h	4	384	ny	f	k	4
385	r	f	k	4	386	v	b	r	4	387	f	p	r	3
388	g	d	b	3	389	h	t	k	3	390	l	ty	k	3
391	m	f	t	3	392	r	f	h	3	393	r	ty	s	3
394	zs	b	l	3	395	zs	d	r	3	396	f	s	p	2
397	h	t	h	2	398	l	f	sz	2	399	l	z	r	2
400	m	v	m	2	401	ny	f	t	2	402	r	zs	d	2
403	t	f	h	2	404	b	z	b	1	405	dz	g	b	1
406	f	k	r	1	407	g	z	b	1	408	h	t	n	1
409	j	d	b	1	410	j	dz	b	1	411	j	sz	f	1
412	l	dzs	d	1	413	l	ty	p	1	414	m	f	p	1
415	n	dz	b	1	416	n	dz	d	1	417	n	dz	zs	1
418	n	dzs	b	1	419	n	dzs	d	1	420	n	dzs	g	1
421	n	dzs	gy	1	422	n	t	s	1	423	n	zs	b	1
424	n	zs	gy	1	425	ny	f	h	1	426	ny	f	p	1
427	ny	f	s	1	428	ny	f	sz	1	429	ny	ty	h	1
430	ny	ty	f	1	431	ny	ty	r	1	432	r	c	cs	1

sorszám	C	C	C	előford.	sorszám	C	C	C	előford.	sorszám	C	C	C	előford.
433	r	d	z	1	434	r	dz	b	1	435	r	dzs	b	1
436	r	dzs	d	1	437	r	f	p	1	438	r	ty	f	1
439	ty	t	r	1	440	v	d	b	1	441	z	d	g	1
442	z	g	b	1	443	zs	b	r	1	444	zs	g	r	1
445	zs	d	b	1										

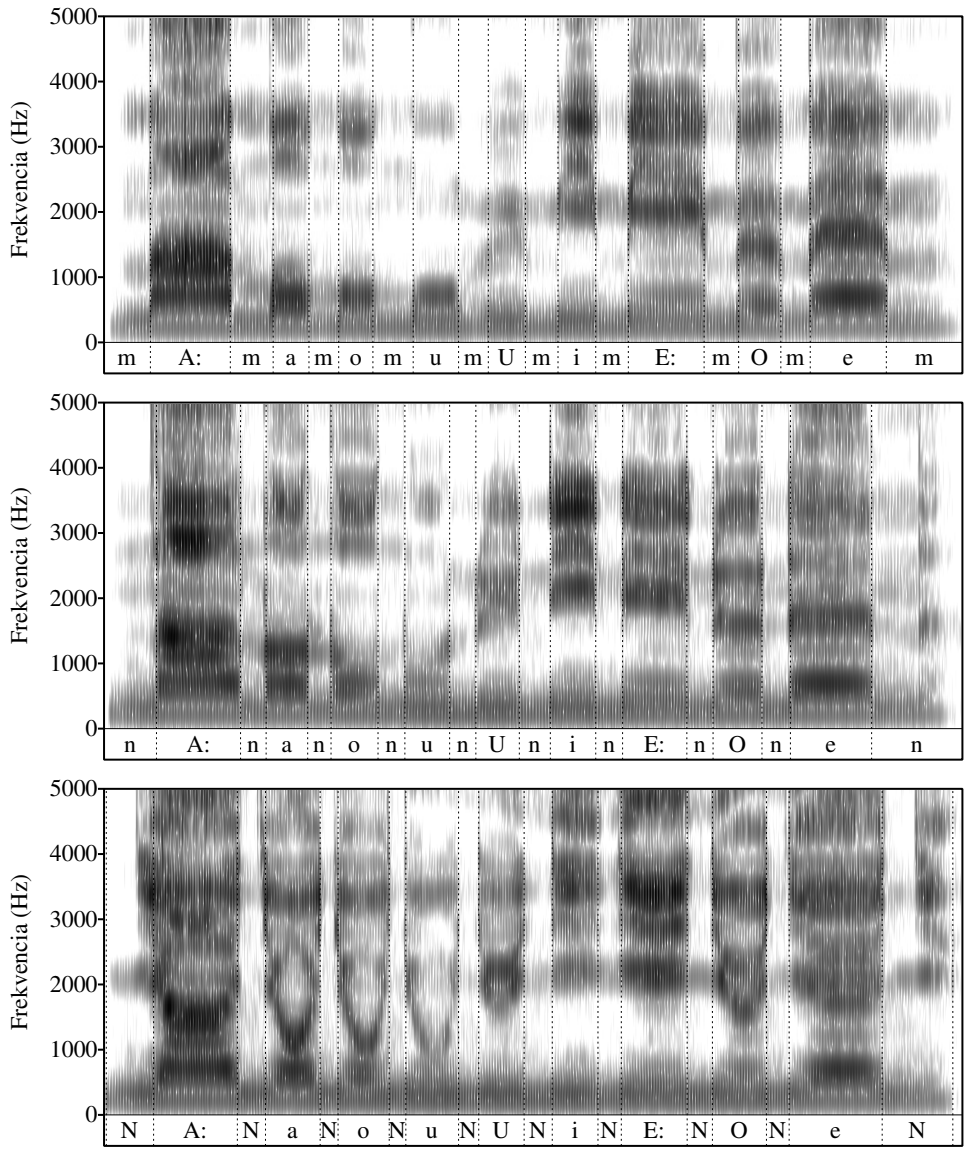
F.3. CCCC hangkapcsolatok

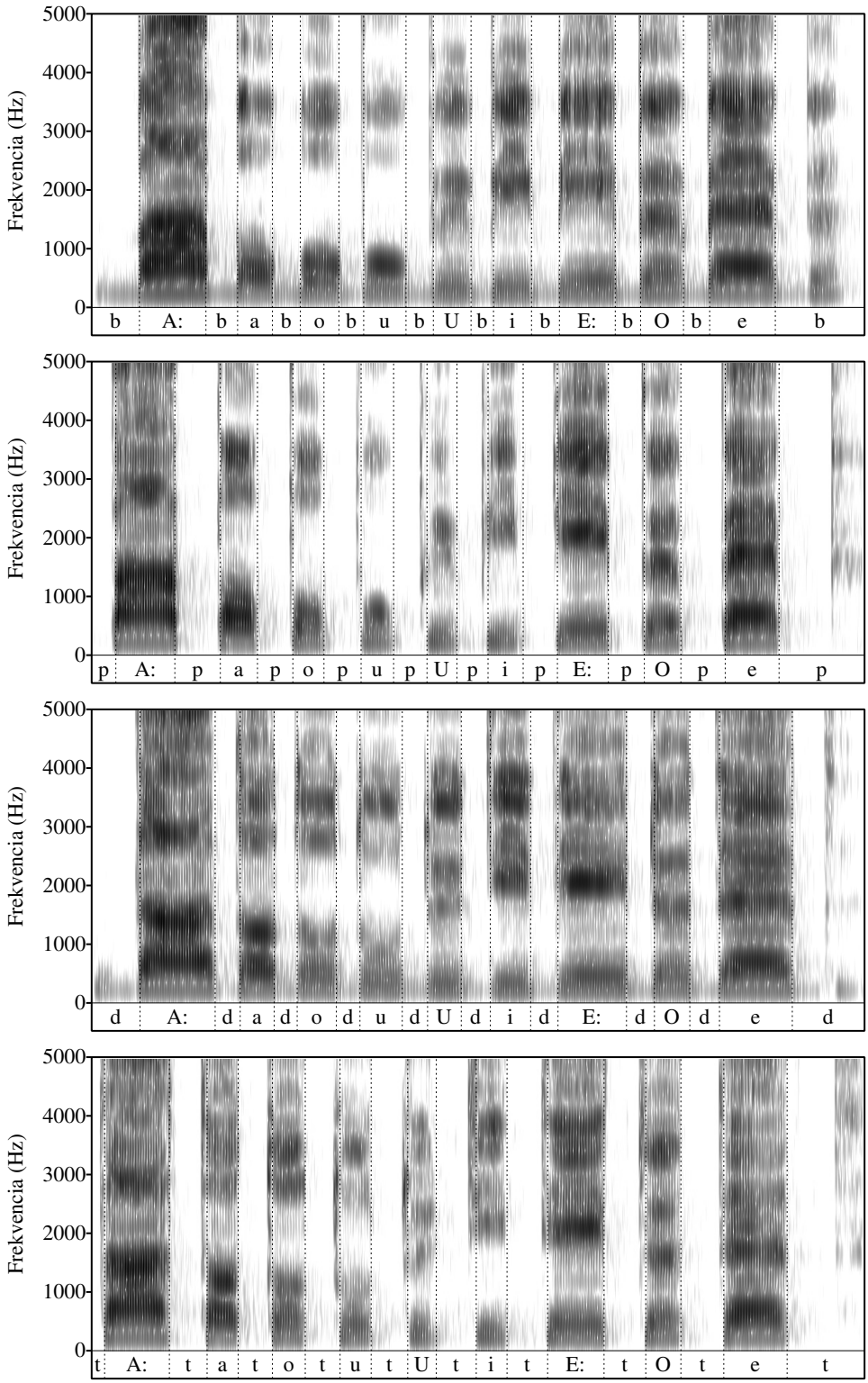
A CCCC hangkapcsolatok gyakorisági sorrendje 2 millió szóból álló szövegkorpuszból számolva, a leggyakoribbtól a legritkábbakig. A hangokat a betűjelükkel jelöltük.

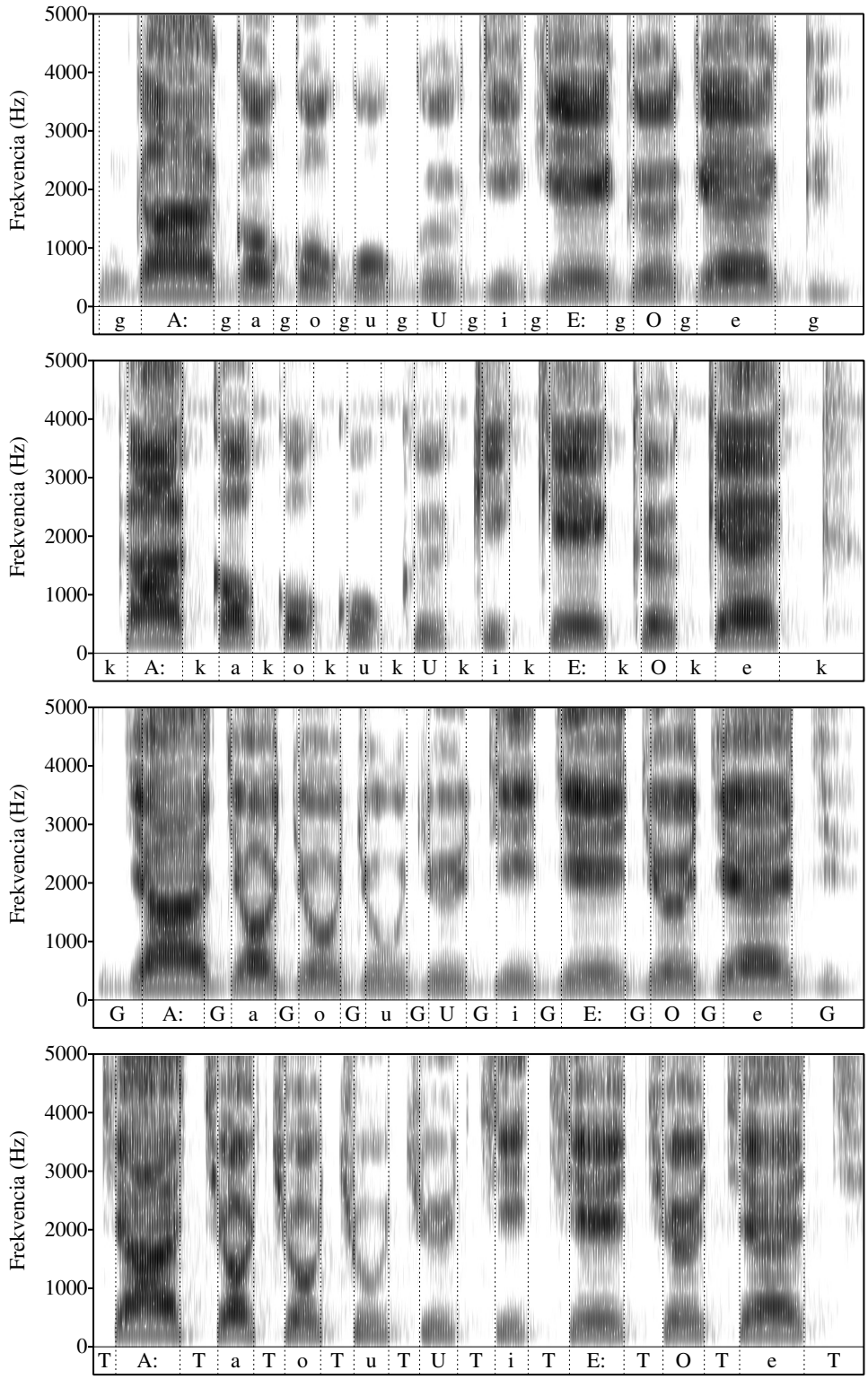
sorszám	C	C	C	C	előford.	sorszám	C	C	C	C	előford.	sorszám	C	C	C	C	előford.
1	n	s	t	r	1237	2	l	t	s	t	232	3	j	s	t	r	154
4	r	s	t	r	153	5	r	t	p	r	141	6	n	k	p	r	131
7	n	sz	p	l	61	8	ny	s	t	r	56	9	l	m	s	t	54
10	r	t	k	v	42	11	r	s	p	r	41	12	k	s	t	r	37
13	l	m	p	r	37	14	r	sz	t	v	36	15	m	s	t	r	35
16	n	c	p	r	34	17	n	t	s	p	33	18	r	sz	t	r	31
19	l	m	k	r	29	20	r	t	k	l	28	21	n	s	p	r	26
22	l	s	t	r	25	23	l	m	k	l	20	24	r	m	p	r	20
25	p	s	t	r	19	26	r	ny	p	r	19	27	r	c	k	r	16
28	j	l	t	r	15	29	l	m	d	r	13	30	l	m	t	r	12
31	n	t	p	r	11	32	l	cs	p	r	10	33	n	d	g	r	10
34	n	k	f	r	10	35	n	k	s	t	9	36	n	k	sz	t	9
37	r	t	p	l	9	38	l	m	s	l	8	39	n	c	k	l	8
40	r	sz	t	f	8	41	r	t	k	r	8	42	r	t	sz	t	7
43	l	t	k	r	6	44	l	t	p	r	6	45	n	s	k	r	6
46	r	c	p	r	6	47	k	s	p	r	5	48	l	s	p	r	5
49	r	t	p	sz	5	50	k	sz	t	r	4	51	n	c	t	r	3
52	n	cs	p	r	3	53	l	f	s	p	2	54	n	k	sz	r	2
55	ny	v	b	l	2	56	r	d	g	r	2	57	r	m	p	l	2
58	j	sz	p	r	1	59	k	t	s	p	1	60	l	cs	p	sz	1
61	l	dzs	d	r	1	62	l	f	p	sz	1	63	l	t	s	p	1
64	n	k	s	l	1	65	n	k	s	p	1	66	n	k	sz	n	1
67	ny	f	k	l	1	68	ny	f	k	r	1	69	ny	f	p	r	1
70	ny	f	s	t	1	71	p	sz	t	r	1	72	r	dzs	b	l	1
73	r	m	t	r	1	74	z	d	g	r	1						

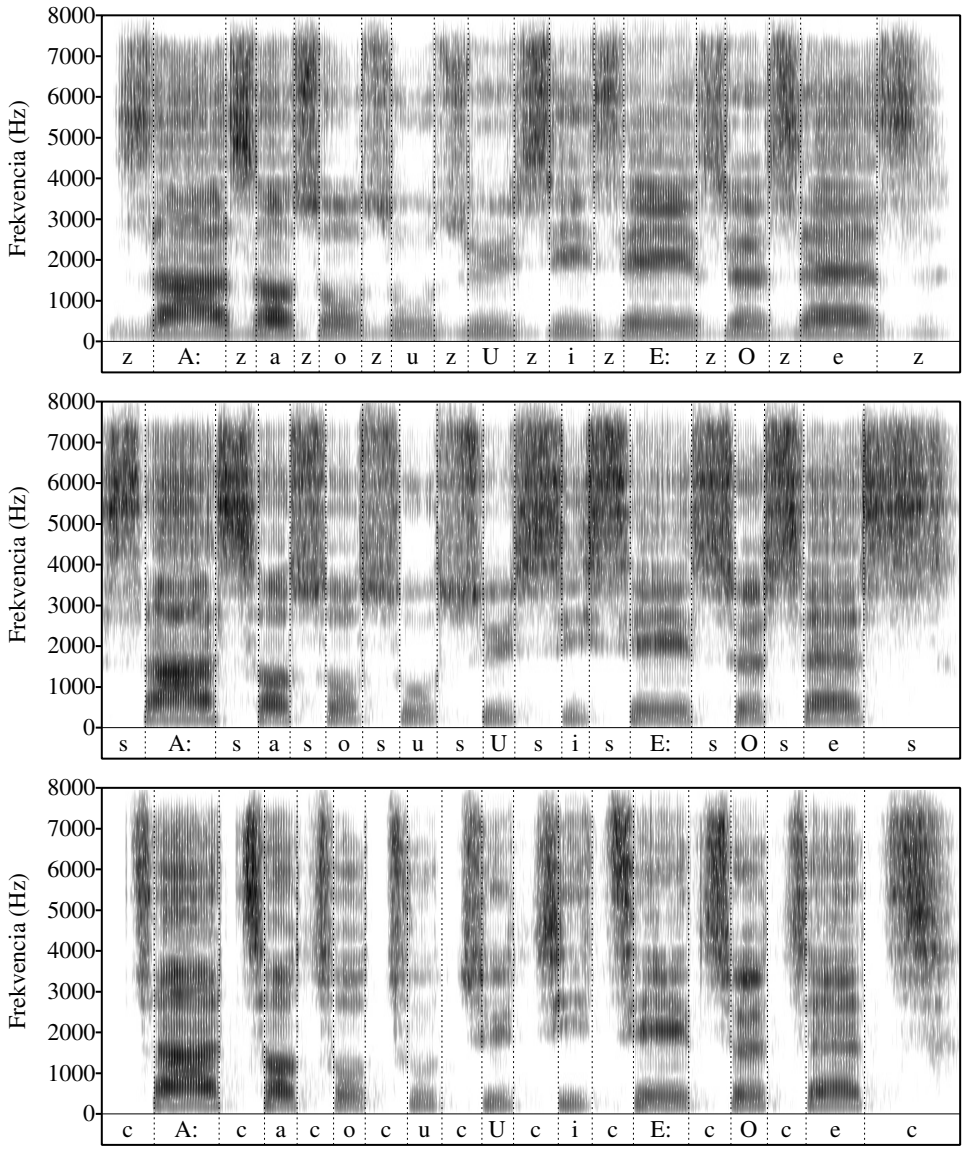
F.4. CVC hangkapcsolatok spektrogramjai

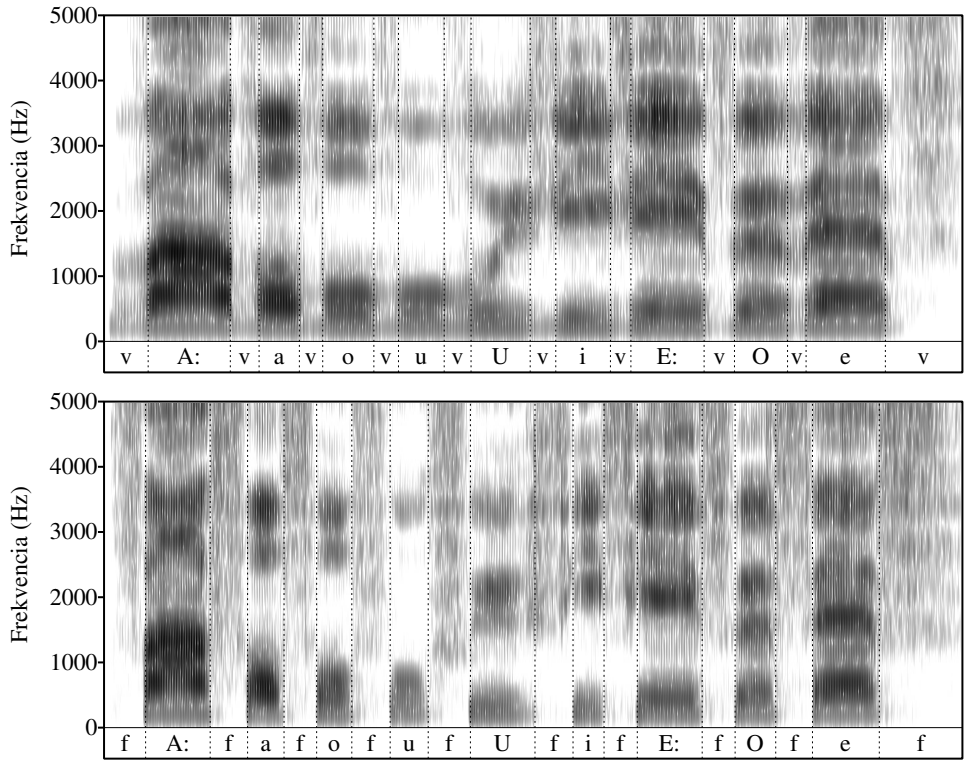
A magyar C1-V-C1 típusú hangkapcsolatok spektrális és időszerkezetét mutatjuk be a koartikuláció függvényében emberi ejtésből, férfi bemondó hangjából. Először a nazális hangokra utána pedig az orális hangokra.

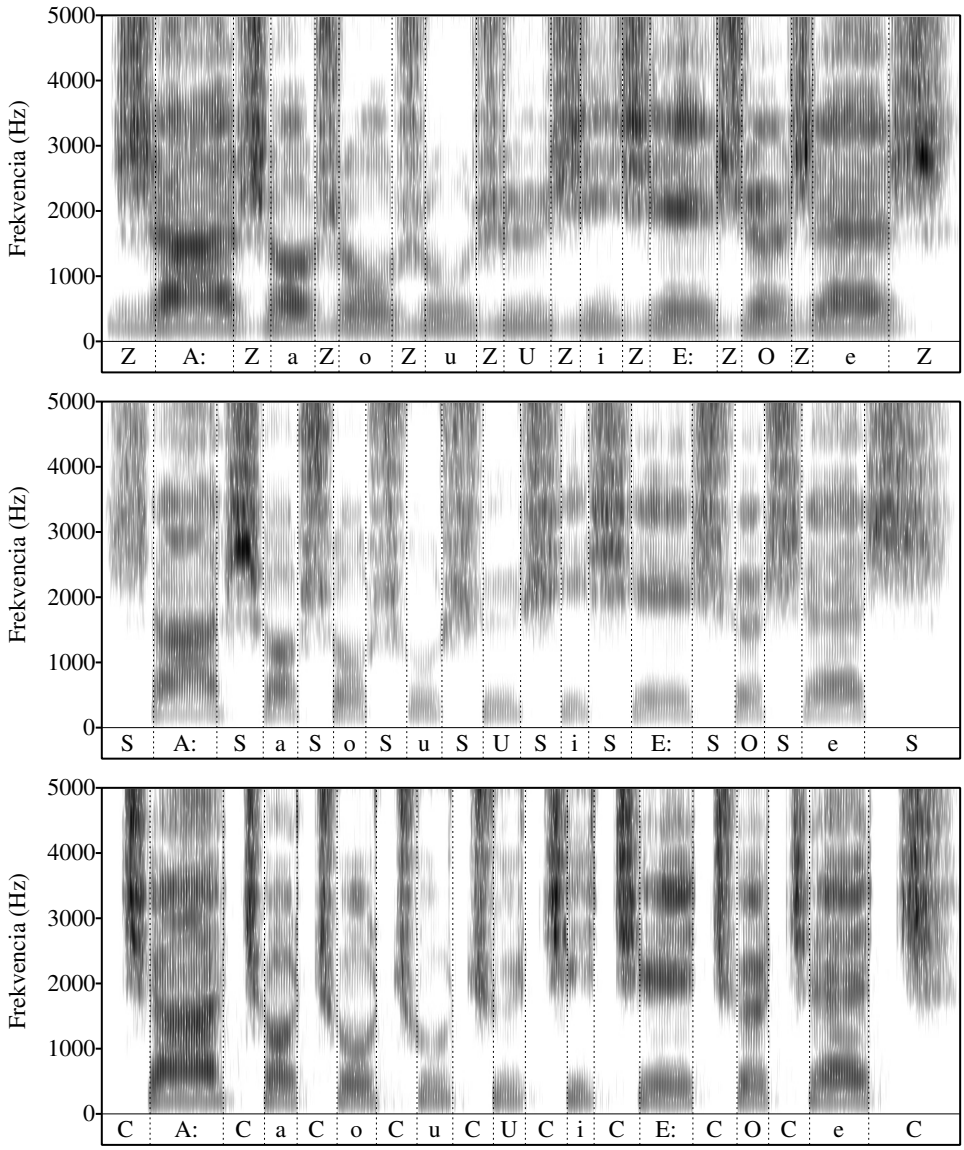


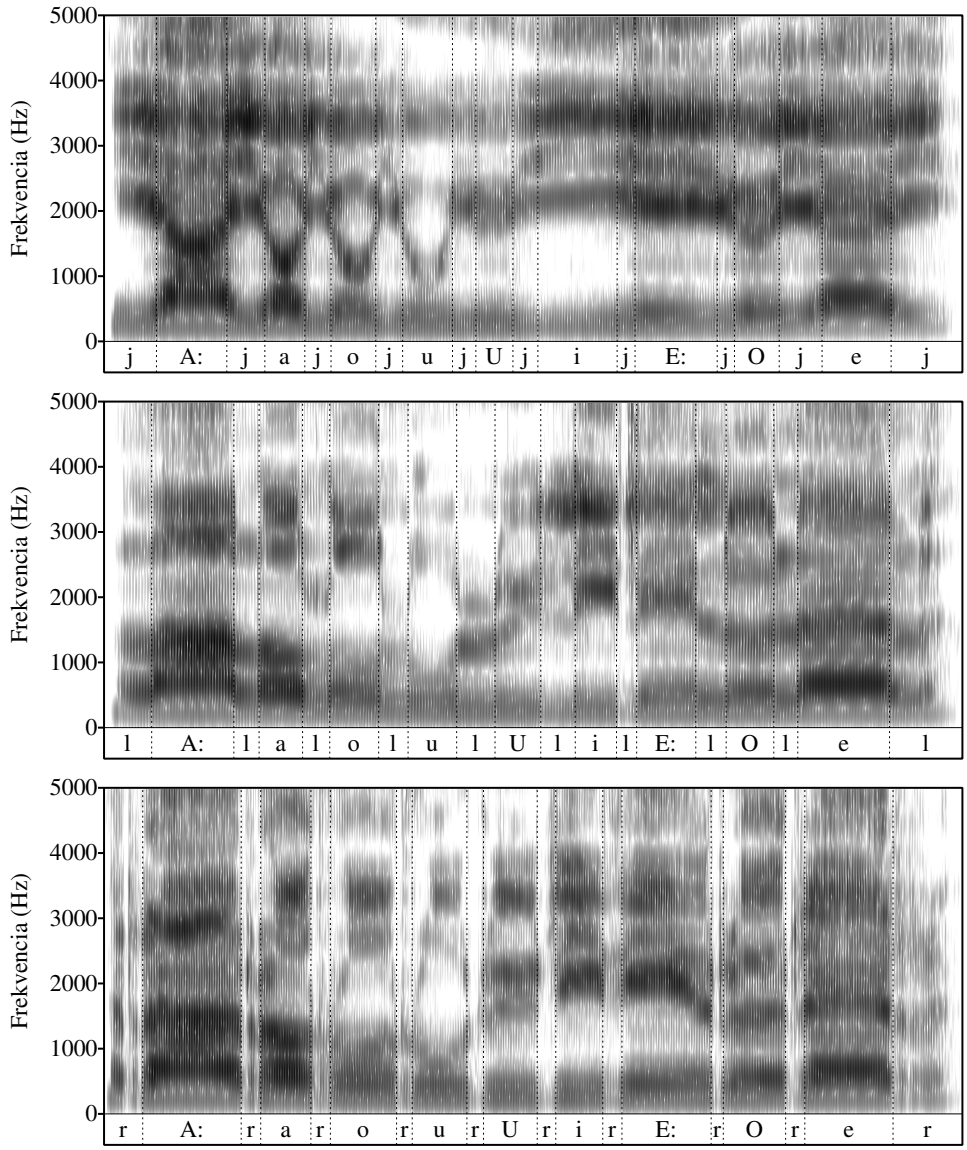

Frekvencia (Hz)
5000
4000
3000
2000
1000
0
b A: b a b o b u b U b i b E: b O b e b
p A: p a p o p u p U p i p E: p O p e p
d A: d a d o d u d U d i d E: d O d e d
t A: t a t o t u t U t i t E: t O t e t


Frekvencia (Hz)
5000
4000
3000
2000
1000
0
g A: g a g o g u g U g i g E: g O g e g
k A: k a k o k u k U k i k E: k O k e k
G A: G a G o G u G U G i G E: G O G e G
T A: T a T o T u T U T i T E: T O T e T


Frekvencia (Hz)
8000
7000
6000
5000
4000
3000
2000
1000
0
z A: z a z o z u z U z i z E: z O z e z
s A: s a s o s u s U s i s E: s O s e s
c A: c a c o c u c U c i c E: c O c e c


Frekvencia (Hz)
5000
4000
3000
2000
1000
0
v A: v a v o v u v U v i v E: v O v e v
Frekvencia (Hz)
5000
4000
3000
2000
1000
0
f A: f a f o f u f U f i f E: f O f e f


Frekvencia (Hz)
5000
4000
3000
2000
1000
0
Z A: Z a Z o Z u Z U Z i Z E: Z O Z e Z
Frekvencia (Hz)
5000
4000
3000
2000
1000
0
S A: S a S o S u S U S i S E: S O S e S
Frekvencia (Hz)
5000
4000
3000
2000
1000
0
C A: C a C o C u C U C i C E: C O C e C


Frekvencia (Hz)
5000
4000
3000
2000
1000
0
j A: j a j o j u j U j i j E: j O j e j
l A: l a l o l u l U l i l E: l O l e l
r A: r a r o r u r U r i r E: r O r e r

Tárgymutató

A kiadásért felelős
az Akadémiai Kiadó Zrt. igazgatója
Felelős szerkesztő: Tárnok Irén
Termékmenedzser: Egri Róbert
LaTeX: Zainkó Csaba, Fegyó Tibor
A kézirat verziója: 1090-1089:1090(mx)
A nyomdai munkálatokat az Akadémiai Nyomda végezte
Felelős vezető: Ujvárosi Lajos
Martonvásár, 2010
Kiadványszám: TK090031
Megjelent 46 (A/5) ív terjedelemben