Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Távközlési és Médiainformatikai Tanszék

A gépi beszéd-előállítás természetességének növelése rejtett Markov-modell alapú szövegfelolvasó rendszerben

Ph.D. disszertáció

BME-VIK Informatikai Tudományok Doktori Iskola

Csapó Tamás Gábor
okl. mérnök-informatikus

Témavezető:
Németh Géza, Ph.D.

Budapest, 2013

Nyilatkozat önálló munkáról, hivatkozások átvételéről

Alulírott Csapó Tamás Gábor kijelentem, hogy ezt a doktori értekezést magam készítettem és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Budapest, 2013. november 15.

_____

Csapó Tamás Gábor

Nyilatkozat nyilvánosságra hozatalról

Alulírott Csapó Tamás Gábor hozzájárulok a doktori értekezésem interneten történő nyilvánosságra hozatalához az alábbi formában*:

- korlátozás nélkül
- elérhetőség csak magyarországi címről
- elérhetőség a fokozat odaítélését követően 2 év múlva, korlátozás nélkül
- elérhetőség a fokozat odaítélését követően 2 év múlva, csak magyarországi címről

Budapest, 2013. november 15.

_____

Csapó Tamás Gábor

* a megfelelő választást kérjük aláhúzni

Kivonat

Csapó Tamás Gábor „A gépi beszéd-előállítás természetességének növelése rejtett Markov-modell alapú szövegfelolvasó rendszerben” című PhD értekezéséhez

A gépi szövegfelolvasás célja, hogy írott szöveget alakítsunk át emberihez hasonló beszéddé. A mai megoldásokban előtérbe kerültek a statisztikai parametrikus módszerek. Gyakran a rejtett Markov-modell alapú rendszert alkalmazzák erre a célra a beszéd forrás-szűrő modelljének használatával. Természetes beszédben a hangszalagok kváziperiodikus rezgése hosszabb-rövidebb időszakokra szabálytalanná (irregulárissá) válhat, azaz ingadozások jelenhetnek meg a periódusonkénti amplitúdóban és/vagy az alapfrekvenciában. Ez érdes, rekedtes hangot eredményezhet, amely a természetes beszéd szerves része. Kutatásaim kezdetéig a jelenség beszédszintézisre gyakorolt hatását nem vizsgálták részletesen. A mai beszédtechnológiai módszerek további gyengesége a forrás-szűrő modell használata során, hogy a feltételezés szerint a forrás és a szűrő tökéletesen szétválasztható. Ez azonban nem mindig teljesül, és nemlineáris csatolás jöhet létre a forrás és a szűrő közötti kölcsönhatás miatt. Az utóbbi néhány évben kimutatták, hogy az alsó légúti rendszer is hozzájárul a beszédhangok alakításához és vizsgálata segíti a kölcsönhatás megértését.

Disszertációmban a fenti területeken született eredményeimet ismertetem. Először egy új gerjesztési modellt mutatok be, amely a beszéd paraméterekre bontására és abból történő visszaállítására alkalmas. Ezután ismertetek egy félautomatikus irreguláris-reguláris transzformációs eljárást, amely az új modellen alapul. Percepciós és akusztikai teszttel igazolom, hogy a transzformációs módszer alkalmas irreguláris beszéd javítására.

A továbbiakban megmutatom, hogy az új gerjesztési modell illeszthető a statisztikai parametrikus keretrendszerhez. A rejtett Markov-modell alapú beszédszintézist kiegészítem az új modell használatával és igazolom, hogy ez javítja a gépi beszéd minőségét. Ahhoz, hogy a beszédben frázishatárokon előforduló irreguláris zöngét beszédszintézisben modellezzem, két alternatív kiegészítést javasolok. Az első egy szabály alapú modell, míg a második adatvezérelt megközelítés. Percepciós és akusztikai tesztek során mindkét modell javítja az alaprendszert kellemesség, eredeti beszélőhöz való hasonlóság és rekedtesség szempontjából.

Ezután bemutatok egy új modellt, amely a magyar magánhangzó formánsok és az alsó légúti rezonanciák (szubglottális rezonanciák) kapcsolatát vizsgálja. A modellt alkalmazom automatikus magánhangzó osztályozóban, amely szubglottális rezonancia alapú formáns normalizálást használva nagyobb pontosságot eredményez egy döntési fa alapú referencia osztályozónál.

A dolgozatban javasolt modellek és módszerek hozzájárulhatnak a természetesebb, expresszív és személyre szabott beszédszintézis rendszerek kialakításához. Az irreguláris zönge megfelelő modellezése a beszédtechnológiában egyrészt javíthatja a rekedtes hangokat, másrészt alkalmas lehet kifejező (expresszív) beszédszintézisre. A bemutatott modellek felhasználhatóak lehetnek nagyméretű beszédadatbázisok automatikus javítására is.

Abstract

of the PhD Thesis of Tamás Gábor Csapó,
"Increasing the naturalness of synthesized speech
in hidden Markov-model based text-to-speech synthesis"

The goal of text-to-speech synthesis is to convert written text to human-like speech. State-of-the-art systems frequently use statistical parametric methods. Often the hidden Markov-model based framework is applied for this purpose with the source-filter model of speech production. In natural speech the quasi-periodic vibration of vocal folds might become irregular for shorter or longer periods of time and fluctuations appear in the period-by-period amplitude and/or fundamental frequency. This can result a rough, creaky voice, which is integral part of natural speech. Until the start of my research it has not been extensively investigated in speech synthesis yet. In current speech technology it is almost exclusively assumed that the source and filter can perfectly be separated. However, according to recent research, nonlinear coupling occurs due to the interaction between the source and the filter. The lower airways also contribute to the shaping of speech sounds and investigating them helps the understanding of the separation and interaction.

In this thesis I present my results in the above topics. First I introduce a new excitation model which can decompose speech to parameters and restore the signal from them. After that I present a semi-automatic irregular-to-regular transformation method using the new model. A perception and an acoustic experiment have shown the suitability of the proposed transformation method to create regular speech from irregular speech.

Next I show that this new excitation model fits well in the statistical parametric speech synthesis framework. I extend the hidden Markov-model based speech synthesis with the novel model and show that it results in improved quality. To model the irregular voice typically occurring in phrase boundaries of speech, two alternative solutions are proposed for statistical parametric speech synthesis. The first one is a rule-based model, while the second is a data-driven approach. In perception and acoustic tests both methods are found to improve the baseline excitation in pleasantness, similarity to the original speaker and creakiness.

After that I propose a new model that investigates the relation between Hungarian vowel formants and the resonances of the lower airways (subglottal resonances). The model is applied in an automatic vowel classifier that is using subglottal resonance based formant normalization and results in improved accuracy compared to a baseline decision-tree based classifier.

The proposed models and methods may contribute to building natural, expressive and personalized speech synthesis systems. The proper modeling of irregular voice in speech technology can enhance creaky voices or express emotions with synthesized speech. The presented models could also be used for automatically correcting large speech databases.

Tartalomjegyzék

Kivonat **5**

Abstract **6**

Tartalomjegyzék **7**

Ábrák jegyzéke **10**

Táblázatok jegyzéke **12**

Rövidítések **13**

Jelölések **15**

Előszó **16**

1. A témakör bemutatása és a problémafelvetés **18**

1.1. Emberi és gépi beszédkeltés 18

1.2. Statisztikai parametrikus beszédszintézis 21

1.3. Beszédkódolás és gerjesztési modellek a statisztikai parametrikus beszédszintézisben 23

1.3.1. Beszédkódolás 23

1.3.2. Impulzus-zaj modell 24

1.3.3. Kevert gerjesztés alapú modellek 25

1.3.4. Glottális forrásjel modellek 25

1.3.5. Harmonikus-zaj modell 26

1.3.6. Maradékjel alapú modellek 27

1.4. Irreguláris zöngeképzés 28

1.4.1. Irreguláris zöngeképzés előfordulása és detekciója 29

1.4.2. Reguláris beszéd transzformációja irreguláris beszéddé 29

1.4.3. Irreguláris zöngeképzés a beszédszintézisben 30

1.5. Szubglottális rezonanciák hatása a beszédre 31

1.5.1. Kvantális elmélet 32
1.5.2. Szubglottális rezonanciák elemzése és alkalmazása 32

2. Kutatási célkitűzések 35

3. Módszertan 36
3.1. Felhasznált beszédkorpuszok 36
3.2. Felvételi körülmények 37
3.3. Alkalmazott eszközök és szoftverek 38
3.4. Meghallgatásos tesztek 39
3.5. Szignifikancia vizsgálatok 40

4. Újszerű gerjesztési modell kidolgozása 41
4.1. Új, MGC maradékjel kódkönyv alapú gerjesztési modell kidolgozása 42
4.1.1. Analízis 42
4.1.2. Szintézis 45
4.2. Az új gerjesztési modell felhasználása irreguláris zöngeképzés javítására 48
4.2.1. Transzformáció 48
4.2.2. Meghallgatásos teszt 51
4.2.3. Akusztikus elemzés 52
4.3. Összegzés 55

5. A gépi beszéd-előállítás természetességének növelése 56
5.1. Az új gerjesztési modell illesztése rejtett Markov-modell alapú szövegfelolvasóhoz 56
5.1.1. HMM-TTS alaprendszer impulzus-zaj modellel 57
5.1.2. Az új gerjesztési modell beépítése HMM-TTS-be 57
5.1.3. Meghallgatásos teszt 59
5.1.4. Irreguláris zönge kezelése az alaprendszerben 62
5.2. Az új gerjesztési modell felhasználása irreguláris beszéd gépi előállítására . . . 62
5.2.1. Szabály alapú irreguláris zönge modell kidolgozása 62
5.2.2. Meghallgatásos teszt a szabály alapú modell vizsgálatára 65
5.2.3. Adatvezérelt irreguláris zönge modell kidolgozása 68
5.2.4. Meghallgatásos teszt az adatvezérelt modell vizsgálatára 70
5.2.5. Akusztikus elemzés 71
5.3. Összegzés 74

6. Szubglottális rezonanciák elemzése a magyar beszédben 76
6.1. Kísérlet a szubglottális rezonanciák beszédre vonatkozó hatásának vizsgálatára 77

6.1.1. A magyar magánhangzók rendszere szubglottális rezonanciák szempontjából . . . 78
6.1.2. Modell a szubglottális rezonanciák beszédre vonatkozó hatására 78
6.1.3. Beszélőnkénti elemzés . . . 79
6.1.4. Normalizált elemzés . . . 81
6.1.5. Optimális kategóriahatárok vizsgálata . . . 83
6.2. Automatikus, szubglottális rezonancia-normalizáció alapú magánhangzó osztályozó kidolgozása . . . 83
6.2.1. Döntési fa alapú referencia osztályozó . . . 84
6.2.2. Szubglottális rezonancia-normalizálás alapú osztályozó . . . 84
6.2.3. A két osztályozó összehasonlítása . . . 86
6.3. Összegzés . . . 87

7. Összefoglalás és tézisek **89**
7.1. Az eredmények alkalmazhatósága . . . 96

Köszönetnyilvánítás **98**

Irodalomjegyzék **100**

A szerző tudományos közleményei **109**
A tézispontokhoz kapcsolódó tudományos közlemények . . . 109
A szerző további tudományos közleményei . . . 110

Ábrák jegyzéke

1.1. Az emberi hangképzés: hangképző és artikulációs szervek. 19
1.2. Általános szövegfelolvasó megvalósítási sémája. 19
1.3. A HMM-TTS rendszer általános felépítése. 22
1.4. A HTS rendszerben lévő alap impulzus-zaj gerjesztés. 25
1.5. Reguláris és irreguláris zöngével képzett beszéd: a „cipő” szó két változata. . . 28
1.6. Az alsó légúti (szubglottális) rendszer. 31
1.7. A kvantális elmélet szerinti nemlineáris kapcsolat az artikulációs és akusztikai paraméterek között. 33

4.1. Beszédjel analízise és szintézise az MGC maradékjel kódkönyv alapú módszerrel. 43
4.2. Példa a beszédjelből számított maradékjelre és a meghatározott periódusokra egy zöngés szakaszon. 44
4.3. Az $rt0$ paraméter számítása egy ablakozott maradékjel kódkönyv elemre. . . . 45
4.4. Példa az analízis során kinyert paraméter értékekre egy hosszabb beszédmintán. 46
4.5. Példa a szintetizált beszédjelre és az összefűzött maradékjelre a 4.2. ábra beszédmintáján. 46
4.6. Az MGC maradékjel kódkönyv alapú gerjesztési modellt felhasználó irreguláris-reguláris transzformáció működése. 49
4.7. A „cipő” szó hullámformái és maradékjelei (eredeti reguláris és irreguláris, valamint transzformált változatok). 50
4.8. Az irreguláris-reguláris transzformációval módosított szavak szubjektív elemzésének eredménye. 52
4.9. Az első két harmonikus és az első három formáns frekvenciájának és amplitúdójának mérése az FFT spektrum alapján. 54
4.10. Az irreguláris-reguláris transzformációval módosított szavak akusztikus elemzésének eredménye. 54

5.1. A HMM-TTS rendszer kiegészítése az új, MGC maradékjel kódkönyv alapú gerjesztési modellel (HTS-CDBK) 58

5.2. Az „ilyen” szó szintetizált és természetes gerjesztőjele valamint beszéd hullámformája. 60
5.3. A HTS-PN és HTS-CDBK beszédszintézis rendszerek szubjektív összehasonlításának eredménye. 61
5.4. A szabály alapú irreguláris zönge modell szintézis része (HTS-CDBK+Irreg-Rule). 64
5.5. A „Mihály” szó szintetizált változatai (alaprendszer, HTS-CDBK+Irreg-Rule, HTS-CDBK+Irreg-Data). 66
5.6. A HTS-CDBK alaprendszerrel és HTS-CDBK+Irreg-Rule szabály alapú irreguláris zönge modellekkel szintetizált szavak szubjektív összehasonlításának eredménye. 67
5.7. Az adatvezérelt irreguláris zönge modell szintézis része (HTS-CDBK+Irreg-Data). 69
5.8. A HTS-CDBK alaprendszerrel és HTS-CDBK+Irreg-Data adatvezérelt irreguláris zönge modellekkel szintetizált szavak szubjektív összehasonlításának eredménye. 72
5.9. A HTS-CDBK-Irreg-Rule és HTS-CDBK+Irreg-Data irreguláris zönge modellekkel szintetizált szavak szubjektív összehasonlításának eredménye. 72
5.10. Az irreguláris zönge modellekkel szintetizált szavak akusztikus elemzésének eredménye. 73

6.1. Szubglottális jel spektrogramja és LPC spektruma Log_FF2 beszélő „adaba” szava alapjána. 78
6.2. Magyar magánhangzók formánstere a szubglottális rezonanciákkal kiegészítve. 80
6.3. Négy beszélő formánsainak és szubglottális rezonanciáinak kapcsolata logatom felvételek alapján. 80
6.4. Szubglottális rezonanciák szerint normalizált formáns hisztogramok logatom beszéd alapján. 82
6.5. ROC elemzés eredménye a szubglottális rezonanciák magánhangzó csoportokra elkülönítésének vizsgálatára. 82
6.6. Példa a formáns alapú döntési fára. 85
6.7. A tisztán formáns alapú döntési fa és SGR-normalizált formáns alapú automatikus osztályozók pontosságának összehasonlítása. 87

Táblázatok jegyzéke

3.1. A PPBA adatbázisból az elemzésekhez kiválasztott beszélők hanganyagának adatai. 36
3.2. A meghallgatásos tesztek összesített tesztelői adatai. 39

5.1. A HTS-PN és HTS-CDBK rendszerek paramétereinek összehasonlítása. 59

6.1. Négy beszélő logatom felvételein mért szubglottális rezonancia értékek mediánjai. 77
6.2. A magyar magánhangzók fonológiai osztályozása. 79
6.3. A magyar magánhangzók artikulációs tulajdonságai. 79
6.4. Hat beszélő olvasott beszéd felvételein mért szubglottális rezonancia értékek mediánjai. 84

Rövidítések

ANOVA	ANalysis Of VAriance / Varianciaanalízis
BEA	BEszélt nyelvi Adatbázis
C	Consonant / Mássalhangzó
CELP	Code-Excited Linear Prediction
CMOS	Comparative Mean Opinion Score
DSM	Deterministic plus Stochastic Model / Determinisztikus-sztochasztikus modell
EGG	Electoroglottograph
GCI	Glottal Closure Instant
GSS	Glottal Spectral Separation
HMM	Hidden Markov-model / Rejtett Markov-modell
HNM	Harmonic plus Noise Model / Harmonikus-zaj modell
HNR	Harmonics-To-Noise Ratio / Harmonikus-zaj arány
HTS	HMM-based Speech Synthesis System (H-Triple-S)
IAIF	Iterative Adaptive Inverse Filtering
IPA	International Phonetic Alphabet / Nemzetközi Fonetikai Ábécé
LF	Liljencrants-Fant
LPC	Linear Predictive Coding / Lineáris Predikciós Kódolás
MELP	Mixed Excitation Linear Prediction
MGC	Mel-Generalized Ceptstrum / Mel-Általánosított Kepsztrum
MGLSA	Mel-Generalized Log Spectral Approximation
MOS	Mean Opinion Score
MVF	Maximum Voiced Frequency
MSD	Multi-Space Distribution / Többterű eloszlás
OQ	Open Quotient / Nyitott hányad
PCA	Principal Component Analysis / Főkomponensanalízis
PN	Pulse-Noise / Impulzus-zaj
PPBA	Preciziós, Párhuzamos magyar Beszédadatbázis
PSOLA	Pitch Synchronous Overlap and Add / Zöngeszinkron átlapoló összegzés
QT	Quantal Theory / Kvantális elmélet

RAPT	Robust Algorithm for Pitch Tracking
RMS	Root Mean Square / Négyzetes átlag
RMSE	Root Mean Squared Error / Átlagos négyzetes hiba
ROC	Receiver Operating Characteristics
SEDREAMS	Speech Event Detection using the Residual Excitation And a Mean-based Signal
SGR	Subglottal Resonance / Szubglottális rezonancia
SPTK	Speech Signal Processing Toolkit
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum
TL	Spectral Tilt / Spektrális lejtés
SVM	Support Vector Machine / Szupport vektor gép
TTS	Text-To-Speech / Gépi szövegfelolvasás
VCO	Voicing Cut-off Frequency
V	Vowel / Magánhangzó
WI	Waveform Interpolation

Jelölések

A1	Első formáns amplitúdója
A2	Második formáns amplitúdója
A3, A3*	Harmadik formáns amplitúdója (*: korrigált érték)
B1	Első formáns sávszélessége
F0	Alapfrekvencia
F1	Első formáns frekvenciája
F2	Második formáns frekvenciája
F3	Harmadik formáns frekvenciája
FF1, FF2, FF3, FF4	PPBA adatbázis négy férfi beszélője
Fn1	Sg1-normalizált első formáns
Fn2	Sg2-normalizált második formáns
Fn3	Sg3-normalizált második formáns
H1, H1*	Első harmonikus (*: korrigált érték)
H2, H2*	Második harmonikus (*: korrigált érték)
HTS-CDBK	Maradékjel kódkönyv gerjesztésű HTS
HTS-CDBK+Irreg-Rule	Szabály alapú irreguláris zönge modell HTS-ben
HTS-CDBK+Irreg-Data	Adatvezérelt irreguláris zönge modell HTS-ben
HTS-HUN	A HTS rendszer magyar nyelvű változata
HTS-PN	Impulzus-zaj gerjesztésű HTS
gain	Maradékjel periódus energiája
Log_FF1, Log_FF2	Logatom felvételek két férfi beszélője
Log_NO1, Log_NO2	Logatom felvételek két női beszélője
NO3	PPBA adatbázis egyik női beszélője
rt0	Maradékjel periódus csúcsok leírásának paramétere
Sg1	Első szubglottális rezonancia
Sg2	Második szubglottális rezonancia
Sg3	Harmadik szubglottális rezonancia
Spo_FF1 ... Spo_FF5	Spontán beszéd felvételek öt férfi beszélője
Spo_NO1	Spontán beszéd felvételek egy női beszélője

Előszó

Az információs társadalomban az ember-gép kapcsolat kutatásába illeszkedik a beszéd gépi előállításának minél jobb minőségű megvalósítása. A felhasználó és a gép között beszéd segítségével megvalósuló kommunikáció igen fontos, ha a felhasználó keze és látása lekötött (pl. autóvezetés közben), illetve sérülés miatt nem használható (pl. látássérültek), továbbá ha az igénybe vett szolgáltatás telefonvonalon keresztül érhető el (pl. intelligens tudakozó, hírolvasás mobil eszközön). Az expresszív, érzelmeket imitáló gépi beszéd akkor lehet előnyös, ha hosszabb szöveg felolvasásában szeretnénk a monotonitást csökkenteni (pl. hangoskönyvek esetén). Az adott beszélő hangján megszólaló, személyre szabott gépi szövegfelolvasó rendszerek hasznosak lehetnek azon felhasználóknak is, akik sérülés vagy betegség miatt elvesztették hangképzési lehetőségüket.

A beszéd képzésének számos egyszerűsített modelljét hozták létre, melyek nagyrészt a forrás-szűrő szétválasztáson alapulnak [1]. A gégének, vagyis annak a hangképző szervnek, amit forrásnak tekintünk, durva modellje lehet akár egy egyszerű impulzussorozat a zöngés szakaszokban és fehér zaj a zöngétlen részeken. A toldalékcső (szájüreg, orrüreg, stb.), azaz a szűrő modellezésére is sokféle eljárást dolgoztak ki. A gépi szövegfelolvasás egyik legújabb technológiája, a statisztikai parametrikus beszédszintézis is sok esetben a forrás-szűrő modellt használja [2]. A toldalékcső modellezése már elérte azt a szintet, ahol a további minőség javulás csak nagy befektetett energiával érhető el és a kutatás nem ezen a ponton kritikus [3]. A forrásjel modellezésére azonban még nem született kiforrott technika, melynek segítségével a statisztikai parametrikus beszédszintézis hangkarakterisztikája általános körülmények között is elérné az elemkiválasztásos rendszerek[1] [4] nyújtotta természetességet. A forrás modellezése ma is aktív kutatási terület, amivel számos kutató foglalkozik.

A legtöbb beszédtechnológiai módszert idealizált beszéd feldolgozására készítették el. Ideális zöngés beszédet feltételezve a hangszalagok kváziperiodikus módon rezegnek, azaz az egyes zöngeperiódusok között csak kis változások figyelhetők meg. A természetes beszédben azonban a beszélők időnként ettől különböző zöngeképzéssel beszélnek, és a beszédjelben az ideálistól lényegesen eltérő jellegzetességű (pl. kiugró vagy erősen lecsökkent amplitúdó-

[1] Az elemkiválasztásos beszédszintézis lényege, hogy az élő személy hangjának rögzítésével kialakított beszédkorpuszból minél hosszabb elemeket (szavakat, szókapcsolatokat) egymás után fűzve próbálja meg a szöveghez tartozó beszédet előállítani.

jú) zöngeperiódusok is megfigyelhetők. Ugyan már léteznek módszerek ezen jelenségek elemzésére, detektálására és transzformációjára [5], de az ideálistól eltérő beszéd (pl. irreguláris zöngeképzés) szintézisben történő modellezésével és az ehhez kapcsolódó transzformációs eljárásokkal keveset foglalkoztak.

A fenti forrás-szűrő szétválasztáson alapuló modellek azt feltételezik, hogy a forrás és a szűrő tökéletesen szétválasztható az emberi beszédkeltés során. Azonban ez nem mindig teljesül, és nemlineáris csatolás jöhet létre a forrás és a szűrő közötti interakció miatt. Az utóbbi néhány évben kimutatták, hogy a gége és a felette lévő szervek mellett az alsó légúti rendszer (pl. tüdő, légcső, hörgők) is befolyásolja a beszédet [6]. Eszerint az alsó légúti (szubglottális, azaz gége alatti) rendszer hozzájárul a magánhangzók megkülönböztető jegyek szerinti elkülönüléséhez [7], azaz szerepet játszik a beszédhangok egymástól való megkülönböztetésében. Az alsó légúti rezonanciák beszédtechnológiai felhasználási lehetőségeit eddig csak kezdeti kísérletekben vizsgálták.

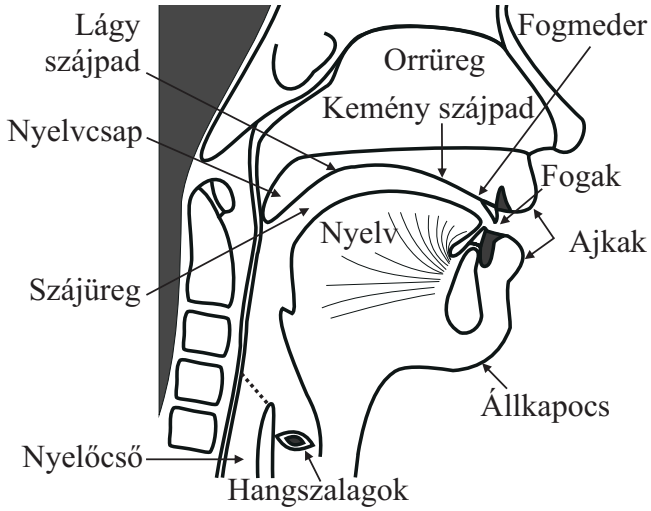
1. fejezet

A témakör bemutatása és a problémafelvetés

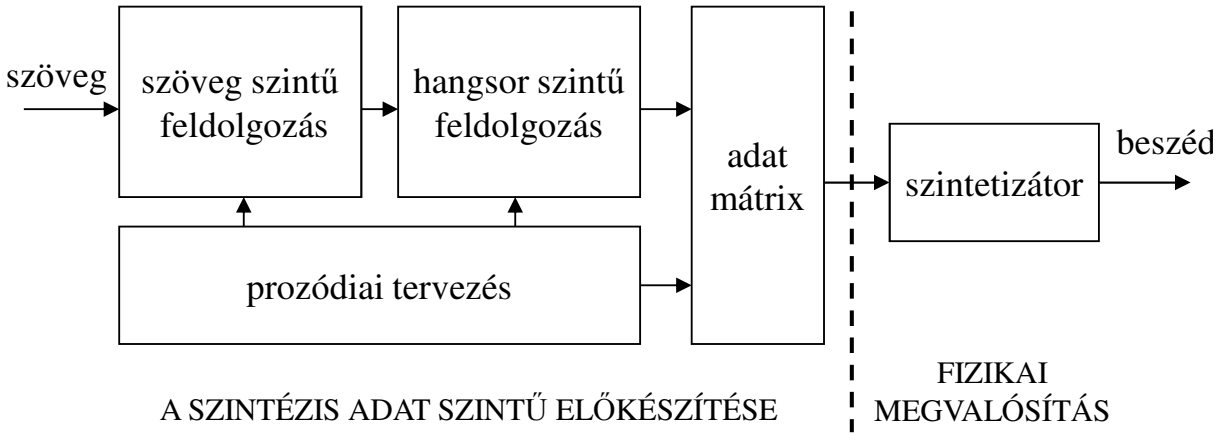
1.1. Emberi és gépi beszédkeltés

Az emberi beszédkeltés során a tüdőből kiáramló levegő a gégén keresztül jut el az artikulációs csatornába, amelynek segítségével ki tudjuk alakítani az egyes beszédhangokat [9, 22. oldal]. A beszédkeltés alapvető szervei a tüdő, a hörgők, a légcső (együttesen gége alatti, azaz szubglottális tér), a gége a hangszalagokkal (glottális tér, angolul *glottis*), illetve a garat, a szájüreg és az orrüreg (vagyis a gége feletti, azaz szupraglottális tér) [8, 19. oldal]. A beszéd folyamatát és az artikulációs szervek működését az agy vezérli. Zöngés beszéd esetén a gégében lévő hangszalagok a szubglottális térben megnövekedett nyomás miatt ismétlődően kinyílnak és záródnak nyomásingadozást eredményezve, ezáltal a megszaggatva a levegő kiáramlását és periodikus forrásjelet létrehozva. Zöngétlen hangok esetén a gége fúvó-, zár- vagy H-állásban van, melynek eredménye turbulens zaj-szerű gerjesztés vagy hirtelen zárfelpattanás [9, 27. oldal]. A gégében keletkező glottális forrásjelet az artikulációs csatorna formálja, melyet az 1.1. ábra mutat be. Az artikulációs csatorna alakját a nyelvcsap állításával, a nyelv vízszintes és függőleges mozgatásával, az állkapocs állításával, valamint a száj alakjának változtatásával tudjuk befolyásolni, ami különböző beszédhangok képzéséhez vezet [8].

A beszédszintézis nem más, mint emberihez hasonló beszéd előállítása mesterséges módon, tipikusan számítógép segítségével. Mivel a módszerek az emberi beszédkeltést próbálják valamilyen módon modellezni, a technológiát gépi beszédkeltésnek vagy gépi beszéd-előállításnak is nevezhetjük. Amennyiben a bemenet írott szöveg, gépi szövegfelolvasóról (angolul *Text-To-Speech*, TTS) beszélünk. A szöveget a beszédszintetizátor különböző lépéseken keresztül alakítja át emberihez hasonló hangzású beszéddé, melyre az 1.2. ábra mutat példát. Általános szö-



1.1. ábra. Az emberi hangképzés: hangképző és artikulációs szervek. Forrás: [8, 24. oldal].



1.2. ábra. Általános szövegfelolvasó megvalósítási sémája. A működés két fő lépésből áll: bemeneti szövegből szimbolikus információ létrehozása (bal oldal), majd ez alapján hangfájl szintetizálása (jobb oldal). Forrás: [10, 303. oldal] alapján, módosítva.

vegfelolvasó esetén ezek a lépések a bejövő szöveg feldolgozása, előkészítése a szintézishez (hangsor szintű feldolgozás és prozódia tervezés), valamint a beszéd létrehozása szintézissel [8].

A beszédszintetizátorok főbb generációit megkülönböztetjük működésük alapján: formánsszintézis, elemösszefűzés, elemkiválasztás és statisztikai parametrikus beszédszintézis [11]. A formánsszintézis volt az első olyan technológia, mellyel szöveget automatikusan érthető beszéddé lehetett alakítani. A rendszer az emberi beszéd formánsainak modellezésével próbálja létrehozni a beszédhangot. Ez kis számítástechnikai kapacitást igénylő megoldás (memóriaigénye akár 2–10 kB is lehet). Mivel a formánsszintézishez szükséges paraméterek megfelelő hangolása automatikus módszerekkel távol áll a tökéletestől, az ilyen rendszerek hangzása a sokszor megfelelő érthetőség ellenére meglehetősen „robotos”, ami háttérbe szorította őket.

Az elemösszefűzéses beszédszintézis során természetes beszédből kivágott hullámforma elemeket fűznek össze (angolul *concatenative synthesis*). Attól függően különböztetjük meg az elemösszefűzéses rendszereket, hogy mekkora a felhasznált elemek mérete: ez lehet diád (két félhang kapcsolata) vagy triád (környezetfüggő hang). A gyakorlatban a diádos-triádos beszédszintetizátorokkal közepes számítás- és memóriaigény (20–100 MB) mellett is jól érthető gépi beszédet lehet előállítani [11].

Az elemösszefűzéses technológia továbbfejlesztése az elemkiválasztásos beszédszintézis (angolul *unit selection*) [4]. Az újdonság itt egyrészt az, hogy nagyobb korpusz, vagyis beszédadatbázis áll rendelkezésre, amelyben egy-egy elem többször, többféle formában is előfordulhat. Másrészt ezek az elemek hosszabbak: szavak vagy akár szókapcsolatok is lehetnek. A kimeneti beszéd létrehozása során a rendszer minél hosszabb olyan elemeket keres a korpuszban, amelyek a bemeneti szöveghez illeszkednek. A diádos/triádos rendszerekhez képest az elemek hosszabbak, így kevesebb összefűzési pont lesz az előállított beszédben. Mivel a korpuszban egy adott hangsorhoz tartozó beszédelem többféle formában (különböző dallammal, intenzitással) is előfordulhat, ezek közül a legtermészetesebbet választva javítható a szintetizált beszéd minősége. Ugyanakkor a rendszer minőségét az is befolyásolja, hogy a szintetizálandó szöveg és a beszédkorpusz mennyire van közel egymáshoz: nem illeszkedő témájú bemeneti szöveg esetén zavaró ugrások jelenhetnek meg a beszédben. Az elemkiválasztás számításigénye nagy a megfelelő összefűzendő elemek keresése miatt, és a szükséges tárhely mérete is lényegesen nagyobb a többi beszédszintetizátor technológiához képest (mintegy 100 MB–5 GB). Az elemkiválasztásos rendszerek fő korlátja az, hogy csak egyféle hangon tudnak megszólalni, mivel a beszédkorpuszbeli hangsorozatokat használják. Így különböző beszédstílusok szintetizálásához egyre nagyobb adatbázis szükséges, amelynek előállítása meglehetősen költséges.

A statisztikai alapú parametrikus beszédszintetizátor rendszerek egyre népszerűbbé váltak az elmúlt évtizedben, ami többek között a számítástechnika fejlődésének köszönhető. Az itt alkalmazott technika leggyakrabban a rejtett Markov-modell (angolul *Hidden Makov-Model*, HMM), amelyről már régen kimutatták empirikusan, hogy jól alkalmazható beszédfelismerés-

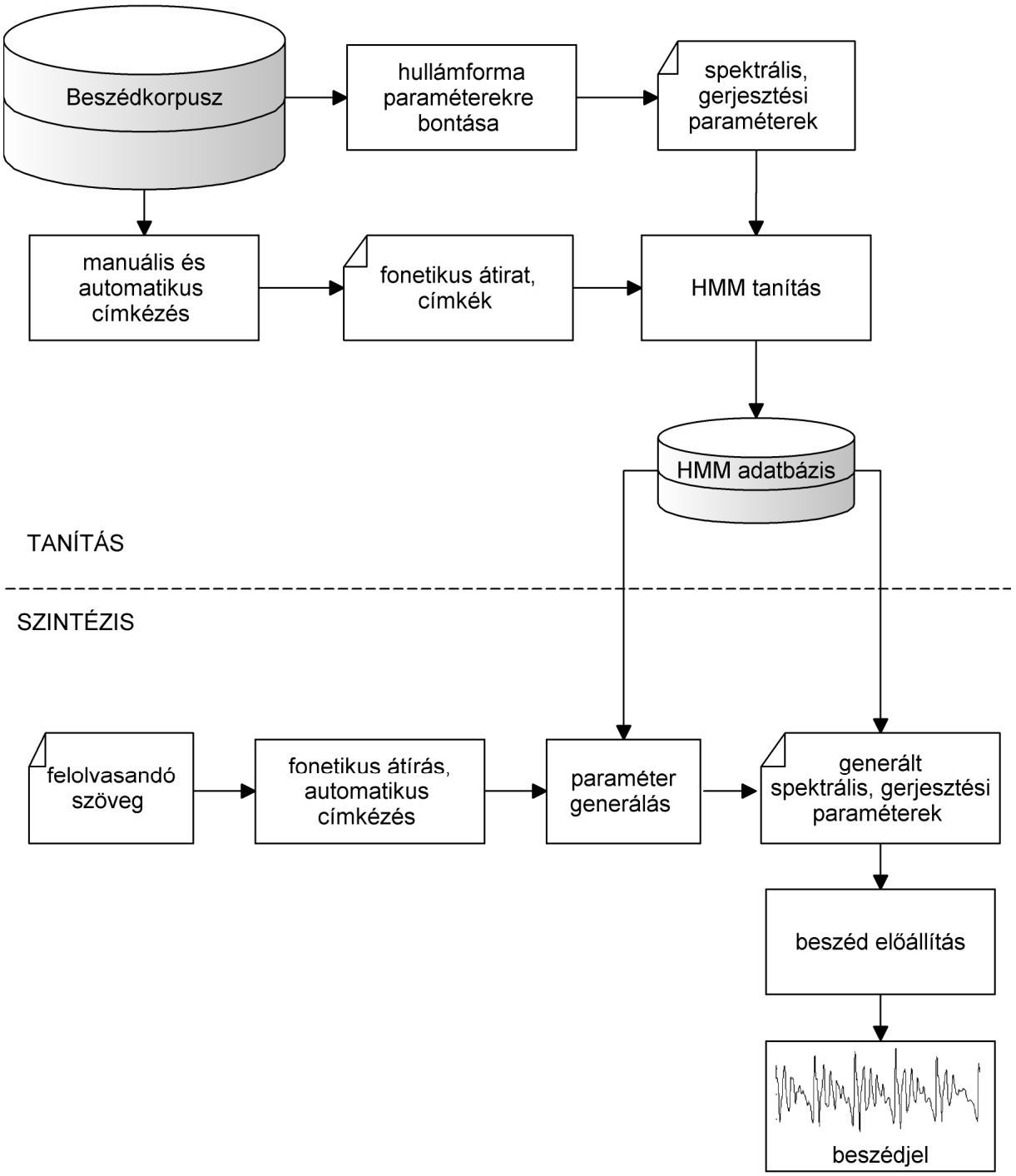
ben [12]. Az elemkiválasztásos beszédszintetizátor nagyméretű futásidejű adatbázisával szemben az új technológia alkalmazásához elég egy betanító korpusz, amelyből a rendszer a gépi tanulás során környezetfüggő HMM-eket majd modellparamétereket állít elő, a kimeneti hullámforma generálása pedig ezek alapján lehetséges. A betanítás hasonlóan történik, mint a beszédfelismerésnél (hiszen a HMM-eket eredetileg erre használták), míg a tényleges szintézis a felismerés inverze, aminek eredménye a hullámforma. Ezzel a módszerrel lehetővé válik különböző beszédstílusok, érzelmek modellezése a HMM paraméterek megfelelő módosításával. A statisztikai parametrikus beszédszintézissel a következő fejezetben részletesebben is foglalkozunk, mivel a disszertációban ismertetett kutatás ehhez a módszerhez kapcsolódik.

1.2. Statisztikai parametrikus beszédszintézis

Az előző fejezet ismertetése szerint a legkorszerűbb beszédszintézis technológiák egyike a rejtett Markov-modell alapú szövegfelolvasó (angolul *Hidden Makov-Model based Text-To-Speech*, HMM-TTS), amely a statisztikai parametrikus beszédszintetizátorok családjába sorolható [3, 13]. Ennek egyik kutatási eszköze a nyílt forráskódú HTS (*HMM-based Speech Synthesis System, H-Triple-S*) rendszer [2]. Az elmúlt években a HMM-TTS nagy népszerűségre tett szert számos előnyös tulajdonsága miatt: flexibilis, alacsony memóriaigényű és nem tartalmaz olyan zavaró akusztikai torzításokat, mint a korábbi elemkiválasztásos rendszerek [3].

A HMM-TTS általános felépítésére az 1.3. ábra mutat példát. Az ábra szaggatott vonal feletti részén láthatóak a tanítás lépései, amelyeket előzetesen, a tipikusan néhány óra hosszúságú beszédkorpusz rendelkezésre állása esetén el lehet végezni. A beszédkorpuszból részben automatikus címkézés és fonetikus átírás, részben manuális javítások segítségével készül el a fonetikus átirat és a környezetfüggő címkézés. A statisztikai parametrikus beszédszintézis során nem közvetlenül a beszédadatbázis hullámformáin végzünk átalakításokat, hanem a beszédet először paraméter-sorozatokra bontjuk, amelyeket gépi tanuló algoritmus kezel a továbbiakban. A paraméterekre bontás lehetséges módszereit az 1.3. fejezetben ismertetem részletesen. A címkék és paraméterek alapján elvégezhető a HMM-ek tanítása. A tanítás eredménye a kisméretű HMM adatbázis, amely a szintézisben használható fel. A szintézis során (1.3. ábra szaggatott vonal alatti része) felolvasandó szöveghez automatikus fonetikus átirat és környezetfüggő címkézés készül, majd a HMM adatbázis alapján a címkézett szöveghez megfelelő paramétereket generál a rendszer. A generált gerjesztési és spektrális paraméterekből beszéd visszaállító eljárással (1.3. fejezet) készül a szintetizált beszéd hullámforma [15]. A tanítás folyamata több napig is eltarthat egy korszerű számítógépen, míg a szintézis valós időben is működhet.

Fontos kiemelni, hogy nem feltétlenül alkalmas a beszéd tetszőleges paraméterű felbontása a gépi tanulásra. A HTS nyílt forráskódú beszédszintetizátor alap változata a beszédet például az alapfrekvenciával és a spektrum egy reprezentációjával jellemzi, melyek a tapasztalatok szerint használhatók HMM tanításra. Más, bonyolultabb paraméterekre bontás azonban nem



1.3. ábra. A HMM-TTS rendszer általános felépítése. A szaggatott vonal feletti rész a tanítási fázis, a szaggatott vonal alatti a szintézis fázis. Négyzetek jelölik az eljárásokat; a behajtott sarkú négyzetek a paramétereket és fájlokat jelzik. Forrás: [14] alapján, módosítva.

feltétlenül vezet eredményes gépi tanulásra [16]. Az eredménytelen gépi tanulást többek között az okozhatja, ha a paraméterek nem rendelhetők hozzá fonémákhoz, vagy nagyobb nyelvi egységekhez. Ez esetben a HMM-ek tanítása nem sikeres, és a HMM adatbázis rendelkezésre állása nélkül nem lehetséges a beszéd szintézise. Egy kutatás szerint a Gauss eloszlású paraméterek általában megfelelőek a HMM-TTS céljaira [16].

A rejtett Markov-modell alapú beszédszintézis aktív kutatási terület. Számos kutató foglalkozik különböző résztémákkal: beszéd paraméterekre bontása, beszélő adaptáció, beszélő interpoláció, felügyelet nélküli tanítás (melyről például Tóth munkájában olvashatunk részletesebben [14]). A következő fejezetben a beszéd paraméterekre bontásával foglalkozom.

1.3. Beszédkódolás és gerjesztési modellek a statisztikai parametrikus beszédszintézisben

A rejtett Markov-modell alapú beszédszintézis egyik fontos kutatási területe a beszéd paraméterekre bontása (analízise) és paraméterekből történő visszaállítása (szintézise). Ezek az eljárások nagymértékben befolyásolják a szintetizált beszéd gépiességét, minőségét. Erre a feladatra beszédkódoló eljárásokat lehet alkalmazni, azonban figyelembe kell venni, hogy a paramétereken végzendő gépi tanulást össze kell hangolni az alkalmazott beszédkódolóval. A gyakorlatban a forrás-szűrő szétválasztáson alapuló gerjesztési modelleket találták megfelelőnek a feladatra, melyeket az 1.3.2–1.3.6. fejezetekben mutatok be. A HTS rendszerben eddig alkalmazott gerjesztési modellek nagy részét Hu és társai elemzése is összefoglalja [17].

1.3.1. Beszédkódolás

A szakirodalomban számos beszédkódoló módszerről olvashatunk, melyeknek célja a beszéd paraméterekre bontása és kódolása azért, hogy a távközlési csatornán minél kisebb sávszélesség mellett lehessen átvinni jól érthető beszédet [8, 244. oldal]. A kódolási technikákat három csoportba lehet osztani: hullámforma-kódolás, parametrikus vagy forráskódolás, és hibrid kódolás. A hullámforma kódolás tetszőleges sávkorlátozott jel digitális tárolására alkalmas, és a jel redundanciájának csökkentésével törekszik az alakhűség megtartására. A parametrikus kódolás esetén forrásmodellt (beszédkeltési modellt) alkalmaznak, ami miatt ez a kódolási forma csak beszédjelre alkalmazható. A hibrid kódolás az előző kettő előnyeit ötvözi [8]. Ezen kódoló típusokból a tapasztalatok szerint a parametrikus kódolók felelnek meg a beszédszintézis céljaira.

A parametrikus kódolók családjába tartozik az LPC kódoló (*Linear Predictive Coding*, [18, 264. oldal]), a MELP (*Mixed Excitation Linear Prediction*, [19]), a CELP (*Code-Excited Linear Prediction*, [18, 299. oldal]) jellegű eljárások és ezek kombinált illetve javított változatai. Ezen beszédkódolók egy részét sikerrel alkalmazták statisztikai parametrikus beszédszintézisben is:

az LPC kódoló integrálását az 1.3.2. fejezet, a MELP kódoló alkalmazását az 1.3.3. fejezet mutatja be. CELP jellegű kódolókkal ugyan a korábbiaknál jobb beszédminőség érhető el, azonban kezdeti kísérleteink szerint ez nem alkalmas a gépi tanulórendszerbe történő integrálásra. A CELP kódoló kódkönyv indexe ugráló értékeket tartalmaz, ami nem modellezhető HMM-ekkel, és a sikeres gépi tanításhoz újfajta megközelítés lenne szükséges. Emiatt a későbbiekben újszerű parametrikus kódolást alkalmazunk.

1.3.2. Impulzus-zaj modell

A legtöbb HMM-TTS rendszer a beszéd forrás-szűrő szétválasztásán alapul [1]. Eszerint az $u_G(n)$ gerjesztő jelen, amely a gégében lévő glottális forrásjel egyszerűsített modellje a lineáris $v(n)$ rendszerrel végzünk spektrális szűrést, amely a toldalékcső modellje [16]. A forrásjel és a spektrális szűrő konvolúciójaként kapjuk meg az $u_L(n)$ beszéd-szerű jelet:

$$u_L(n) = u_G(n) * v(n), \tag{1.1}$$

melyet z tartományba transzformálhatunk:

$$U_L(z) = U_G(z) \cdot V(z). \tag{1.2}$$

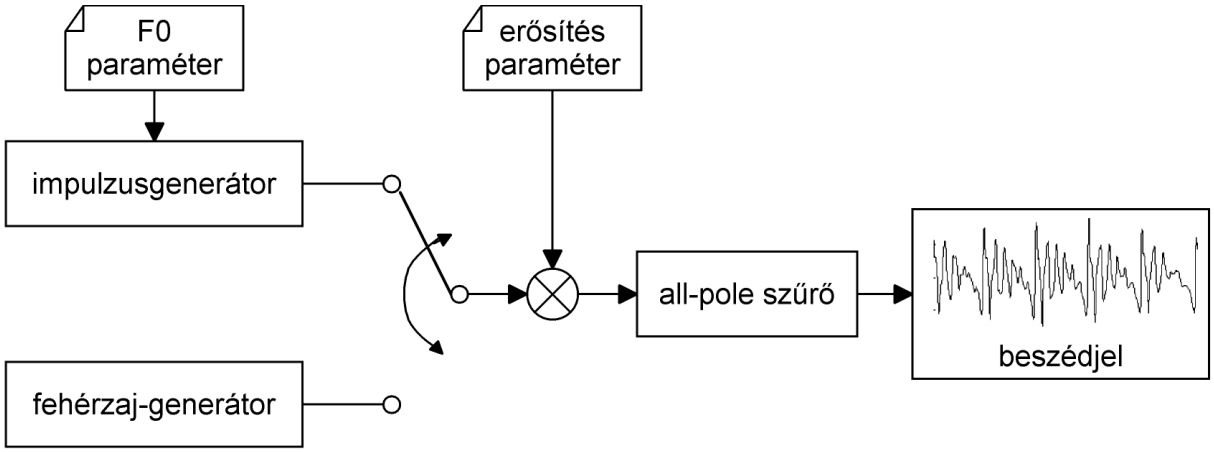
A toldalékcső $V(z)$ átviteli függvényét megfelelően lehet modellezni egy all-pole szűrővel [20]:

$$V(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k \cdot z^{-k}}, \tag{1.3}$$

ahol a G és $\{\alpha_k\}$ a toldalékcső alakjától függenek [20]. Az all-pole szűrőben a pólusokkal csak a rezonanciák modellezése lehetséges. A nazális hangokban lévő anti-formánsok szintéziséhez zérusokra is szükség lenne, amit a fenti egyszerű modellben nem szokás használni.

A forrásra a legegyszerűbb modell az impulzus-zaj módszer (angolul *pulse-noise*), melyre az 1.4. ábra mutat példát: a zöngés szakaszokat alapfrekvencia-függő ($F0$) impulzussorozattal, a zöngétlen részeket sávkorlátozott fehér zajjal modellezzük. A forrás kereteinek összefűzése és erősítés után all-pole szűréssel kapjuk meg a beszédjelet. A lineáris predikciós beszédkódolók legegyszerűbb változataiban (pl. LPC-10, [18, 264. oldal]) is ezt a modellt valósították meg. A HTS rendszer impulzus-zaj gerjesztésű változatában (HTS-PN) a beszéd hullámformák leírása az $F0$ forrás és MGC (*Mel-Generalized Cepstrum*) spektrális paraméterekkel történik.

Az alap HTS rendszerben lévő egyszerű impulzus-zaj gerjesztés azonban a HMM-TTS szintézis minőségét „zizegőssé”, robotossá teszi az elemkiválasztásos rendszerek tiszta, csengő hangjához képest (zizegős beszéden az egyszerű beszédkódolók által eredményezett fémes,



1.4. ábra. A HTS rendszerben lévő alap impulzus-zaj gerjesztés. Forrás: [18, 264. oldal] alapján, módosítva.

gépies, robotos hangot értem; angol megfelelője: *buzzy*). Azért, hogy ezt a jelenséget kiküszöböljék, számos továbbfejlesztett gerjesztési modellt javasoltak a szakirodalomban, melyeket különböző kategóriákba sorolhatunk az alkalmazott modell típusa és a gerjesztő jel szerint.

1.3.3. Kevert gerjesztés alapú modellek

A kevert gerjesztés [21], a kétsávos gerjesztés [22] és a STRAIGHT beszédkódoló használata [23] rendkívül jó minőségű HMM-alapú gépi beszédet eredményez [24], azonban ezek nehézkesen építhetők be valós idejű alkalmazásokba nagy számításigényük miatt. A kevert gerjesztés (angolul *mixed excitation*) lényege, hogy a forrásjel nem egyértelműen zöngés vagy zöngétlen, hanem ezek sávszűrt szuperpozíciójaként áll elő. A gerjesztésnek még jobb modellje a kevert gerjesztés kiegészítése állapotfüggő szűréssel, amelynek működése az analízis-szintézissel típusú beszédkódolókkal egyezik meg [25, 26]. Maia és társainak legújabb munkája ezt egészíti ki a komplex kepsztrum modellezésével, amely a beszéd kevert fázisú jellegzetességeinek leírására is megfelelő [27].

A kevert gerjesztés azon beszédhangok modellezésére különösen hasznos, amelyek nem egyértelműen zöngések vagy zöngétlenek, hanem ezek keverékeként jönnek létre (pl. zöngés réshangok, illetve gerjesztésváltás két hang között). A HTS-STRAIGHT rendszert széles körben használják (pl. [14]), mivel kutatási célra a HTS-PN-hez hasonlóan szabadon hozzáférhető.

1.3.4. Glottális forrásjel modellek

A glottális (azaz gégében lévő) forrásjel leírása és paraméterekre bontása már régóta aktív kutatási terület. Cabral és társai a glottális forrás deriváltjának Liljencrants-Fant által kidolgozott (LF) akusztikus modelljét [28] használják a gerjesztőjel előállítására [29]. Az LF

modell használata mellett egy erős érv az, hogy az LF hullámformának magasabb frekvenciákon csökkenő jellegű a spektruma, ami jobban hasonlít a valódi glottális forrásjelhez, mint az impulzus-zaj alapú vagy kevert gerjesztés [30]. Cabral és társai a továbbiakban glottális spektrális szétválasztást (*Glottal Spectral Separation*, GSS) is alkalmaznak, amelynek lényege, hogy a spektrum burkoló számítása során a glottális forrás hatásait megpróbálják minimálisra csökkenteni [31]. A végső rendszert HTS-LF-nek nevezik, amely a kísérletek szerint kis mértékben jobb, mint a HTS-STRAIGHT szintetizátor [32].

Raitio és társai a korábban kidolgozott glottális inverz szűrés eljárást (*Iterative Adaptive Inverse Filtering*, IAIF, [33]) használják fel és integrálják a HTS rendszerbe, melyet GlottHMM-nek neveznek [34, 35]. Az egyetlen pulzust felhasználó technikát [34] kiegészítik egy pulzus elem könyvtárral [36] és elemkiválasztással [37], ami pulzus összefűzés alapú hibrid parametrikus-elemkiválasztásos rendszert eredményez [38]. A legújabb kísérletekben viszont megmutatták, hogy a pulzusok átlagának felhasználása hasonlóan jó eredményre vezet, mint a pulzus könyvtárból történő komplex elemkiválasztás [39].

Az LF-paramétereket Lanchantin, Degottex és társaik a szintetizált beszéd levegősségének[1] állítására is sikerrel használják [40, 41]. Így tehát beszéd transzformációjára és expresszív beszéd szintézisére is lehetőség van az LF-modell alapú rendszerrel. A modellt Gauss zaj hozzáadásával egészítik ki, így a kevert gerjesztéshez közelítve a technikát [42]. A kiegészített módszer alkalmas HMM-alapú beszédszintézisre, beszéd levegősségének és dallamának módosítására is.

Összességében a glottális forrást alkalmazó rendszerek jó minőségű zöngés beszédet tudnak létrehozni, de a zöngés-zöngétlen átmenetek kezelése nem teljesen megoldott és stabilitási problémák fordulhatnak elő.

1.3.5. Harmonikus-zaj modell

Néhány módszer a Harmonikus-Zaj Modell (*Harmonic Plus Noise Model*, HNM) alkalmazását javasolja a HTS környezetben és a paraméterek közé veszi a maximális zöngés frekvenciát (*Maximum Voiced Frequency*, MVF) [43, 44] vagy zöngés vágási frekvenciát (*Voicing Cut-off Frequency*, VCO) [45, 46, 47], melyek a harmonikus és sztochasztikus részek elválasztására szolgálnak. Ezekben a rendszerekben a harmonikus részeket szinuszos jellel modellezik, míg a sztochasztikus részek Gauss-zaj megfelelően szűrt változatából állnak elő. Erro és társai rendszere (AhoTTS) a HTS-STRAIGHT-hez hasonló minőség létrehozására képes kisebb számításigény mellett [44].

Az MVF és VCO alapú rendszerek előnye, hogy a spektrum felsőbb frekvencia sávjaiban sztochasztikus zajt alkalmazva csökkenthető a szintetizált beszéd zizegőssége.

[1] Levegős beszéden (angolul *breathy*) azt a zöngeminőséget értjük, amikor a hangszalagok nem teljes záródása miatt nagymértékű aszpirációs zaj jelenik meg a beszédben, és az alsó harmonikusok felerősödnek. Érzetileg fátyolos hangnak nevezhetjük.

1.3.6. Maradékjel alapú modellek

Számos gerjesztési modell foglalkozik a beszédből származtatott maradékjellel. Ezen megoldások nagy előnye, hogy a maradékjel közvetlenül, automatikusan kinyerhető a beszédjelből lineáris predikció alapú inverz szűréssel, így nem kell például külön EGG (*Electoroglottograph*) felvételt rögzíteni és a glottális forrásjel becslése sem szükséges.

Az egyik ilyen modellben Wen és Tao a maradékjel paraméterekkel történő leírására az amplitúdó spektrumot használja, illetve zéró-fázisú kritériumot alkalmaz a maradékjel szintetizálásakor [45]. A módszert továbbfejlesztik spektrum normalizálással és kódkönyv építéssel, majd megmutatják, hogy a javasolt rendszerrel a HTS-STRAIGHT-hez hasonló minőség érhető el [48].

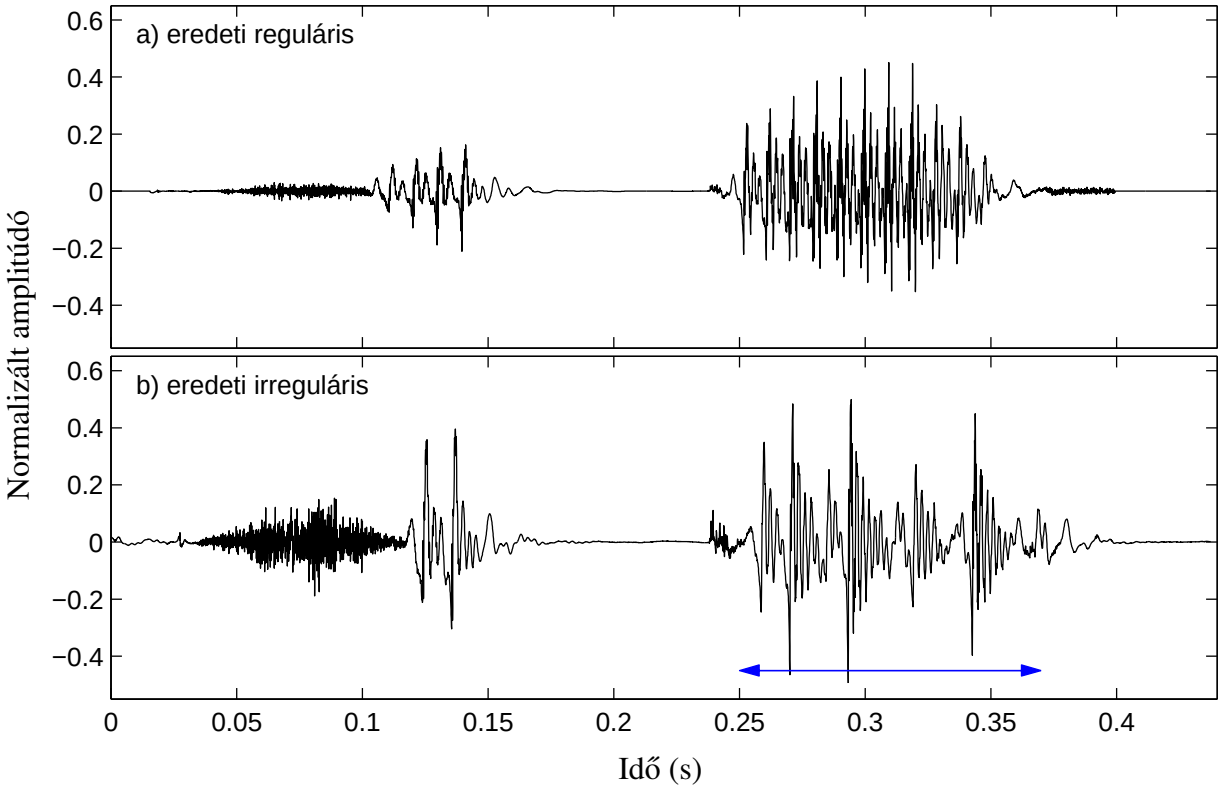
Sung és társai egy másik gerjesztési modellben a maradékjelből karakterisztikus hullámformákat vágnak ki, és hullámforma interpolációt (*Waveform Interpolation*, WI) alkalmaznak [49]. A modellt kiegészítik a lassan változó és gyorsan változó hullámforma fogalmának bevezetésével, ami alacsonyabb spektrális torzítást eredményez [50]. Emellett az idő- és frekvenciatartománybeli null-kitöltés hozzáadása a WI modellhez tovább csökkenti a spektrális torzítást [51].

Drugman és kollégái zöngeszinkron maradékjel kódkönyv építést használnak, majd PCA (*Principal Component Analysis*) eljárással tömörítik a kódkönyvet [52]. A módszer egyszerűsítéseként bevezetik a determinisztikus-sztochasztikus modellt (*Deterministic Plus Stochastic Model*, DSM), amely a korábbi kódkönyv alapú eljárással szemben egy „sajátmaradékjel” újramintavételezésével állítja elő a maradékjel periódusokat [53]. Itt a determinisztikus rész az alacsony frekvenciás zöngés tartomány, míg a magasabb frekvencia komponensekben sztochasztikus zajt alkalmaznak a HNM modellhez hasonlóan. A szerzők szerint a PCA-val tömörített sajátmaradékjel használatával a modell nagyon egyszerű és mégis jó minőségű: a maradékjel paraméterekkel történő leírására elegendő az alapfrekvencia. A HTS rendszerbe történő integrálás mellett a DSM modell nem csak beszédszintézisre, hanem alapfrekvencia módosításra és beszélő azonosításra is alkalmas [54, 55]. Nurminen és kollégái egy kezdeti kísérletben azt is megmutatták, hogy a maradékjel spektrumának modellezése tovább javíthatja a szintetizált beszéd minőségét [56].

A maradékjelen alapuló módszerek előnye az automatikus inverz szűrés mellett, hogy könnyen kiegészíthetőek a normáltól eltérő zöngeminőségű beszéd modellezésére.

A statisztikai parametrikus beszédszintézis alapmódszereit és a legtöbb fenti gerjesztési modellt ideális beszédre[2] dolgozták ki és optimalizálták. Azon beszélők esetén várhatóan nem eredményez jó minőséget, akiknél gyakran előfordul az ideálistól lényegesen eltérő zöngeképzés. Ennek egyik oka lehet az irreguláris fonáció, melyet a következő fejezetben ismertetek.

[2]Ideális beszéden azt értem, amikor a zöngés szakaszokban a hangszalagok rezgése kváziperiodikus vagy a beszéd mértékletesen megjelenő irreguláris zöngét tartalmaz.



1.5. ábra. Reguláris és irreguláris zöngével képzett beszéd: a „cipő” szó két változata FF3 beszélőtől. Vízszintes nyíl jelöli az irreguláris zöngét.

1.4. Irreguláris zöngeképzés

Az emberi beszédben ideális (más néven reguláris vagy modális) zöngeképzés esetén a hangszalagok kváziperiodikusan rezegnek. A gégében azonban hosszabb-rövidebb időtartamra instabilitás léphet fel, ami a hangszalagok irreguláris rezgését okozza. Ez eltér a modális zöngeképzéstől, és irreguláris fonációnak, glottalizációnak, érdes zöngének vagy recsegő beszédnek nevezik [5]. A kifejezés angol elnevezései: *irregular phonation, glottalization, creaky voice, vocal fry, laryngealization*. A jelenség a zöngeperiódusok hosszának és/vagy amplitúdójának hirtelen megváltozásából adódik. Az irreguláris fonáció előfordul egészséges és patologikus beszélők esetén is [57], általában szakaszhatárokon (pl. mondat végén) [58] vagy magánhangzó-magánhangzó kapcsolatokban [59, 60]. Gyakran kíséri extrém alacsony alapfrekvencia és a glottális pulzusok gyors lecsökkenése [61]. Érzetileg recsegő, érdes jellegű beszédet jelent [62]. Az 1.5. ábra egy példát mutat a modális és glottalizált beszéd közti különbségre: vízszintes nyíl jelöli az irreguláris zöngével képzett szakaszt. Az irreguláris szakaszban jól látható az amplitúdó ingadozás a reguláris kváziperiodikus hullámmal szemben.

Léteznek megoldások illetve kezdeti kísérletek az irreguláris zönge detekciójára [63, 64, 65, 66, 67], modális beszéd irregulárissá transzformálására [19, 62, 68] és érdes zöngével kiegészített beszédszintézisre [69, 70, 71]. A továbbiakban ezeket tekitjük át.

1.4.1. Irreguláris zöngeképzés előfordulása és detekciója

A glottalizáció előfordulása függ a prozódiai szerkezettől (gyakran egybeesik prozódiai határokkal, például szünetek [72] és hangsúlyos szótagok [73]), valamint információt hordoz a beszélő személyről, nyelvjárásáról, hangulatáról, érzelmi állapotáról és arról, hogy a hangszalagok egészségesek-e [74, 75]. A glottalizáció akár a beszédhangok 15%-ában is előfordulhat egy-egy beszélő esetén, így egyáltalán nem elhanyagolható jelenség [63]. Az irreguláris fonáció problémákat okozhat a beszédanalízis módszerekben (pl. $F0$ mérés és spektrális analízis). A fentiek miatt az irreguláris zöngével képzett beszéd megfelelő modellezése hozzájárulhat a természetesebb, érzelmeket imitáló és személyre szabott beszédszintetizátor rendszerek elkészítéséhez.

A zöngeminőség osztályozók általában néhány, a beszédjelen mért akusztikai paraméter alapján hoznak döntést arról, hogy a zöngét reguláris vagy irreguláris zöngével képezték-e. Surana négy akusztikai jegyet használ, és szupport vektor gép (*Support Vector Machine*, SVM) alapú osztályozást alkalmaz [64]. Ishi és társai három másik jegy bevezetését javasolják, amelyek a beszédjel nagyon rövid szakaszában számolt teljesítményén alapulnak, és egyszerű küszöbértéket használnak a döntéshez [65]. Bőhm egyesíti az előző két osztályozót és algoritmikus finomhangolással valamint SVM alapú osztályozással javítja a pontosságot [5, 63]. Kane és társai bemutatnak egy újszerű algoritmust, amely a lineáris predikciós maradékjel két új akusztikai paraméterét használja döntési fa alapú osztályozóval [67].

A fenti automatikus osztályozó eljárásokkal a reguláris és az irreguláris zöngével képzett beszéd közel tökéletesen elkülöníthető egymástól. Saját kísérleteinkben a Kane és társai által bemutatott irreguláris zönge detektort használjuk [67].

1.4.2. Reguláris beszéd transzformációja irreguláris beszéddé

Az irreguláris fonáció első modelljeit a szövegfelolvasók területén formánsszintézisben készítették el, másolás-szintézis kísérletekben [68]. Más kezdeti módszerek egyszerűen a beszédjel jitter és shimmer[3] értékének növelésével próbáltak érdes jellegű hatást elérni [19].

Bőhm létrehozott egy reguláris-irreguláris transzformációs módszert, amely az egyes glottális ciklusok amplitúdóját skálázza [5, 62]. Az eljárás a beszédet először zöngeszinkron módon ablakozza, a periódusokat megszorozza egyéni, kézzel beállított skálázó faktorokkal és végül átlapolt összeadással készíti el a módosított beszédjelet a PSOLA eljáráshoz (*Pitch Synchronous Overlap and Add*, [76]) hasonló módon. A skálázó faktorok erősíthetik, gyengíthetik, eltörölhetik, vagy nem változtatják az egyes ciklusokat. A módszer kiegészítéseként elkészült egy félautomatikus eljárás is, amely stilizált pulzus minták másolásával egyszerűsíti a transzformációt [5, 62]. A kísérletek szerint a reguláris-irreguláris transzformáció eredményeként kapott minták

[3] A jitter az alapfrekvencia, a shimmer az amplitúdó ingadozását jellemzi.

érzeti érdessége megfelel a természetes glottalizált mintákénak. Emellett egy objektív elemzés eredménye alapján három jellegzetes akusztikai jegy szempontjából is megfelelően módosítja az eljárás az eredetileg reguláris mintákat.

A transzformáció másik irányára, az irreguláris beszédminták regulárissá alakítására (amely egyik kutatási területem) nem találtunk eljárást a szakirodalomban.

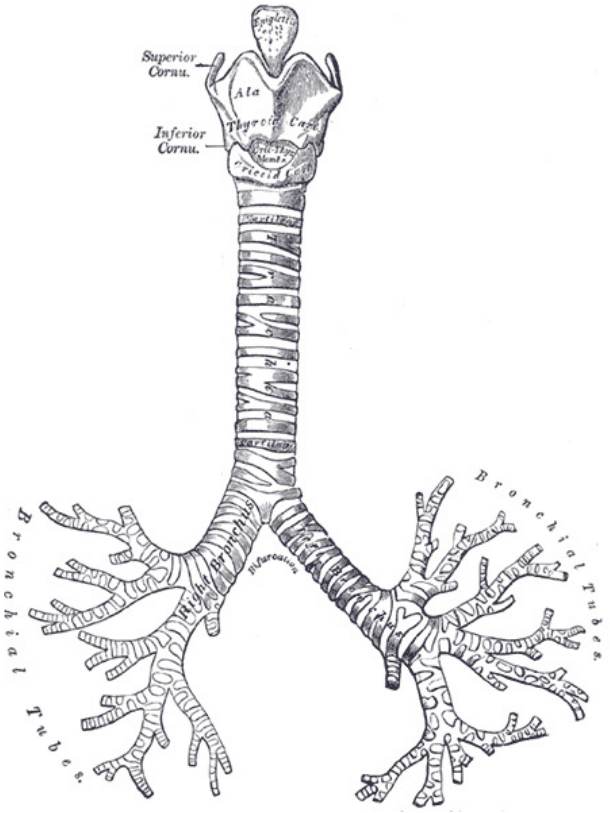
1.4.3. Irreguláris zöngeképzés a beszédszintézisben

A beszédszintézis rendszerekben a korpusz felvétele során általában bejelölik és kerülik a glottalizációval képzett beszédrészeket, mivel ezek a későbbi automatikus feldolgozást megzavarhatják. Silén és társai bemutatnak egy finn nyelvű elemkiválasztásos beszédszintetizátor rendszert, amely foglalkozik az irreguláris zöngeképzéssel [69]. A korpusz felvételénél azt vették észre, hogy a beszélők általában a szakaszok határa előtt képzik irreguláris zöngével a beszédet. Az elemkiválasztás súlyait ez alapján úgy módosítják, hogy a glottalizált részek a szintetizált mondatokban is csak a megfelelő (például mondat végi) pozícióba kerülhessenek. Zainkó és társai expresszív beszéd elemzése során észrevették, hogy a szomorú érzelmet a beszélők sokszor glottalizált beszéddel is próbálják jelezni, így az irreguláris zönge megfelelő modellezése segítheti az expresszív beszédszintézis rendszereket [77].

Statisztikai parametrikus beszédszintézisben eddig csak kezdeti kísérleteket végeztek az irreguláris zöngeképzés modellezésére [70, 71, 78]. A rejtett Markov-modell alapú beszédszintézisben Silén és társai módszerének lényege, hogy robusztus $F0$ mérést alkalmaz megbízható zöngésség detekcióval és kétsávos zönge kezeléssel, ezáltal eltüntetve a glottalizált beszédrészleteket a szintetizált beszédből [70]. Így viszont a beszélőre jellemző irreguláris fonáció teljesen elveszik a beszédszintézis kimenetéből, és az eljárás nem foglalkozik a megfelelő hangszín visszaállításával.

Drugman és társai a DSM modell [53] továbbfejlesztésével analízis-szintézis kísérletekben bemutatják, hogy a maradékjel periódusokban előforduló másodlagos impulzusok jelenléte megfelelően modellezi az irreguláris beszédet [71]. A módszer a szintetizált maradékjelet a beszélő első sajátmaradékjele alapján állítja elő: csak a zárt szakasz hosszát módosítja újramintavételezéssel a cél $F0$-nak megfelelően. Ezáltal a nyílt szakasz[4] nem változik és megmarad a hegyes jellege. Másolás-szintézis kísérletek és egy szubjektív teszt szerint ez a kiegészítés tovább javítja az alap DSM gerjesztési modellt. Drugman ezután megvizsgálja, hogy a HTS rendszer mely környezetfüggő címkéi lehetnek hasznosak a glottalizáció előfordulásának előrejelzésére és új paraméterfolyamokat is hozzáad a rendszerhez, amelyek segítik az automatikus döntést az irreguláris zönge helyéről [78]. Raitio és társai egyesítik a fenti módszereket és bemutatnak egy

[4] A nyílt és a zárt szakasz a zönge képzése során a gége két jól elkülöníthető állására utal.



1.6. ábra. Az alsó légúti (szubglottális) rendszer. Forrás: [80].

irreguláris zönge előrejelzésére és szintézisére alkalmas rendszert a DSM és GlottHMM modellek kiegészítéseként [79]. Eredményeik szerint a glottalizált minták használata kis mértékben érdesebbé tette a szintetizált beszédet, míg nem javította az alaprendszer természetességét.

1.5. Szubglottális rezonanciák hatása a beszédre

Beszédhangjaink akusztikai minőségét nem csak a gége és a felette lévő szervek határozzák meg, hanem a gége alatti (szubglottális) légzőszervek bizonyos tulajdonságai (pl. tüdő térfogata, légcső hossza) is befolyásolják azt. A korábban ismertetett forrás-szűrő modell [1] a forrás és szűrő közötti nemlineáris csatolást nem modellezi megfelelően [81]. A kutatások szerint a gége, a hangszalagok, a szubglottális tér és a szupraglottális tér ugyanis nemlineáris kölcsönhatásban állhat egymással [82]. A gége alatti tér, azaz alsó légúti rendszer szintén hozzájárul a beszédhangok alakításához, melynek felépítésére az 1.6. ábra mutat példát. A szubglottális rendszer rezonanciái (szubglottális rezonancia, *subglottal resonance*, SGR) pólus-zérus párokat alkotnak, amelyek a formánsokhoz hasonlóan alakítják a zöngés hangok spektrumát. A pólusok erősítik, a zérusok gyengítik a rezonanciafrekvencia körüli harmonikusokat. Mivel az alsó légúti szervek közül a légcső és a hörgők fiziológiai méretei viszonylag keveset változnak a be-

széd során, a rezonanciafrekvenciák közel állandóak egy-egy ember beszédében. Az első három szubglottális rezonancia tipikus értéke férfiak esetén 600, 1500 és 2300 Hz körül mérhető [6]. Női és gyermek beszélőknél az értékek valamivel magasabbak.

Az utóbbi években több nyelvre (amerikai angol [7], spanyol [83], német [84] és koreai [85]) megmutatták, hogy az alsó légutak rezonanciái a magánhangzókat és a mássalhangzókat a frekvenciaszerkezetük szerint diszkrét csoportokra bontják, melyek jellegzetes kategóriáknak feleltethetőek meg (fonológiai megkülönböztető jegyek, [6]). Ezen kategóriákat már számos elméleti megközelítés segítségével próbálták magyarázni, melyek közül az egyik legsikeresebb a kvantális elmélet (*Quantal Theory*, QT) [86].

1.5.1. Kvantális elmélet

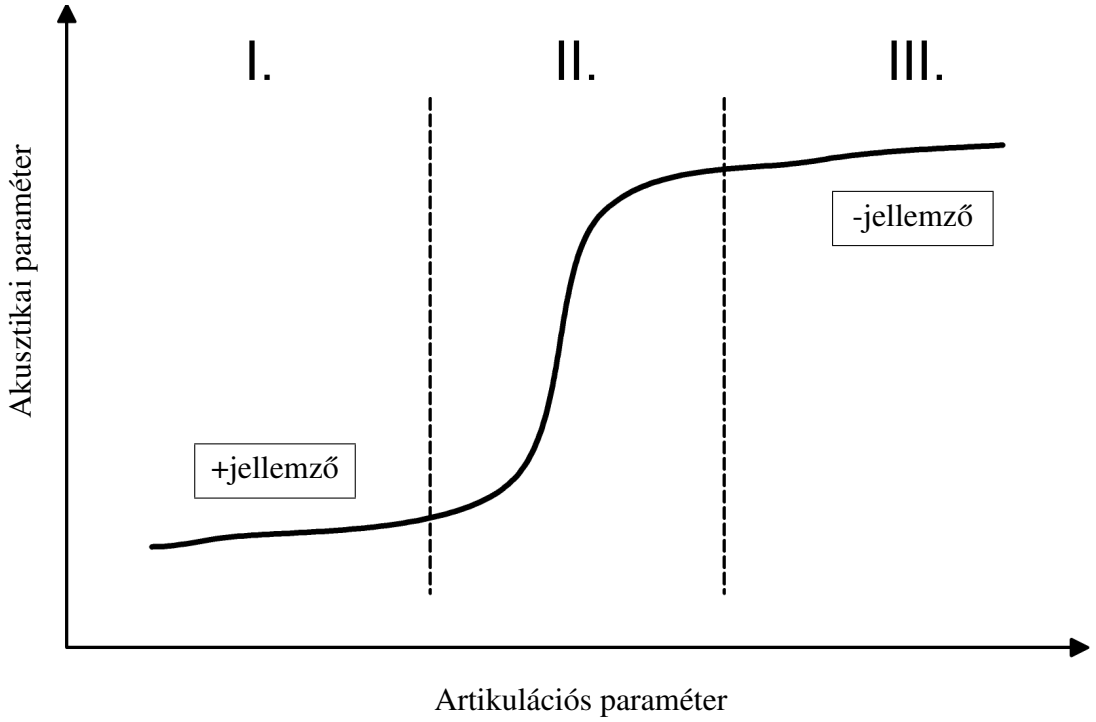
Stevens kvantális elmélete azon alapul, hogy a beszédhangokban mérhető akusztikai paraméterek és a beszélő által változtatott artikulációs helyzetek jellegzetes nem-monoton módon változnak, azaz az artikulációs tér egyes részeiben lévő kis változások nagy akusztikus változáshoz vezetnek, míg más, nagyobb artikulációs változtatások csak kisebb akusztikus változással járnak [6]. Az 1.7. ábrán látható módon a bináris jellemzőkhöz két diszkrét stabil állapot (I. és III.) tartozik, ahol a beszédkeltő rendszer akusztikai kimenete viszonylag érzéketlen az artikulációs paraméterek változására. Az átmeneti tartományban (II.) az akusztikai kimenet nagy mértékben változik az artikulációs mozgás hatására, és a feltételezések szerint az artikulációs szervek próbálják kerülni ezt a nem stabil állapotot. Eszerint a beszédhangokat kvantális jellemzőkkel lehet reprezentálni. Az egyik kvantális stabil állapothoz a [+jellemző] érték, a másik oldalhoz a [-jellemző] érték tartozik; a megkülönböztető jegy két értéke között pedig a határ rész húzódik.

A megkülönböztető jegyek egyik jó példája az elöl és hátul képzett magánhangzók esete: a két magánhangzó csoportot a nyelv vízszintes mozgása különbözteti meg. A [+/- hátul képzett] kvantális jegyhez ([+/-*back*]) tehát az I. stabil állapot tartozik, amikor a nyelv hátul van (magyarban [ɔ, o, oː, u, uː][5] magánhangzók); a III. állapot pedig az az eset, amikor a nyelv elöl van (magyarban [aː, ɛ, eː, i, iː, ø, øː, y, yː] magánhangzók). A II. átmeneti állapothoz nem köthető egyik magyar magánhangzó sem.

1.5.2. Szubglottális rezonanciák elemzése és alkalmazása

A kvantális elmélet szerint a magánhangzók artikulációja során van néhány olyan megkülönböztető jegy, amelyekre a toldalékcső és a szubglottális rendszer közötti akusztikai csatolás is hatással van. Ha egy formánsnak és egy szubglottális rezonanciának egymáshoz közeli a frekvenciája, akkor a formáns körüli spektrumot lényegesen módosíthatja az SGR jelenléte: gyak-

[5] A fonémák jelölésére IPA (*International Phonetic Alphabet*) ábrázolást használunk, `http://www.langsci.ucl.ac.uk/ipa/IPA_chart_(C)2005.pdf`



1.7. ábra. A kvantális elmélet szerinti nemlineáris kapcsolat az artikulációs és akusztikai paraméterek között. I. és III. régiók: stabil állapotok, II. régió: átmeneti tartomány. Az egyik kvantális stabil állapothoz a [+jellemző] érték, a másik oldalhoz a [-jellemző] érték tartozik. Forrás: [86] alapján, módosítva.

ran többszörös formáns-szerű csúcshoz vagy a formáns gyengüléséhez vezet, formánsmenetben történő megszakadás fordulhat elő, illetve ezek kombinációja is megjelenhet. Azt is kimutatták, hogy a szubglottális rezonanciák eltüntethetik a közeli spektrális csúcsokat, különös tekintettel a második formáns ($F2$) és a második szubglottális rezonancia ($Sg2$) esetére [6]. A SGR-ek frekvenciájának környezete tehát akusztikai szempontból előnytelen. Emiatt azt feltételezik, hogy a beszéd képzése során próbáljuk elkerülni azokat az artikulációs helyzeteket, amikor a formánsok és szubglottális rezonanciák között interakció léphetne fel. A formánsok próbálják elkerülni az SGR értékeket, ami magánhangzó csoportok elkülönüléséhez vezet. Az állítások szerint amerikai angolban az $Sg2$ egy természetes elválasztó az elöl képzett (*front*: [i, ɪ, ɛ, æ]) és hátul képzett (*back*: [ɑ, ʌ, o, ʊ, u]) magánhangzók között [6, 7, 87]. Az első szubglottális rezonancia ($Sg1$) hatása általában kevésbé erős, mint az $Sg2$-é, részben az alacsony frekvenciás akusztikai veszteségek miatt. Mégis azt vették észre, hogy az első formáns ($F1$) tekintetében az $Sg1$ elválasztó szerepet játszik az alsó állású (*low*: [æ, ɔ, ɑ]) és nem-alsó állású (*non-low*: [i, ɪ, ɛ, e, o, ʊ, u]) magánhangzók között [85]. Lulich azt találta, hogy amerikai angolban a harmadik szubglottális rezonancia ($Sg3$) sokszor az elöl képzett feszes (*tense*: [i, e]) és laza (*lax*: [ɪ, ɛ, æ]) magánhangzók között helyezkedik el [7]. A magánhangzók mellett a mássalhangzók közül a zárhangok képzési helye is kapcsolatban van az SGR-ek értékével [7]. Ezen állítások alapján a kvantális elmélet kiegészíthető a szubglottális rezonanciákra vonatkozó hipotézisekkel.

A szubglottális rezonanciák közvetlen mérése nehézkes lenne az invazív módszerek miatt, de közel pontos mérési eredményeket lehet elérni, amennyiben egy gyorsulásmérő eszközt szorítunk a nyakhoz, a gége előtti porcokhoz [87, 88]. Ezen eljárás során a mért jelben jelen van a szövetek csillapító hatása is, azonban a jel elegendően tiszta a szubglottális rezonanciák méréséhez [88]. A gyorsulásmérő által rögzített jelben (szubglottális jel) a spektrum burkoló csúcsaiként mérhetőek az $Sg1$, $Sg2$ és $Sg3$ értékek. A legtöbb vizsgálat során viszonylag kevés adaton végezték az elemzéseket, egyedül amerikai angol nyelvre áll rendelkezésere nagyobb beszédkorpusz, melyet 50 beszélővel rögzítettek [89]. Németben eddig két dialektus 12 beszélőjére végeztek vizsgálatot [84], koreaiban néhány felnőtt és 10 gyermek szubglottális rezonanciáit elemezték [85], valamint 20 kétnyelvű spanyol-angol gyermek beszédét és SGR-eit is tanulmányozták [83].

Az eddigi eredmények szerint a szubglottális rezonanciák a formánsmenetekben a folytonosság megszakadását okozhatják [88], észrevehetőek a beszédpercepció számára [90], valamint Wang és társai kutatásai szerint hasznosak lehetnek az automatikus beszélő normalizálásban [83, 91, 92]. Kezdeti kísérletekben Lulich és Chen bemutatták, hogy az $Sg2$ segítségével létre lehet hozni olyan automatikus osztályozó eljárást, amely mássalhangzó-magánhangzó kapcsolatokat tud az artikuláció szerinti kategóriákba sorolni [93, 94]. A szubglottális rezonanciák ismerete hasznos lehet beszédfelismerésben is azért, mert az SGR-ek közel konstansak [95]. Arsikere és társai néhány kutatása bemutatta, hogy korrelációs kapcsolat van a beszédjel bizonyos tulajdonságai és az SGR-ek között, így a szubglottális rezonanciák számíthatóak közvetlenül a beszéd mikrofonnal felvett jeléből is [95, 96, 97]. Emellett a szubglottális rezonanciák elemzése alapján lehetőség van a beszélő magasságának automatikus becslésére is [98, 99].

Az SGR-eket beszédszintézis környezetben eddig csak kezdeti kutatásokban vizsgálták. Gorbunov és Makarov artikulációs beszédszintetizátorban modellezi a szubglottális régiót: a korábbi modelleket kiegészítik a légcső, hörgők és tüdő modelljével [100]. Hiroya és társai bevezetnek egy módszert, amely a szubglottális rezonanciák hatását el tudja tüntetni a toldalékcső spektrum számítása közben, és megmutatják japán beszédszintézis mintákon, hogy az eljárás eredményes [101].

A fentiek szerint a szubglottális rezonanciák vizsgálata aktív kutatási terület, eddig azonban a magánhangzó formánsok és SGR-ek kapcsolatát csak néhány nyelvre vizsgálták. A szubglottális rezonanciák beszédhangokra kifejtett szerepével kapcsolatban magyar nyelvre korábban nem történt kutatás.

2. fejezet

Kutatási célkitűzések

Kutatásaimmal a rejtett Markov-modell alapú gépi szövegfelolvasók természetességének növeléséhez és a beszédképzés forrás-szűrő modelljének pontosításához kívánok hozzájárulni. Konkrét céljaim a kutatás során:

1) a statisztikai parametrikus beszédszintézisben a gépi beszéd természetességének növelése,

2) irreguláris zöngeképzés elemzése és ennek javítása, rekedtes beszéd hangzásának kellemesebbé tételére,

3) irreguláris beszédmodellek létrehozása beszédszintézisben, amelyekkel expresszív és személyre szabható gépi szövegfelolvasó rendszerek készíthetőek,

4) az emberi beszédképzésben a forrás-szűrő közti kölcsönhatás pontosabb megismerése, különös tekintettel a szubglottális rendszer hatására.

Ezeket a kutatási célokat azért választottam, mert számos kihívást tartalmaznak és a kutatásommal hozzá tudok járulni az ember-gép kapcsolat természetesebbé tételéhez. Munkám során a kísérleteket magyar nyelvű beszédkorpuszokon végeztem, de az eredmények nagy része könnyen alkalmazható más nyelvekre is, mert a 4. és az 5. fejezetek módszerei nem tartalmaznak nyelvfüggő elemeket. A 4. fejezetben az 1) és 2) kutatási célokkal, az 5. fejezetben az 1) és 3) célokkal foglalkozom, míg a 6. fejezetben a 4) kutatási célt teljesítem.

3. fejezet

Módszertan

Kutatásom során a létrehozott módszerek eredményességét kísérleti úton vizsgáltam. Ehhez nagyméretű beszédadatbázisokat használtam fel. A modelleket és módszereket szoftver eszközökkel valósítottam meg, majd az eredményeket meghallgatásos és akusztikai tesztekkel ellenőriztem.

3.1. Felhasznált beszédkorpuszok

A beszéd analízisével, szintézisével és az irreguláris zöngeképzéssel kapcsolatos kísérleteket (4. és 5. fejezetek) a PPBA adatbázisból kiválasztott 5 magyar anyanyelvű beszélőn végeztük [102]. Négy férfitól (FF1, FF2, FF3 és FF4) és egy nőtől (NO3) származó, professzionális körülmények között rögzített, 44,1 kHz-es mintavételezéssel 16 biten digitalizált hangfelvételt használtunk fel. Az adatbázis beszélőnként közel ugyanazt az 1940 mondatot tartalmazza, amely nagyjából 2 órányi tiszta hangfeltételt jelent (pontos adatok a 3.1. táblázatban találhatóak). Az adatbázishoz szöveges címkézés, automatikus fonetikus átirat és ennek kézzel javított változata, valamint hanghatárjelölés is tartozik. A precíziós annotálás eredményeként az adatbázis megbízható, 99,9%-os: a hanghatárok pontossága 10 ms-on belül van és a fonetikai átirat pontosan megfelel a hanghullámnak.

3.1. táblázat. A PPBA adatbázisból az elemzésekhez kiválasztott beszélők hanganyagának adatai.

Beszélő	Mondatok száma	Időtartam
FF1	1936	190 perc
FF2	1938	137 perc
FF3	1941	170 perc
FF4	1938	214 perc
NO1	1937	128 perc

A beszéd analízisével és szintézisével kapcsolatos módszereket magyar mintákon teszteltük és validáltuk, de az itt alkalmazott eljárások nyelvfüggetlenek és várhatóan más nyelvre is hasonló módon alkalmazhatóak.

3.2. Felvételi körülmények

A szubglottális rezonanciák vizsgálatához (6. fejezet) a beszéd és szubglottális felvételek egy részét a kutatás során rögzítettük magyar anyanyelvű beszélőkkel. Részben 4 beszélő logatom felvételein [C4], részben a BEA adatbázis [103] 6 beszélőjétől származó spontán beszéd felvételeken és ugyanezen beszélők olvasott beszéd felvételein [J4] végeztük az elemzéseket.

Az első szubglottális rezonanciákat elemző kísérlethez logatom felvételek során az akusztikai adatokat két magyar anyanyelvű férfitól és két nőtől rögzítettük (életkor: 22–38 év, jelölés: Log_FF1, Log_FF2, Log_NO1, Log_NO2). A beszélők /ɔCVCɔ/ típusú logatomokat olvastak fel, amiben a vizsgálandó magánhangzó két zárhang között szerepelt (C: mássalhangzó, V: magánhangzó). Az első mássalhangzó [b, d, g] volt; a második mássalhangzó férfiak esetén fixen [b], a nők esetén fixen [d] volt. A cél magánhangzó a második (hangsúlytalan) szótagba került. A logatomokban minden magyar magánhangzó ([ɔ, aː, o, oː, u, uː, ɛ, eː, i, iː, ø, øː, y, yː]) szerepelt. A női beszélők minden logatomot ötször (15 minta magánhangzónként), a férfiak háromszor ismételtek (9 minta magánhangzónként). A felvételeket csendes szobában végeztük, Monacor EMC 100 kondenzátor mikrofonnal. A szubglottális jelet K&K HotSpot típusú gyorsulásmérő eszközzel rögzítettük, melyet a nyakon a gégénél lévő pajzsporchoz szorítottunk. A beszéd és a gyorsulásmérő jelet is 8 kHz-es mintavételezéssel digitalizáltuk két külön csatornán Terratec DMX 6 Fire USB külső hangkártyával, a Wavesurfer programmal. Az itt használt felvételekhez nem állt rendelkezésre címkézés; a szöveges és fonetikus átírást valamint a hanghatárok címkézését a kutatás során készítettük el automatikus eszközökkel és manuális javítással. A hanghatárok automatikus meghatározásához a MAUS kényszerített felismerő programot használtuk fel[1].

A második szubglottális rezonanciákat elemző kísérlethez felhasznált akusztikai adatok egyrészt hat magyar anyanyelvű beszélő spontán beszéd anyagából [103], másrészt ugyanezen beszélők gyorsulásmérő felvételeiből álltak (5 férfi és 1 nő, életkor: 25–35 év, jelölés: Spo_FF1 – Spo_FF5, Spo_NO1). A BEA spontán beszéd adatbázis vizsgált részében a beszélők kötetlen témában beszélgettek csendesített szobában az interjúztatóval 3-10 percen keresztül, amelyből csak a cél beszélő hanganyagát használtuk fel. A szöveges átírás után a fonetikus átírást és a hanghatárok bejelölését a BME-TMIT kényszerített felismerő programjával végeztük el, majd a hanghatárokat manuálisan javítottunk[2].

[1] A többes szám a kutatásban részt vevő többi személyre utal: Bárkányi Zsuzsanna, Gráczi Tekla Etelka, Bőhm Tamás és Steven M. Lulich. A felvételek készítését, a manuális méréseket és a kézi javításokat közösen végeztük.

[2] A többes szám a kutatásban részt vevő többi személyre utal: Gráczi Tekla Etelka, Bárkányi Zsuzsanna, Beke András és Steven M. Lulich. A manuális méréseket és a kézi javításokat közösen végeztük.

3.3. Alkalmazott eszközök és szoftverek

Kutatásaim során a következő eszközöket és szoftvereket használtam fel:

BME-TMIT kényszerített felismerő: hanghatárok automatikus címkézése [104],

GLOAT / SEDREAMS: beszédjel felbontása zöngeszinkron periódusokra [105],
`http://tcts.fpms.ac.be/~drugman/Toolbox/`

HTS: paraméterek tanítása HMM-ek segítségével [2],
`http://hts.sp.nitech.ac.jp/`

HTS-HUN: a HTS rendszer magyar változata [15],

Matlab: beszédjel analízise és szintézise, ROC elemzés, t-teszt,
`http://www.mathworks.com/products/matlab/`

MAUS: kényszerített felismerő, hanghatárok automatikus címkézése,
`http://www.phonetik.uni-muenchen.de/forschung/Verbmobil/VM14.7eng.html`

Praat: alapfrekvencia mérése; formánsok mérése; beszédjel vizuális elemzése [106],
`http://www.fon.hum.uva.nl/praat/`

Snack / getF0: alapfrekvencia mérése a HTS rendszerben,
`http://www.speech.kth.se/snack/`

SoX: beszédjel aluláteresztő szűrése és újramintavételezése,
`http://sox.sourceforge.net/`

SPSS: ANOVA analízis,
`http://www.ibm.com/software/hu/analytics/spss/`

SPTK: spektrális elemzés, inverz szűrés és digitális szűrés,
`http://sp-tk.sourceforge.net/`

Voice_Analysis_Toolkit / creak_detect: irreguláris zönge detektor [67],
`https://github.com/jckane/Voice_Analysis_Toolkit`

VoiceSauce: beszédjel akusztikai paramétereinek korrekciója; HNR számítása,
`http://www.ee.ucla.edu/~spapl/voicesauce/`

Wavesurfer: beszédjel és gyorsulásmérő jel vizuális elemzése és akusztikai mérések [107],
`http://www.speech.kth.se/wavesurfer/`

Weka: döntési fák megvalósítása [108],
`http://www.cs.waikato.ac.nz/ml/weka/`.

3.4. Meghallgatásos tesztek

A transzformációs eljárások és szintézis módszerek eredményességét percepciós (meghallgatásos) kísérletekkel is vizsgáltam. A beszédszintézis területén általánosan elterjedt a módszerek eredményének meghallgatásos teszt alapú értékelése. A kísérletekben többféle típusú tesztet szoktak alkalmazni, melyek közül saját vizsgálataim során a következőket használtam: a tesztelők az egyes hangminták meghallgatása után 1-5 skálás MOS (*Mean Opinion Score*), illetve minta párok esetén 1-3 vagy 1-5 skálás CMOS (*Comparative Mean Opinion Score*) jellegű kérdésekre válaszolnak.

A kísérletek készítése során a szakirodalomban javasolt teszt típusokból indultam ki [109]. A tesztek elején a kísérleti alanyok egy ismertetőt olvashattak az aktuális kísérlet témájáról és menetéről, majd néhány adat (nem, kor, eszköz, beszédtechnológiai ismeretek) megadását kértük tőlük. A tesztek internet alapúak voltak, melyeknek során a mintákat vagy mintapárokat a megadott szempontok és kérdések szerint értékelték. A hangmintákat vagy minta párokat minden tesztelő más-más sorrendben hallgatta meg; a párok esetén a két változat is véletlen sorrendben szerepelt.

Törekedtem arra, hogy a kísérletek felépítése hasonló legyen, a felhasznált hanganyag és a feltett kérdések azonban tesztenként eltérőek. A kísérleti személyekkel kapcsolatban egy összesítés látható a 3.2. táblázatban. Az egyes percepciós kísérletek körülményei és részletei a későbbi fejezetekben olvashatóak.

3.2. táblázat. A meghallgatásos tesztek összesített tesztelői adatai.
Rövidítések: FH = Fejhallgató, HSZ = Hangszóró, BK = Beszédkutató, E = Egyéb.

Fejezet	**Tesztelők**										
	Összesen	**Nem**		**Eszköz**		**Tesztelői kör**		**Életkor (év)**		**Időtartam (perc)**	
		Férfi	**Nő**	**FH**	**HSZ**	**BK**	**E**	**Átlag**	**Szórás**	**Átlag**	**Szórás**
4.2.2.	9	9	0	7	2	3	6	23,67	3,20	6,92	1,39
5.1.3.	15	12	3	5	10	0	15	32,00	9,02	5,08	1,47
5.2.2.	11	9	2	10	1	0	11	23,81	4,31	9,03	2,09
5.2.4.	17	13	4	13	4	7	10	31,76	11,15	17,11	7,01

3.5. Szignifikancia vizsgálatok

A statisztikai elemzések során egymintás t-tesztet, párosított mintás t-tesztet és Tukey-HSD post-hoc teszttel kiegészített egytényezős ANOVA analízist alkalmaztam a Matlab és SPSS programokkal. Az elemzések során kétoldalas $p < 0{,}05$ szignifikancia szint alatt (95% konfidencia szint felett) vetem el a nullhipotézist.

4. fejezet

Újszerű gerjesztési modell kidolgozása

A szakirodalomban számos beszéd analízis-szintézis módszerről olvashatunk, melyeknek célja eredetileg a beszéd paraméterekre bontása és kódolása volt azért, hogy a távközlési csatornán minél kisebb sávszélesség mellett lehessen átvinni jól érthető beszédet (ld. 1.3.1. fejezet). Emellett napjainkban a beszédfeldolgozás területén egyre fontosabb, hogy a beszédjel olyan parametrikus felbontását találjuk meg, amely különböző transzformációkra alkalmazható és gépi tanuló rendszerben is felhasználható. Kezdeti kísérleteink[1] szerint a ma elérhető legjobb beszédkódoló eljárások (pl. CELP, Code-Excited Linear Prediction jellegű kódolók) nem alkalmasak a gépi tanulórendszerbe történő integrálásra (pl. a CELP kódoló kódkönyv indexe ugráló értékeket tartalmaz, ami nem modellezhető egyszerűen HMM-ekkel). Az 1.3. fejezetben ismertetett gerjesztési modellek közül az egyszerűbbek (pl. impulzus-zaj modell) zizegős beszédet eredményeznek. A bonyolultabbakkal (pl. kevert gerjesztés) ugyan jobb minőségű beszéd szintetizálható, de sok esetben nehezen használhatóak fel valós idejű alkalmazásokban nagy számításigényük miatt. A két véglet között olyan gerjesztési modell elkészítését céloztuk meg, melynek minősége megfelelő, és várhatóan használható korlátozott erőforrású eszközben is.

A fejezet bemutat egy újszerű, beszédet paraméterekre bontó, maradékjelen alapuló, nyelvfüggetlen gerjesztési modellt, amely beszéd analízis-szintézisére alkalmas és a paraméterei integrálhatóak a rejtett Markov-modell alapú gépi tanításba. A korábbi eljárások közül vannak ehhez hasonló gerjesztési modellek. A DSM eljárás is maradékjel kódkönyv alapú, azonban ez nem alkalmaz összefűzési költséget az elemkiválasztás során [52]. A GlottHMM rendszerben alkalmaznak ugyan célköltséget és összefűzési költséget is, de ez glottális forrásjel szintjén történik [39]. Ez alapján az itt javasolt modell lényeges pontokban különbözik az ismert korábbi rendszerektől. Emellett új típusú, korábban nem használt paramétereket vezetünk be a mara-

[1] A továbbiakban többes szám első személyt használok a könnyebb olvashatóság érdekében. Saját eredményeimet a 7. fejezetben összegzem.

dékjel leírására. A további fejezetekben ismertetjük a modell alkalmazását arra a célra, hogy az irreguláris zöngével képzett természetes beszéd érzeti érdességét egy transzformációval csökkentsük és a reguláris zöngéjű beszédhez hasonlóvá tegyük.

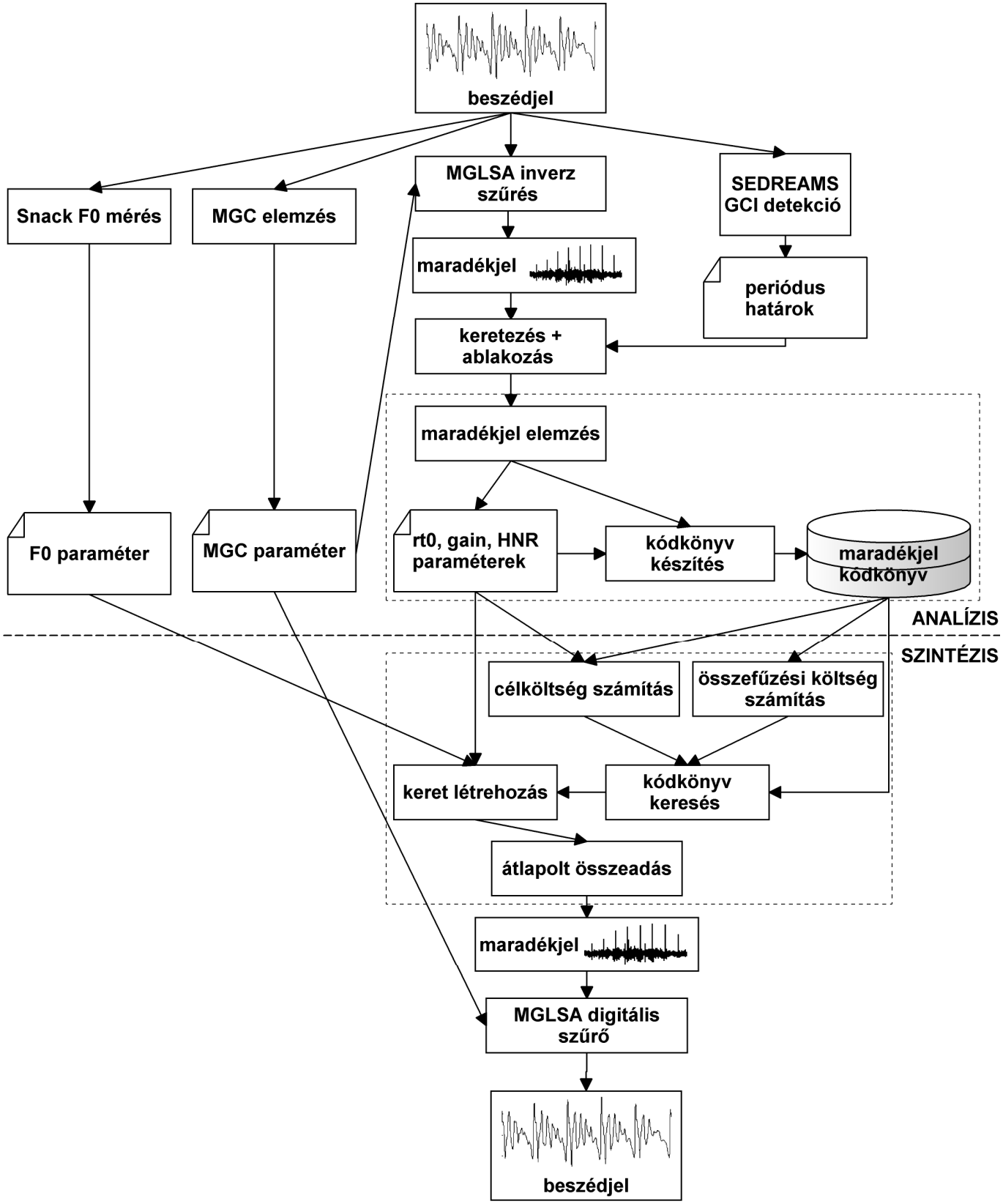
4.1. Új, MGC maradékjel kódkönyv alapú gerjesztési modell kidolgozása

Az irodalmi áttekintésen belül az 1.3. fejezetben ismertettük azokat a gerjesztési modelleket, amelyek alkalmasak a beszéd analízis-szintézis felbontására. A következő részben bemutatjuk egy új gerjesztési modell kidolgozását.

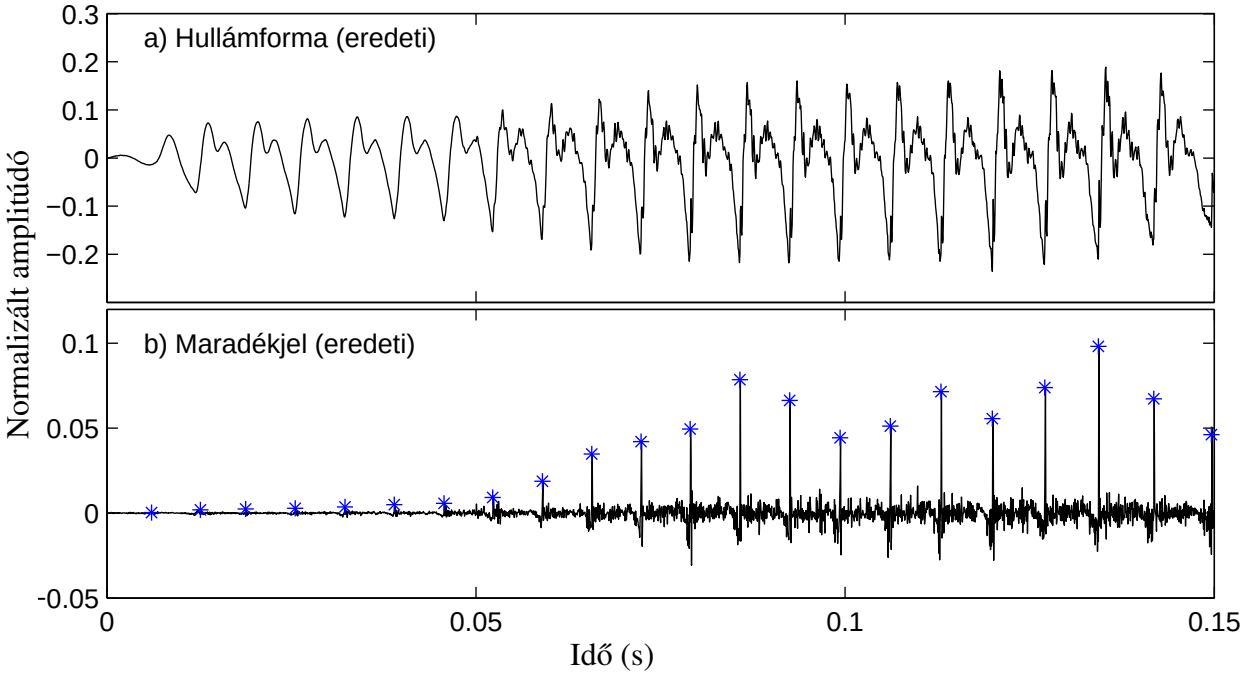
4.1.1. Analízis

Az analízis lépéseit a 4.1. ábra szaggatott vonal feletti része mutatja. Az általunk hozzáadott új eljárások a szaggatott vonalú téglalapon belül láthatóak. Az analízis módszer bemenete beszéd hullámforma, amelyet 7,6 kHz-es aluláteresztő szűrés után 16 kHz mintavételezéssel és 16 bites lineáris PCM kvantálással tárolunk. A módszer először egy zöngeszinkron maradékjel periódusokból álló kódkönyvet épít, majd elvégzi a maradékjel elemzését. A beszéd alapfrekvenciáját 25 ms kerethosszal és 5 ms eltolással mérjük a Snack RAPT $F0$-detektáló algoritmusával [110]. Következő lépésben spektrális elemzést végzünk MGC (*Mel-Generalized Cepstrum*, magyarul *Mel-Általánosított Kepsztrum*) módszerrel [111]. Az elemzéshez 34-ed rendű MGC analízist alkalmazunk $\alpha = 0{,}42$ és $\gamma = -1/3$ paraméterekkel. A maradékjelet, vagyis a beszéd gerjesztését MGLSA (*Mel-Generalized Log Spectral Approximation*) inverz szűréssel számoljuk ugyanazon α, γ és dimenzió paraméterrel [112]. Ezután az SEDREAMS (*Speech Event Detection using the Residual Excitation And a Mean-based Signal*) zöngeperiódus-meghatározó algoritmust alkalmazzuk a zöngés maradékjel periódusainak szétválasztásához [105]. A 4.2. ábra egy példát mutat zöngés beszédszakaszra (a) és maradékjelére (b). A b) ábrán csillagok jelzik a zöngeperiódusok meghatározott helyét a GCI (*Glottal Closure Instant*) pozícióknak megfelelően. A GCI, vagyis a hangszalagok záródásának pillanata a maradékjel periódusokban a legnagyobb amplitúdójú, impulzus-szerű értékkel esik egybe. Azért választottuk az SEDREAMS algoritmust erre a feladatra, mert egy friss elemzés szerint az elérhető GCI számító módszerek közül ez eredményezi a legmagasabb találati arányt és legalacsonyabb téves riasztást, valamint robusztus a hozzáadott zajra és visszhangra [113].

Az analízis további lépéseit a maradékjelen végezzük el 50 ms keretméret és 5 ms eltolás értékekkel. A hosszabb keretméret biztosítja, hogy alacsony $F0$ esetén is található legalább két periódus a keretben. A zöngés szakaszokból zöngeszinkron, két periódus hosszú, Hann-ablakozott maradékjel periódusokat vágunk ki, melyekből egy kódkönyv készül. A kódkönyv elemek leírására a következő paramétereket használjuk:



4.1. ábra. Beszédjel analízise (szaggatott vonal felett) és szintézise (szaggatott vonal alatt) az MGC maradékjel kódkönyv alapú módszerrel. Négyzetek jelölik az eljárásokat és hullámformákat; a behajtott sarkú négyzetek a paramétereket jelzik. A szaggatott vonalú téglalapok mutatják az általunk hozzáadott új eljárásokat.



4.2. ábra. Példa a beszédjelből számított maradékjelre és a meghatározott periódusokra egy zöngés szakaszon: a) beszéd hullámforma b) maradékjel. A b) ábrán a maradékjel kiugró értékei a GCI helyek. A csillagok az SEDREAMS algoritmussal meghatározott periódusok időhatárait jelölik.

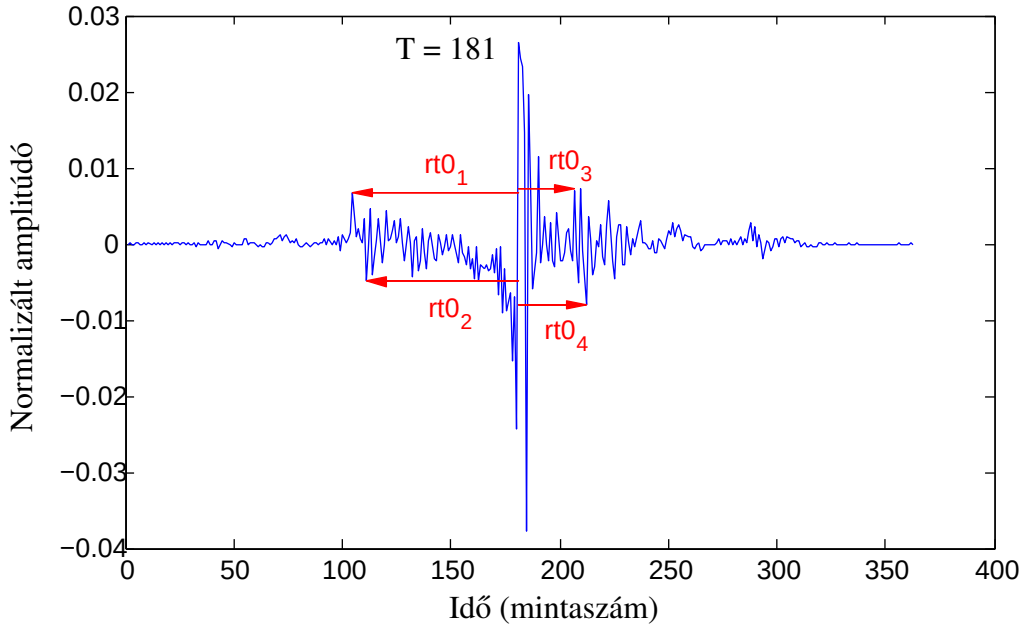
F0: az elem alapfrekvenciája,

gain: az elem energiája:
$gain_i = \sqrt{\sum_{j=0}^{N} r_j^2}$, ahol r_j az i. ablakozott elem j. mintája,

rt0: az ablakozott elemben a kiugró csúcsok pozíciója (példa: 4.3. ábra),

HNR: az elem harmonikus-zaj aránya (HNR, *Harmonics-to-Noise Ratio*) [114].

Minden zöngés kerethez eltárolunk egy kódkönyv elemet az ablakozott jellel és a fenti paraméterekkel együtt. A $gain$ paraméter az ablakozott elem RMS (*Root Mean Square*) energiája. Az $rt0$ paraméter egy négy dimenziós vektor, amelynek célja az ablakozott maradékjel kódkönyv elemben lévő jelentős csúcsok leírása. A 4.3. ábra mutat példát az $rt0$ paraméter értékeinek számítására. A középső ($T = 181$ minta) impulzustól mérjük a többi csúcs előjel nélküli távolságát, azonban az impulzus közelében lévő, a periódus hosszának 10%-án belüli jelentős csúcsokat nem vesszük figyelembe. Ennek az az oka, hogy az elemzéseink szerint a közeli csúcsok használata nem segíti a gépi tanuló rendszerbe való integrálást. Korábban ilyen paramétert használó megoldást egyik módszer sem alkalmazott a maradékjel leírására. A HNR paraméter a keret harmonikus és zaj komponenseinek arányát méri, melynek számítását kepsztrális harmonikus alapon végezzük [114]. A 4.4. ábra megmutatja a fenti paraméterek keretenkénti értékeit



4.3. ábra. Az $rt0$ paraméter számítása egy ablakozott maradékjel kódkönyv elemre. Az $rt0_i$ érték a kiugró csúcsok mintában mért távolságát adja meg az elemben lévő impulzushoz ($T = 181$) képest. Az ábrán lévő értékek: $rt0_3 < rt0_4 < rt0_2 < rt0_1$.

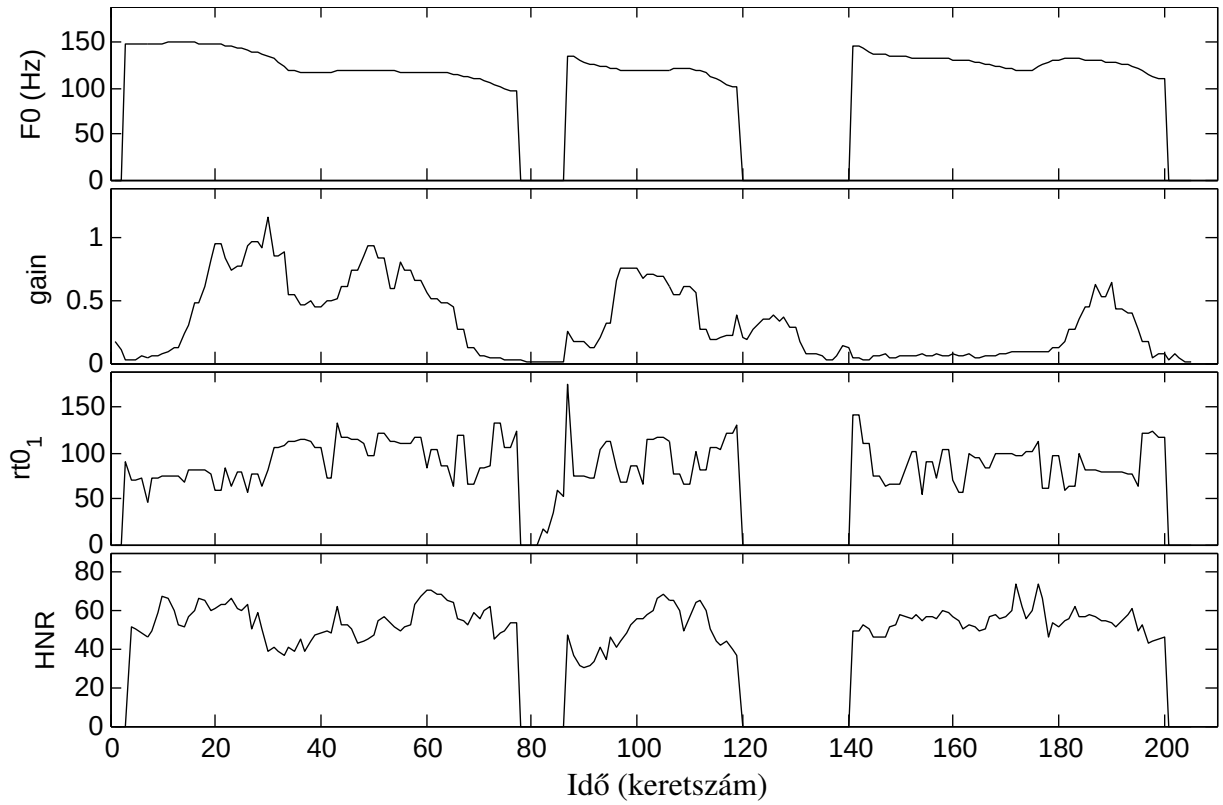
egy hosszabb analizált beszédmintán, a *„Néhány perc múlva”* beszédrészlet alapján. Az $F0$ paraméter a zöngés szakaszokon értelmezett, a zöngétlen helyeken 0 értékű. A $gain$ paraméter az egyes keretek energiáját adja meg, így a nagyobb intentizású beszédhangokban magasabb értékű. Az $rt0_1$ a maradékjelben lévő csúcsokat, míg a HNR a maradékjel zöngés és zöngétlen részeinek arányát mutatja.

A maradékjel kódkönyv készítése során a hasonló, egymáshoz várhatóan illeszkedő elemeket összefűzési költség felhasználásával számítjuk. Ehhez az ablakozott maradékjel elemeket $F0$ szerint normalizáljuk, azaz újramintavételezzük 40 mintára (16 kHz mintavételezés mellett 2,5 ms). Az ablakozott, normalizált maradékjel keretek között RMSE (*Root Mean Squared Error*) távolságot számítunk, ami megadja az egymáshoz való hasonlóságukat. Az összefűzési költséget a szintézis során használjuk fel az elemek összeillesztésekor.

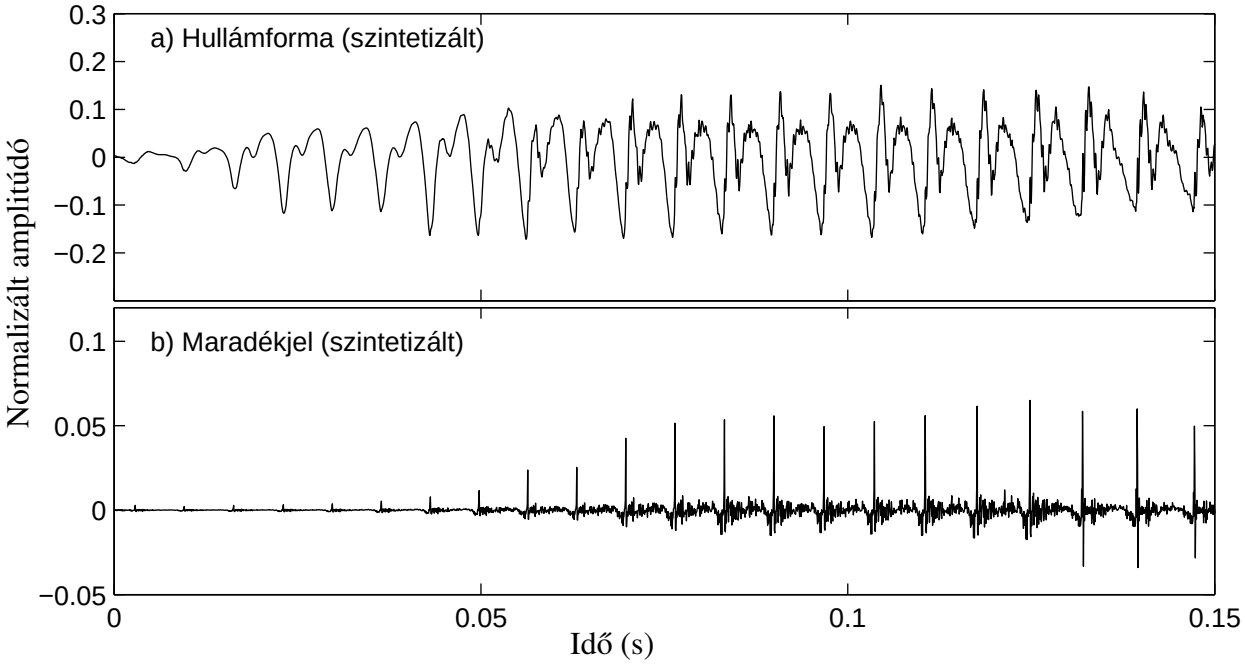
A beszédjel analízise során a fenti paramétereket kinyerjük minden zöngés keretből (azaz ha $F0 > 0$). Zöngétlen keret esetén ($F0 = 0$) csak a $gain$ értéket számoljuk.

4.1.2. Szintézis

A szintézis lépéseit a 4.1. ábra szaggatott vonal alatti része mutatja be. Az általunk hozzáadott új eljárások a szaggatott vonalú téglalapon belül láthatóak. A szintézis bemenete az analízis eredményeként kapott paraméterek ($F0$, $gain$, $rt0$, HNR és MGC) illetve a zöngeszinkron maradékjelek kódkönyve. A visszaállítás során először a maradékjelet állítjuk elő keretenként. Amennyiben a keret zöngés ($F0 > 0$), az $F0$, $rt0$ és HNR paraméterek alapján egy megfelelő, hozzá tartozó elemet keresünk a kódkönyvből. Kézzel beállított súlyozású célköl-



4.4. ábra. Példa az analízis során kinyert paraméter értékekre egy hosszabb beszédmintán: FF3 beszélő *„Néhány perc múlva”* beszédrészlete alapján.



4.5. ábra. Példa a szintetizált beszédjelre és az összefűzött maradékjelre a 4.2. ábra beszédmintáján: a) szintetizált beszéd hullámforma b) szintetizált maradékjel.

séget és összefűzési költséget alkalmazunk az elemkiválasztásos beszédszintézishez hasonlóan [4]. A célköltség az aktuális keret paraméterei és a kódkönyv elemeinek paraméterei közötti négyzetes különbség. Az összefűzési költséget a kódkönyv elemek normalizált változatának átlagos négyzetes különbségeként (RMSE távolság) számítjuk. A legmegfelelőbb kódkönyv elem hosszát a cél $F0$-nak megfelelően beállítjuk törléssel vagy nullák hozzáadásával. Amennyiben a keret zöngétlen ($F0 = 0$), fehér zajt használunk gerjesztésként. Ezután a maradékjelet a Hann-ablakozott periódusok zöngeszinkron átlapolt összeadásával és a zöngétlen részek összefűzésével kapjuk. Az átlapolt összeadást a PSOLA eljáráshoz hasonlóan végezzük [76]. Végül a keretek energiáját a $gain$ paraméter alapján beállítjuk, majd a szintetizált beszédet előállítjuk MGLSA szűréssel az MGC paramétereket felhasználva. A 4.5. ábrán látható a 4.2. ábra analízis-szintézis elemzéseként kapott maradékjele és visszaalakított beszéd hullámformája, amelyek az eredeti jelekhez hasonlóak.

A maradékjel kódkönyv méretének optimális meghatározására előzetes megvalósíthatósági kísérletet végeztünk. A cél az volt, hogy egy viszonylag nagy méretű kódkönyvből kiindulva megtaláljuk azt a legkisebb méretet, amely mellett a visszaállított beszéd minősége nem romlik érezhetően. Ehhez először kb. 30 000 elemből álló kódkönyvet készítettünk és alkalmaztunk, majd fokozatosan csökkentettük a méretét egészen 100 elemig. 10 mondatot elemeztünk a fenti analízis-szintézis módszerrel és a kódkönyvekkel visszaállított beszédminták informális meghallgatása során arra az észrevételre jutottunk, hogy mintegy 6 500 elemű kódkönyv mellett (amely kb. 20 perc beszéd alapján készült) még ugyanolyan a kódolt-dekódolt beszéd minősége, mint a legnagyobb méretű kódkönyvvel. Raitio és társai hasonló eredményre jutottak: egyik kutatásuk során a GlottHMM rendszer glottális forrásjel elemtárának optimális méretét vizsgálták, amelynek eredménye szerint a kb. 7 500 elem nagyságrendű elemtárral még mindig megfelelő a szintetizált beszéd minősége [38, 39].

Az analízis-szintézis eljárás önmagában alkalmas a beszédjel paraméterekre bontására és abból történő visszaállítására, vagyis ez egy beszédkódoló algoritmus. Emellett fontos kiemelni, hogy a paramétereket módosítva lehetőség nyílik a beszéd tulajdonságainak módosítására is. Az $F0$ paraméter növelésével illetve csökkentésével a beszéd dallamát lehetne változtatni, a $gain$ paraméter skálázásával pedig az egyes beszédhangok vagy azon belüli szakaszok relatív intenzitását. Ezt a tulajdonságot kihasználva a 4.2. fejezetben egy beszéd transzformációs eljárást dolgozunk ki. Ezen kívül a paraméterekkel reprezentált beszéd alkalmas a statisztikai parametrikus beszédszintézisben történő felhasználásra, amit az 5. fejezetben ismertetünk.

4.2. Az új gerjesztési modell felhasználása irreguláris zönge-képzés javítására

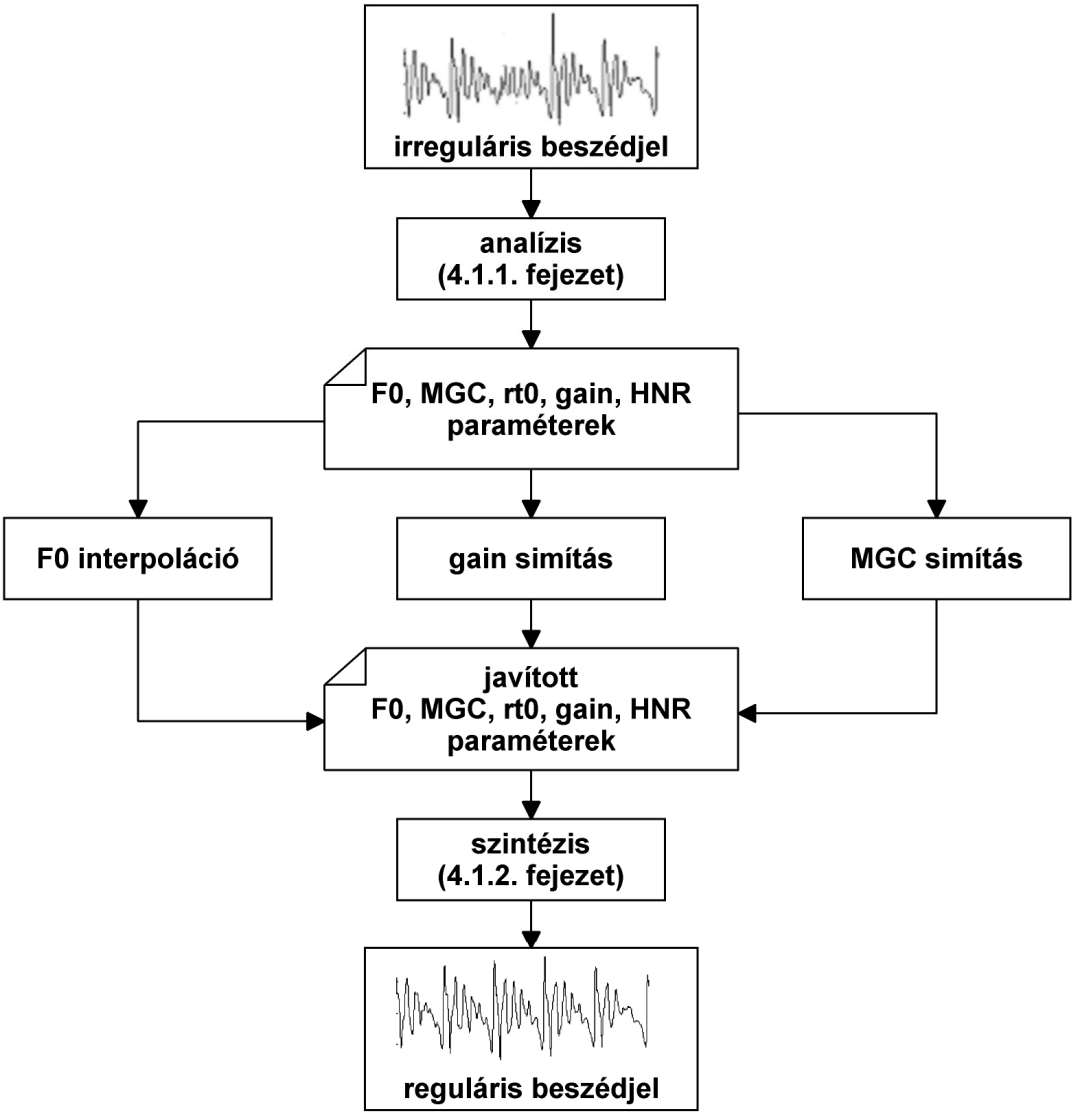
Az 1.4. fejezetben bemutattuk, hogy az irreguláris zöngeképzés kutatása során részletesen foglalkoztak a glottalizáció detekciójával, reguláris-irreguláris transzformációval, és kezdeti kísérletek történtek érdes zöngével kiegészített beszédszintézisre; azonban tudomásunk szerint eddig nem foglalkoztak részletesen az irreguláris-reguláris beszéd transzformációjával.

Az itt kidolgozott eljárás a 4.1. fejezet analízis-szintézis módszerét egészíti ki egy olyan transzformációs eljárássá, amely alkalmas a glottalizált beszéd modálissá alakítására, tehát az irreguláris zöngeképzés javítására. Az analízis hasonlóan történik, mint a fenti analízis-szintézis gerjesztési modellben, azzal a különbséggel, hogy a kódkönyvet csak modális maradékjel szakaszokból építjük, az irreguláris zöngével képzett részeket kihagyva. Az analízis után a paramétereket módosítjuk, majd a 4.1. fejezet szintézisével visszaállítjuk a javított beszédjelet.

4.2.1. Transzformáció

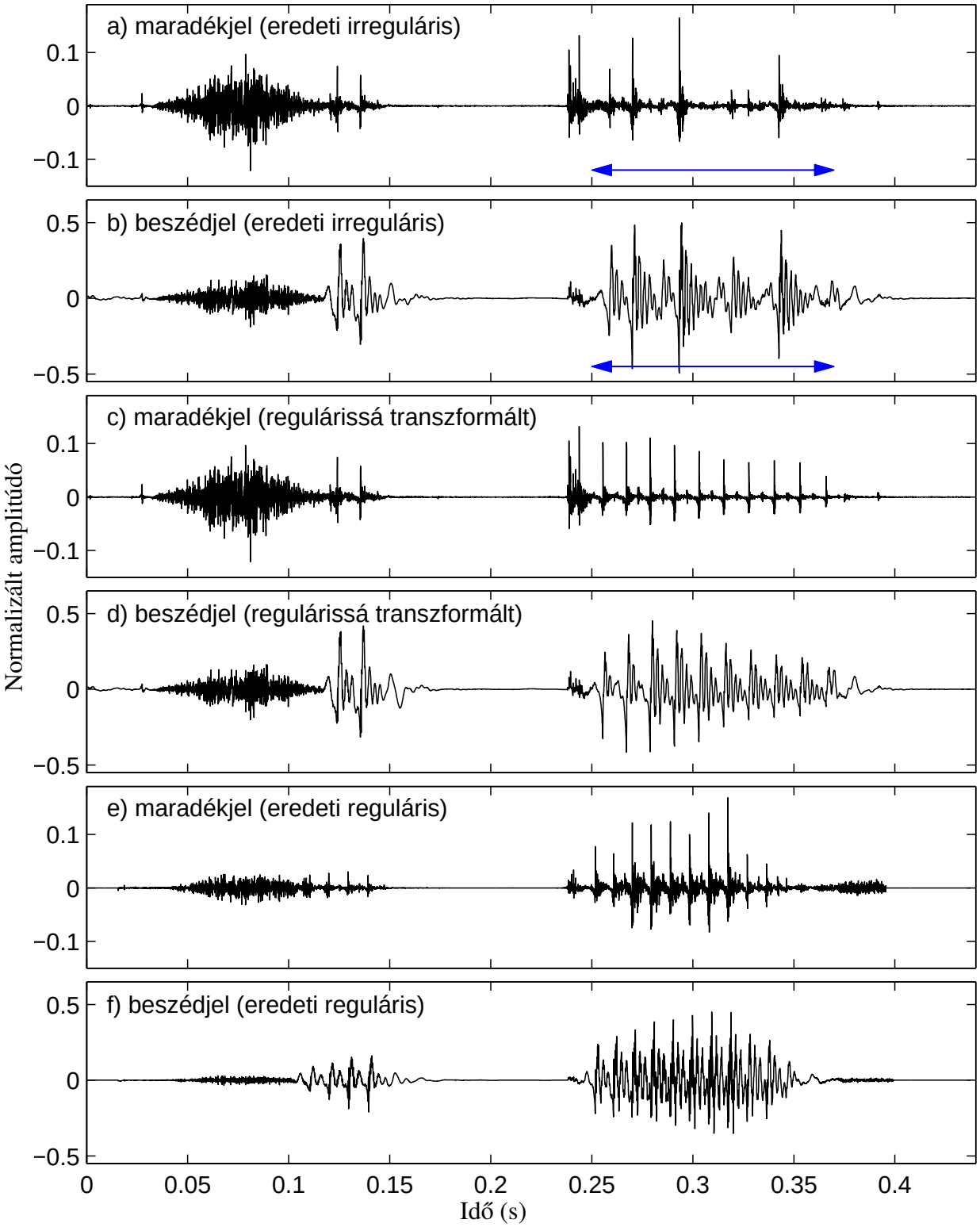
A transzformáció során az eredeti beszédből származtatott maradékjelnek azon szakaszait vizsgáljuk, amelyet irreguláris zönge címkék jeleznek, míg a modális zöngés és zöngétlen maradékjel részeket változatlanul hagyjuk. A transzformációs eljárás működését a 4.6. ábra ismerteti.

A 4.1. fejezetben bemutatott analízis eredményeként kapott $F0$ értékeket interpoláljuk, míg a $gain$ és MGC értékeket simítjuk az irreguláris szakaszokon. A glottalizáció hibákat okozhat az $F0$ detekcióban: a hirtelen alapfrekvencia és amplitúdó változás miatt (ld. 4.7. a és b ábra) előfordulhat, hogy egy eredetileg zöngés keretet zöngétlennek jelöl a detektor, vagy az eredeti érték felét méri. Emiatt a mért $F0$-menetet interpoláljuk azokban a zöngés szakaszokban, ahol az algoritmus nem detektált zöngét. Ezen szakaszokat a beszédmintához tartozó hanghatár jelölés alapján határozzuk meg. Ha egy magánhangzóban egyáltalán nem mért $F0$ értéket a zönge detektor, akkor a mondat dallamának megfelelő ereszkedő $F0$ menetet állítunk be. A kísérletek során minden $F0$-menetet kézzel ellenőriztünk és javítottunk, emiatt a módszer félautomatikus működésű. Az irreguláris fonáció kis perturbációkat okoz a keretenkénti $gain$ és MGC értékekben az irreguláris zöngeperiódusok amplitúdójának hirtelen változása miatt. Emiatt 5-pontos simítást végeztünk ezeken a paramétereken, amely tapasztalataink szerint megfelelőnek bizonyult a perturbációk eltüntetésére. A spektrum MGC reprezentációja alkalmas ilyen interpolációra és nem eredményez instabilitást. A szintézis további lépései megegyeznek a 4.1. fejezetben ismertetett lépésekkel, azaz a paramétereknek megfelelő maradékjel elemeket keresünk a kódkönyvből a célköltség és összefűzési költség felhasználásával, majd ezeket átlapolt összeadással összefűzzük. A zöngés és zöngétlen részeket egyesítve az energia megfelelő beállítása után spektrális szűréssel kapjuk meg a transzformáció kimeneti beszédjelét.



4.6. ábra. Az MGC maradékjel kódkönyv alapú gerjesztési modellt felhasználó irreguláris-reguláris transzformáció működése.

Az irreguláris-reguláris transzformáció eredményére láthatunk egy példát a 4.7. ábrán. Az ábrán észrevehető, hogy a „regulárissá transzformált” (c és d) és az „eredeti reguláris” (e és f) változatoknak hasonló zöngeperiódusai vannak, míg az „eredeti irreguláris” (a és b) jel ettől lényegesen eltérő és periódusonkénti amplitúdó ingadozást tartalmaz. A c-e illetve d-f ábrák közti különbségek azért fordulnak elő, mert ez a szó két különböző realizációja, így kis eltérések láthatóak az egyes beszédhangok időtartamában és átlagos amplitúdójában.



4.7. ábra. A kiejtett és transzformált „cipő” szó hullámformái és maradékjelei FF3 beszélőtől:
a) maradékjel és b) beszédjel eredeti irreguláris záró magánhangzóval
(nyíl jelöli az irreguláris zöngét),
c) maradékjel és d) beszédjel regulárissá transzformált záró magánhangzóval,
e) maradékjel és f) beszédjel eredeti reguláris záró magánhangzóval (a szó másik realizációja).

4.2.2. Meghallgatásos teszt

Hanganyag és módszer, kísérleti személyek

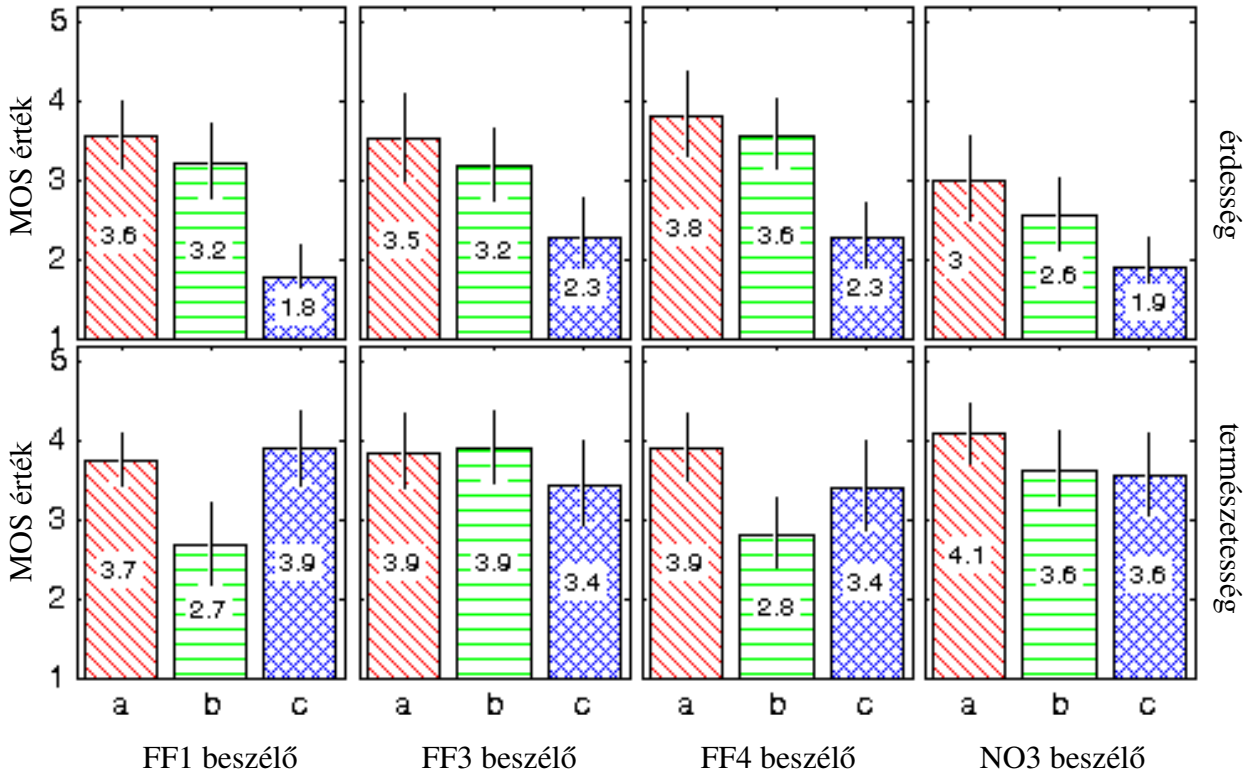
Az irreguláris-reguláris transzformáció működését a PPBA adatbázis négy beszélőjének (3 férfi: FF1, FF3 és FF4 és egy nő: NO3) hanganyagán teszteltük [102]. Mind a négy beszélő gyakran használ irreguláris fonációt, elsősorban szakaszhatárokon és a mondatok végén. Kiválasztottunk 4-4 szót, amelyek reguláris és irreguláris formában is előfordultak az adatbázisban. Ezután az irreguláris változatot transzformáltuk a fenti módszerrel. Bizonyos esetekben csak egy magánhangzó volt glottalizált, míg más mintákban a teljes zöngés szakaszt irreguláris módon ejtette a beszélő. Azokban az esetekben, amelyben az $F0$ interpoláció nem volt megfelelő, kézzel javítottuk az $F0$ menetet, hogy illeszkedjen a mondat ereszkedő dallammenetéhez. A szavak 3-3 változatát (eredeti irreguláris, regulárissá transzformált és eredeti reguláris) meghallgatásos tesztben hasonlítottuk össze. A 4.7. ábra egy példát mutat a teszt egyik szavának 3-3 változatára.

Az internetes meghallgatásos tesztben összesen 48 szót kellett értékelni (4 beszélő $\cdot$ 4 szó $\cdot$ 3 változat) természetesség és érdesség szerint. A teszt megkezdése előtt a kísérleti alanyokat megkértük, hogy hallgassanak meg néhány glottalizált beszédmintát, hogy egyértelműsítsük az „érdes” kifejezés jelentését. A tesztelők minden minta meghallgatása után két MOS-jellegű (*Mean Opinion Score*) kérdésre válaszoltak: 1) „Kérlek jelöld be, hogy mennyire érzed érdesnek (glottalizáltnak) a hallott hangot!” („1 - egyáltalán nem érdes” ... „5 - nagyon érdes”); 2) „Kérlek jelöld be, hogy mennyire érzed természetesnek a hallott hangot!” („1 - nagyon természetellenes” ... „5 - teljesen természetes”). A szavakat a tesztelők véletlen sorrendben hallgatták meg.

A tesztet összesen 9 magyar anyanyelvű tesztelő végezte el (mind a 9 férfi volt). Közülük hárman beszédkutatók voltak, míg a többiek egyetemi hallgatók. Az átlagos életkor 23,67 év volt (szórás: 3,20 év). Heten fejhallgatóval végezték a kísérletet, ketten hangszórón hallgatták a mintákat. Átlagosan 6,92 percig tartott a teszt kitöltése (szórás: 1,39 perc).

A teszt eredményei

A tesztelők értékelését a 4.8. ábra mutatja, melyet párosított mintás t-teszttel hasonlítottunk össze. Az elemzések szerint az eredeti irreguláris és regulárissá transzformált minták szignifikánsan különböznek érdesség szempontjából ($p < 0{,}05$), amennyiben az összes adatot egybe vesszük. A MOS értékeket megvizsgálva azt vehetjük észre, hogy az eredeti irreguláris minták érdessége lényegesen csökkent (de ez még nem éri el az eredeti reguláris minták szintjét). Összességében a módszer szignifikáns mértékben csökkentette az eredeti minták természetességét ($p < 0{,}05$). Amennyiben az eredményeket beszélőnként vizsgáljuk, az érdesség külön-külön



4.8. ábra. Az irreguláris-reguláris transzformációval módosított szavak szubjektív elemzésének eredménye: a) eredeti irreguláris b) regulárissá transzformált c) eredeti reguláris. A függőleges fekete vonalak a 95%-os konfidenciaintervallumot jelölik.

is szignifikánsan csökkent a transzformált mintákon. A természetesség megőrzésében viszont nagy beszélőnkénti különbség látható: az FF3 és NO3 beszélők esetében nem csökkent szignifikánsan a minták természetessége, míg az FF1 és FF4 beszélő esetén igen.

A mintákat utólagosan megvizsgálva az utóbbi két beszélőnél jelentkező természetesség csökkenését valószínűleg az alkalmazott maradékjel kódkönyvek felépítése és a transzformáció során az RMSE alapú összefűzési költség okozhatta. A szintézis lépésben az elemkiválasztásnál előfordult, hogy az összefűzési költség miatt ugyanazon maradékjel periódus ismételten megjelent egymás után többször, így robotos, gépies hangzást eredményezve.

A meghallgatásos teszt eredménye az, hogy az irreguláris-reguláris transzformációs módszer szignifikánsan csökkentette a beszédminták érzeti érdességét, és a négyből két beszélő esetén ezt a természetesség csökkentése nélkül tudta megtenni. A személyfüggés oka az lehet, hogy a glottalizáció különböző megjelenési formái közül a transzformációs algoritmus valószínűleg nem minden esetben tudja megtartani az eredeti beszéd természetességét.

4.2.3. Akusztikus elemzés

A 4.2.2. fejezetben meghallgatásos teszthez kiválasztott beszédmintákon akusztikus elemzést is végeztünk. A zöngeminőségnek számos akusztikai megfelelője van, melyeket a szakirodalomban következetesen használnak [61]. Ez alapján megvizsgáltuk az eredeti irreguláris,

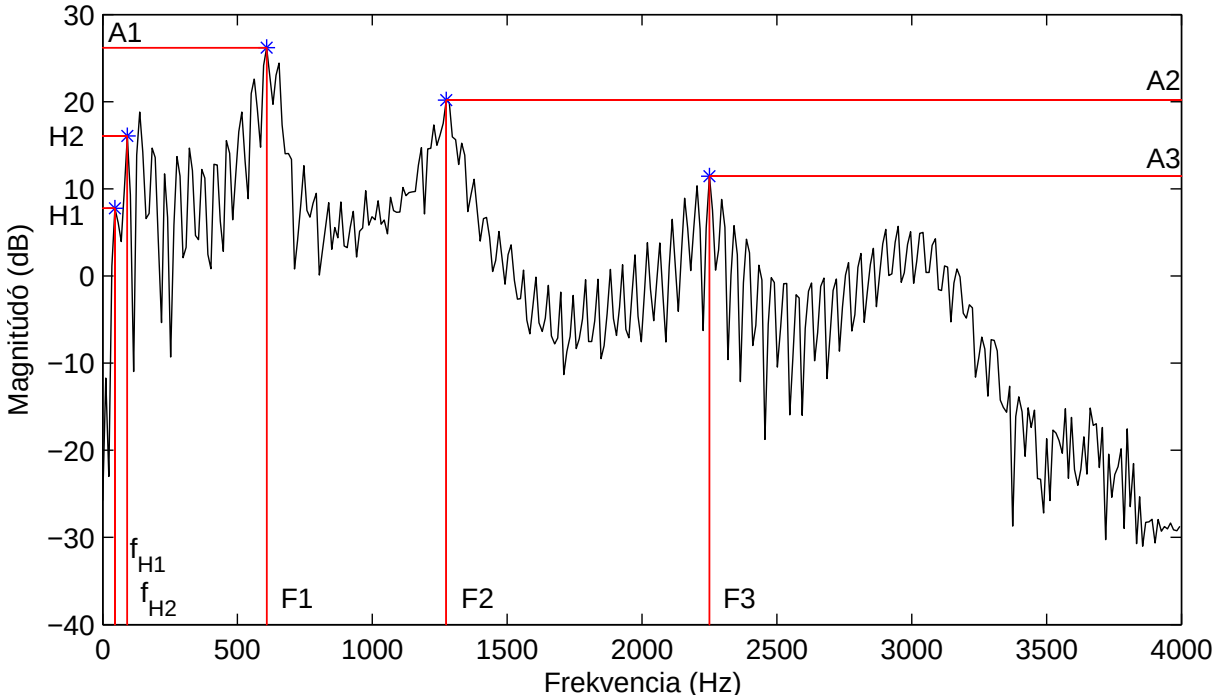
regulárissá transzformált és eredeti reguláris beszédmintákat néhány akusztikai jegy szempontjából. Amennyiben a transzformált mintákon megfelelőek a mért akusztikai paraméterek, az megmagyarázhatja, hogy miért érezték őket a tesztelők kevésbé érdesnek az eredeti irreguláris beszédhez képest.

A szakirodalomból kiválasztottunk három olyan akusztikai jegyet, amelyeket korábban irreguláris és reguláris beszéd megkülönböztetésére használtak [5, 62, 68]. Ezek alapján irreguláris zöngeképzés esetén a hangrés nyitott idejének aránya, vagyis a nyitott hányad (*open quotient*, OQ) alacsonyabb, mint reguláris zöngében. Az első formáns sávszélessége (*first formant bandwidth*, $B1$) nagyobb a glottalizáció során a gégében előforduló nagyobb mértékű akusztikai veszteség miatt. Irreguláris zöngeképzés során a hangszalagok záródása összefüggéstelenebb, azaz a spektrum lejtése (*spectral tilt*, TL) meredekebb, mint reguláris beszédben. A transzformáció hatását az OQ, $B1$, TL akusztikai jellemzőkre mérésekkel vizsgáltuk. A méréseket frekvenciatartományban végeztük, mivel így könnyebb a paraméterek számítása [5]. Holmberg és társai kimutatták, hogy az OQ arányos az első és második harmonikus dB-ben mért különbségével ($H1 - H2$) [115]. $B1$ fordítottan arányos $H1$ és az első formáns amplitúdójának különbségével ($H1 - A1$) [115], míg a TL korrelál $H1$ és a harmadik formáns amplitúdójának különbségével ($H1 - A3$) [115].

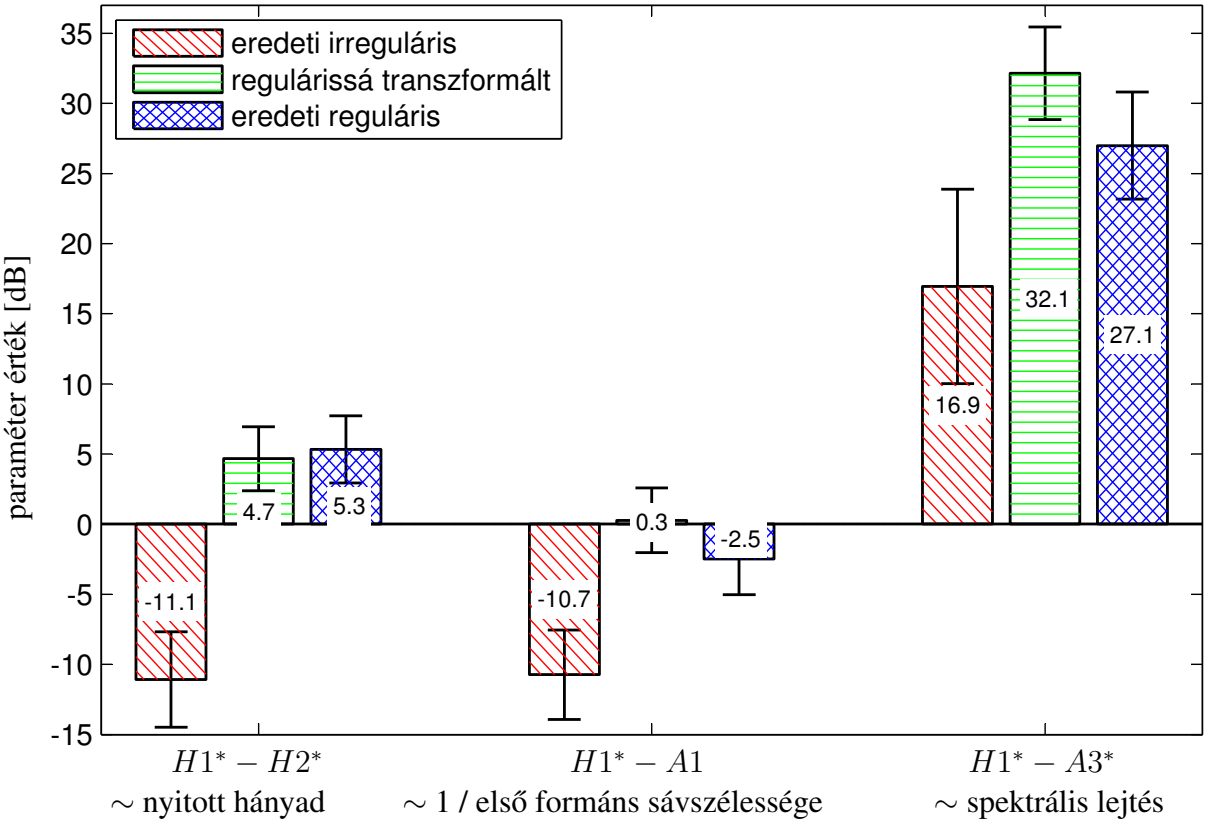
A $H1$, $H2$ és $A3$ értékeket a formánsok hatása befolyásolhatja, ezért az Iseli és társai által javasolt kompenzációt alkalmaztuk [116] a VoiceSauce program implementációjában. Ez alapján a $H1$ és $H2$ értékét az első és második formáns frekvenciája szerint korrigáltuk ($H1^*$ és $H2^*$), valamint az $A3$ értékét az első három formáns frekvenciája szerint kompenzáltuk ($A3^*$). Végül kiszámítottuk az amplitúdó különbségeket: $H1^* - H2^*$, $H1^* - A1$ és $H1^* - A3^*$.

A paramétereket a meghallgatásos teszt mintáin mértük (eredeti irreguláris, regulárissá transzformált és eredeti reguláris változatok). A hangfájlokat először 3,8 kHz-es aluláteresztő szűrésen engedtük át, majd újramintavételeztük 8 kHz-en, ami biztosította, hogy a spektrumban csak a 3,8 kHz alatti tartomány látható. Ezután az eredeti irreguláris változatokból kiválasztottunk egy glottalizált magánhangzót, majd ennek 3-3 pontját jelöltük ki (nagyjából egyenletesen elosztva és a zöngehatároknak megfelelően). A Wavesurfer programot használva 512 pontos FFT analízist végeztünk a Hanning-ablakozott jelen, majd vizuálisan leolvastuk a $H1$, $H2$ amplitúdókat és ezek frekvenciáit, az $F1$, $F2$, $F3$ valamint $A1$ és $A3$ értékeket. Az irreguláris változatokban gyakran erős al-harmonikusok jelentek meg; itt a $H1$ és $H2$ értékét a spektrális csúcsok közül az első kettőben mértük. A formánsok értékeit a legnagyobb spektrális csúcs frekvenciájaként és amplitúdójaként mértük. A mérés módszere a 4.9. ábrán látható.

A három mért akusztikai paramétert a három beszédminta típuson a 4.10. ábra mutatja be. ANOVA elemzést végeztünk, mely szerint a beszédminták típusának szignifikáns hatása volt mindhárom paraméterre ($p < 0{,}0005$). Tukey-HSD post-hoc teszttel hasonlítottuk össze a beszédminta típusok átlagos paramétereinek értékét. Ez alapján megállapítottuk, hogy a $H1^* - H2^*$ megközelítőleg azonos az eredeti reguláris és a transzformált beszédrészleteken



4.9. ábra. Az első két harmonikus (H1, H2) és az első három formáns (F1, F2 és F3) frekvenciájának és amplitúdójának (A1, A2 és A3) mérése az FFT spektrum alapján.



4.10. ábra. Az irreguláris-reguláris transzformációval módosított szavak akusztikus elemzésének eredménye. A függőleges fekete vonalak a 95%-os konfidenciaintervallumot jelölik.

($p = 0{,}938$, n.s. különbség), míg szignifikánsan különböző az eredeti irreguláris mintákhoz képest ($p < 0{,}0005$). A nyitott hányad szempontjából a transzformált változatok tehát közel vannak a modális beszédhez. Az irreguláris zöngével képzett szavak $H1^* - A1$ és $H1^* - A3^*$ különbségei szintén szignifikánsan különbözőek az eredeti reguláris és a transzformált változatokhoz képest ($p < 0{,}0005$ és $p < 0{,}05$), de az eredeti reguláris és a transzformált változatokban közel megegyeznek ($p = 0{,}336$ és $p = 0{,}321$, n.s. különbség). Eszerint a transzformált minták közel vannak az eredeti modális felvételekhez $B1$ és TL tekintetében is. A spektrum lejtés szempontjából viszont a transzformált minták értékei jóval magasabbak a természetes mintákénál, ami magyarázhatja, hogy miért érezhették a meghallgatásos teszt résztvevői a minták egy részét kevésbé természetesnek.

A transzformációs eljárás a vizsgált akusztikai jegyek (nyitott hányad, első formáns sávszélessége és spektrum lejtés) szempontjából tehát a reguláris beszédre jellemző értékek irányába módosítja az irreguláris beszédjelet.

4.3. Összegzés

A jelen fejezetben bemutatott új eredmények tézisszerű összefoglalása és az alkalmazási lehetőségek a 7. fejezetben találhatóak (*I. téziscsoport*).

A 4.1. fejezetben bemutattuk egy újszerű gerjesztési modell kidolgozását, amely a beszédjel paraméterekre bontására és abból történő visszaállítására alkalmas analízis és szintézis lépések során (*I.1. tézis*). A paraméterek módosításával a kimeneti beszéd bizonyos tulajdonságai is változtathatóak. Ezt kihasználva ismertettünk egy transzformációs eljárást a 4.2. fejezetben, amely irreguláris-reguláris beszéd átalakítására alkalmas (*I.2. tézis*). A transzformáció eredményét szubjektív és akusztikai kísérletekben vizsgáltuk. Ezek alapján az irregulárisból regulárissá alakított beszéd kevésbé érdes, mint az eredeti minták; valamint a 4.2.3. fejezet szerint három akusztikai jegy szempontjából közel van az eredeti reguláris beszédhez (*I.3. tézis*).

A gerjesztési modell kidolgozásának célja az is volt, hogy azt beszédszintézisben fel lehessen használni a beszéd minőségének javítására. A következő fejezet (*II. téziscsoport*) foglalkozik annak részleteivel, hogyan integráltuk a modellt statisztikai parametrikus beszédszintézisbe, majd milyen módon egészítettük ki ezt irreguláris zönge szintézisére alkalmas módszerekkel.

5. fejezet

A gépi beszéd-előállítás természetességének növelése újszerű gerjesztési modellel

Az 1. fejezet irodalmi áttekintése során bemutattunk számos gerjesztési modellt, amelyeket statisztikai parametrikus beszédszintézisben alkalmaznak. A módszerek egy része impulzus-zaj vagy kevert gerjesztést használ, más eljárások a glottális forrásjelet próbálják modellezni, bizonyos kísérletekben a harmonikus-zaj modellt fejlesztik tovább, és jónéhány esetben beszéd maradékjel alapú modellt alkalmaznak.

Ebben a fejezetben a 4.1. fejezet maradékjel alapú gerjesztési modelljét statisztikai parametrikus beszédszintézisbe illesztjük. A javasolt rendszert a HTS szabadon hozzáférhető változatával, az impulzus-zaj gerjesztéssel hasonlítjuk össze. Ezután a javasolt rendszert kiegészítjük két alternatív irreguláris zönge modellel.

5.1. Az új gerjesztési modell illesztése rejtett Markov-modell alapú szövegfelolvasóhoz

A 4.1. fejezetben ismertetett gerjesztési modell kidolgozása és a maradékjel paraméterekkel történő leírása során az volt a cél, hogy olyan típusú paramétereket válasszunk, amelyek gépi tanulásra alkalmasak. A HTS rendszer szabadon hozzáférhető változatában két paraméterfolyam írja le a beszédjelet, melyet az új modellben további három paraméterrel egészítettünk ki. Kísérleti úton kimutatjuk, hogy ezek a paraméterek megfelelően modellezhetőek HMM-ekkel, és az új gerjesztési modellel kiegészített beszédszintézis jobb minőséget eredményez, mint az impulzus-zaj modell.

A jobb megértéshez először ismertetjük az alaprendszert az impulzus-zaj modellel.

5.1.1. HMM-TTS alaprendszer impulzus-zaj modellel

A HTS rendszer szabadon elérhető változata az impulzus-zaj gerjesztést használja (HTS-PN, 1.3.2. rész). A HTS-PN rendszerben a tanítási lépés (1.3. ábra, szaggatott vonal felett) kezdeteként a címkézett tanító adatbázisból kinyerjük az $F0$ és az MGC paramétereket. A kísérletek során az alaprendszerben 16 kHz-en mintavételezett beszéd hullámformákat használunk. Az $F0$-számítás a Snack RAPT algoritmusával történik [110], 25 ms-os kerethossz és 5 ms eltolás értékekkel. A 34-dimenziós MGC elemzést $\alpha = 0{,}42$ és $\gamma = -1/3$ paraméterekkel végeztük [111]. Az egyes bemondásokhoz tartozó $log(F0)$ és MGC értékeket valamint első és második deriváltjaikat paraméter fájlokban tároljuk. Ezután a fonetikai átírat alapján környezetfüggő címkézés készül. A tanítás során a HMM-ek betanulják a környezetfüggő címkéknek megfelelő paraméter eloszlásokat[1]. Mivel az $F0$ paraméter változó dimenziójú (csak zöngés szakaszokon értelmezett), ezért ennek modellezésére MSD-HMM (*Multi-Space Distribution* HMM, azaz többterű eloszlású HMM) technikát alkalmaz az impulzus-zaj gerjesztésű alaprendszer. Az időtartamok modellezéséhez minden fonémára beszédállapot időtartam eloszlásokat számít a rendszer. A fonéma-függő állapot időtartamokat Gauss eloszlással modellezzük. A környezetfüggő címkézés és az alkalmazott döntési fák csökkentik az összes lehetséges hangkörnyezet kombinációját. Az egyes paraméterfolyamokat külön döntési fákkal kezeljük [15].

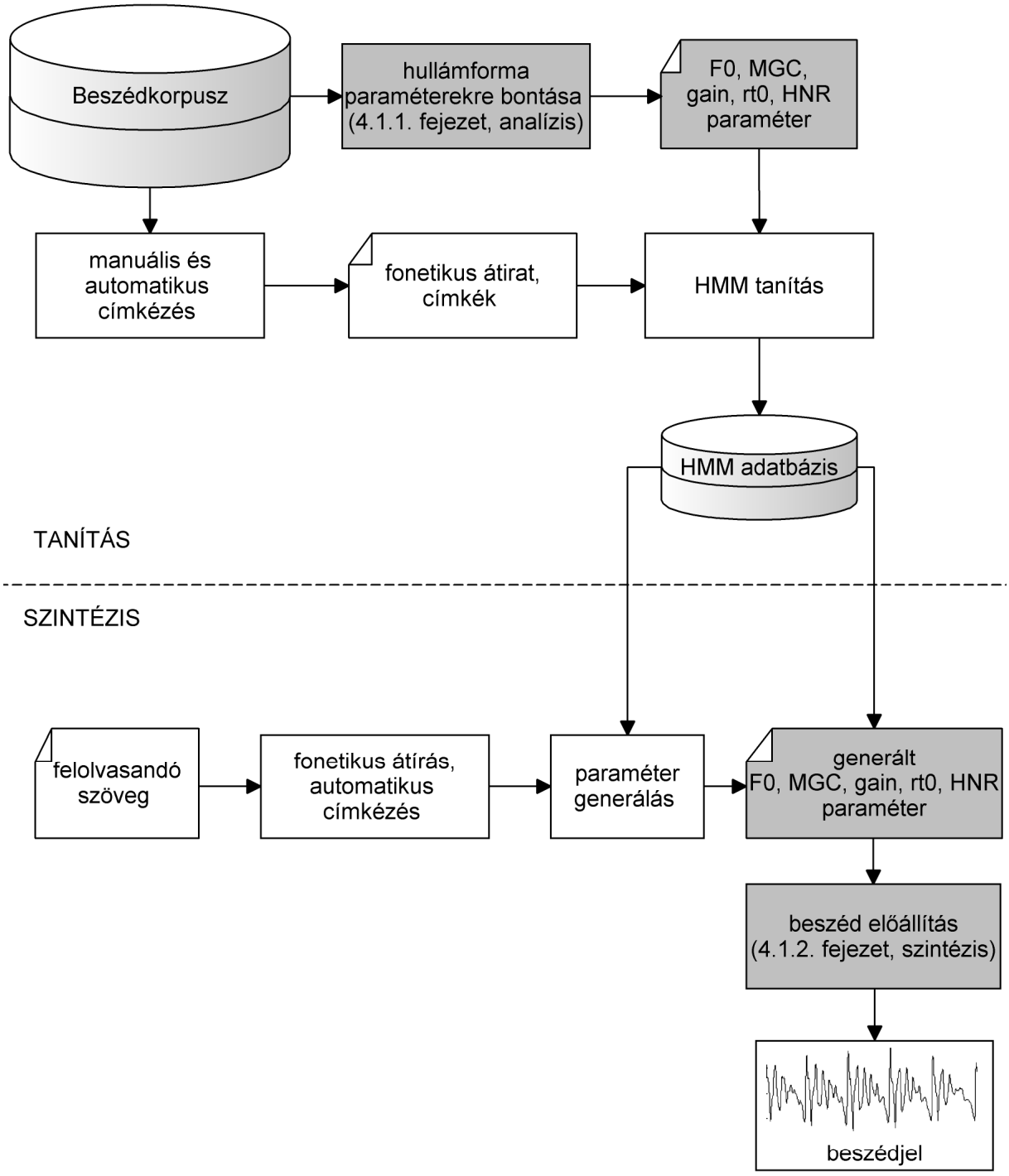
A szintézis lépés (1.3. ábra, szaggatott vonal alatt) során a szöveghez leginkább illeszkedő paramétereket ($F0$, állapot időtartamok és spektrális paraméterek) HMM-ek generálják, majd ezután az impulzus-zaj kódolóval történik a beszédjel visszaállítása. Az 1.3.2. fejezetben leírt módon a zöngés szakaszokon a gerjesztés az $F0$-nak megfelelő távolságú impulzus sorozat, míg zöngétlen szakaszokon fehérzaj. A gerjesztőjelből az MGC paramétereket felhasználó MGLSA szűrés [112] után kapjuk meg a szintetizált beszédet.

5.1.2. Az új gerjesztési modell beépítése HMM-TTS-be

A korábban kidolgozott MGC maradékjel kódkönyv alapú gerjesztési modellt a HTS rendszerbe integráltuk az 5.1. ábrán látható módon. Az új paraméterekkel kiegészítettük a tanítást, majd a szintézis lépést ezen paraméterek alapján valósítottuk meg. Az új, kiegészített rendszert HTS-CDBK-nak nevezzük.

A 4.1.1. fejezet analízis lépésénél leírt paramétereket ($F0$, $gain$, $rt0$, HNR és MGC) kiszámítjuk a tanító beszédadatbázis mondatainak minden 50 ms-os keretére, 5 ms-os eltolással. A paraméterek derivált és második derivált értékeit is eltároljuk a paraméterfolyamban. A változó dimenziójú $log(F0)$, $log(rt0)$ és $log(HNR)$ paramétereket MSD-HMM-mel modellezzük (az $F0$-hoz hasonlóan az $rt0$ és HNR paraméterek valós értékűek a zöngés keretekre, de nem értelmezettek zöngétlen esetben). A logaritmus értékek használata a kísérletek során jobb ered-

[1] A rejtett Markov-modellek tanításáról részletesen olvashatunk Tóth disszertációjában [14].



5.1. ábra. A HMM-TTS rendszer kiegészítése az új, MGC maradékjel kódkönyv alapú gerjesztési modellel (HTS-CDBK). A szaggatott vonal feletti rész a tanítási fázis, a szaggatott vonal alatti rész a szintézis fázis. A szürke hátterű dobozok jelzik az alaprendszer kiegészítéseit.

5.1. táblázat. A HTS-PN és HTS-CDBK rendszerek paramétereinek összehasonlítása.

	HTS-PN		HTS-CDBK kiegészítés		
	$F0$	MGC	$gain$	HNR	$rt0$
dimenzió	1	35	1	1	4
típus	MSD-HMM	HMM	HMM	MSD-HMM	MSD-HMM

ményre vezetett. A többi paramétert ($log(gain)$ és MGC) hagyományos HMM-ek modellezik. Az 5.1. táblázat összefoglalja a HTS-PN alaprendszerben és a HTS-CDBK kiegészített rendszerben használt paramétereket és azok tulajdonságait. A tanítás többi része (pl. környezetfüggő címkézés, döntési fák, időtartamok modellezése) az alaprendszerrel megegyező módon történik.

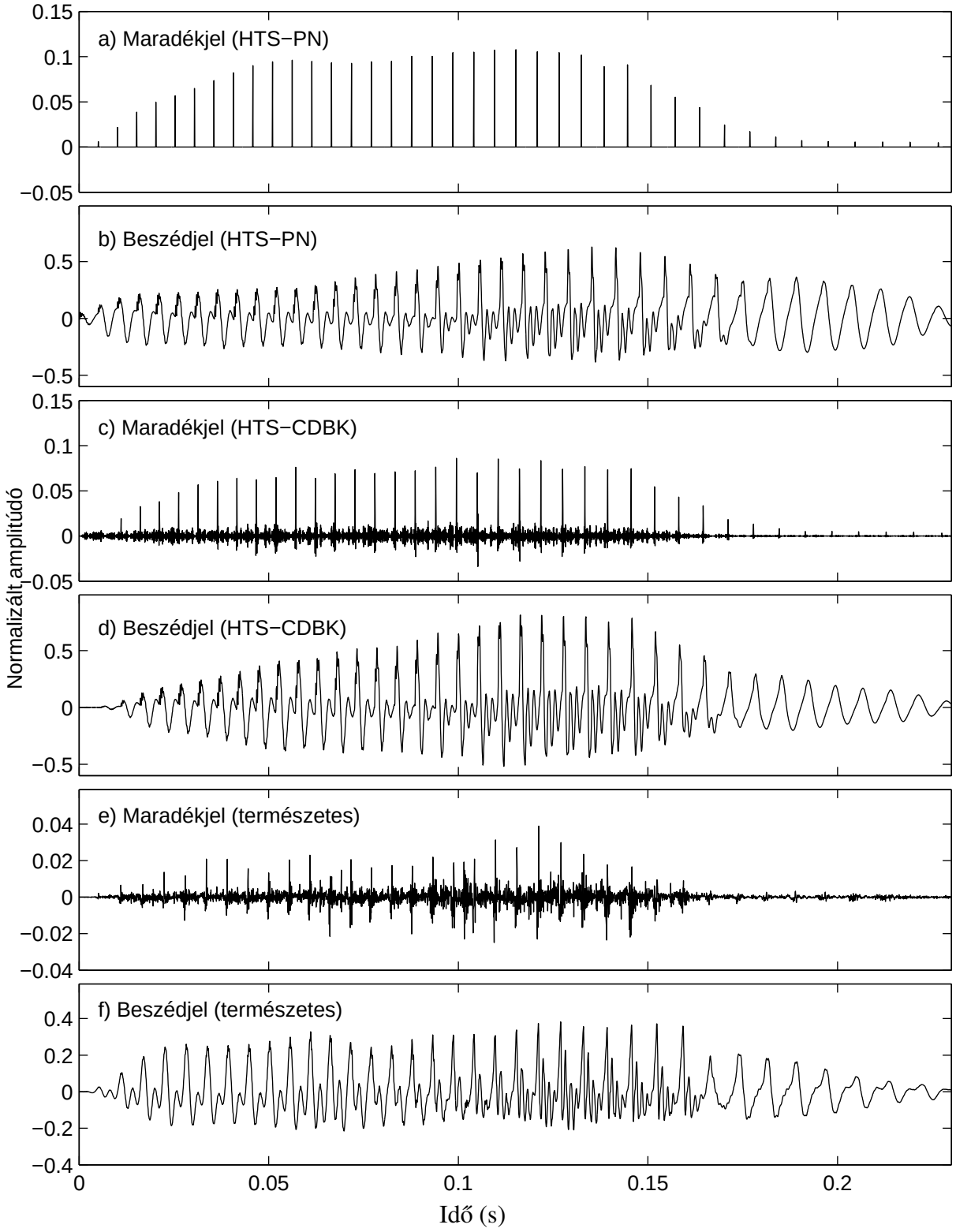
A szintézis a 4.1.2. fejezetben leírthoz hasonlóan megy végbe néhány kiegészítéssel. A gépi tanulás eredményeként kapott $F0$, $gain$, $rt0$ és HNR paraméterek és a maradékjel kódkönyv segítségével előállítjuk a maradékjelet. Ezután 6 kHz-es aluláteresztő szűrést végzünk, és a 6 kHz feletti frekvencia tartományban fehér zajt használunk a HNM alapú modellekhez hasonlóan. Erre a lépésre azért van szükség, mert lényegesen csökkenti a zöngés hangoknál előforduló zizegősséget. Végül a beszédet az MGC paraméterek segítségével szintetizáljuk MGLSA szűrővel.

5.1.3. Meghallgatásos teszt

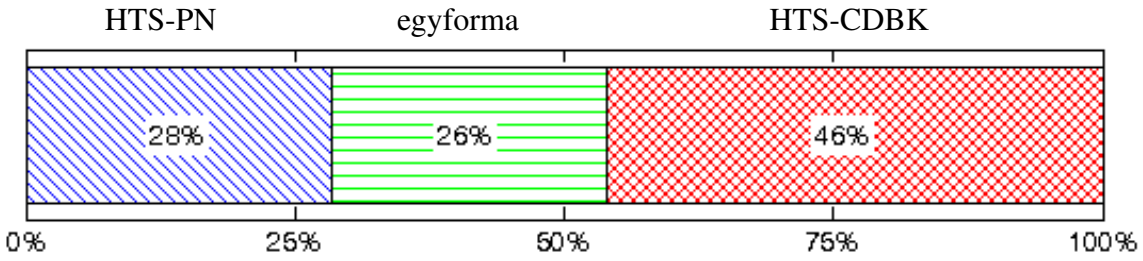
A kísérletek során magyar nyelvű mintákon végeztük a HMM-ek tanítását és minta szövegek szintézisét. Ehhez a nyelvspecifikus lépéseket a HTS-HUN rendszerből kiindulva alkalmaztuk [15].

Hanganyag és módszer, kísérleti személyek

A meghallgatásos tesztekhez szükséges beszédmintákat az alábbiak szerint állítottuk elő. A PPBA adatbázis FF2 férfi beszélőjének hanganyagával végeztünk beszédszintézis kísérleteket. Ehhez a teljes, 137 percnyi (1938 mondat) beszédfelvételt és a hozzá tartozó címkézést használtuk fel beszélőfüggő tanítás keretében. Az eredetileg 44,1 kHz-en tárolt mintákat újramintavételeztük 16 kHz-en 7,6 kHz-es aluláteresztő szűrés után. Alaprendszerként a HTS-HUN egyszerű impulzus-zaj gerjesztésű változatát (HTS-PN) használtuk. Az FF2 beszélő maradékjelei alapján 6 500 elemből álló kódkönyvet készítettünk a HTS-CDBK rendszerben. A 16 kHz-es minták használatára az volt a motiváció, hogy így a HTS-PN és HTS-CDBK rendszerek mintái közvetlenül összehasonlíthatóak. Mindkét rendszerrel 130-130 olyan mondatot szintetizáltunk, amely nem fordult elő a tanító adatbázisban. 20-20 mondatot kiválasztottunk egy meghallgatásos teszthez, melyben a két rendszert hasonlítottuk össze.



5.2. ábra. Az „ilyen” szó szintetizált és természetes gerjesztőjele valamint beszéd hullámformája:
a) maradékjel b) beszédjel a HTS-PN alaprendszerrel,
c) maradékjel d) beszédjel a HTS-CDBK rendszerrel,
e) maradékjel f) beszédjel természetes beszédből.



5.3. ábra. A HTS-PN és HTS-CDBK beszédszintézis rendszerek minőség szerinti szubjektív összehasonlításának eredménye.

Az 5.2. ábra egy példát mutat a rendszerekkel szintetizált valamint természetes gerjesztőjelre és beszédjelre: az a) ábrán egy zöngés szakasz gerjesztőjelében látszik az $F0$-függő impulzussorozat, míg a b) rész mutatja az impulzus-zaj szintézis eredményét a HTS-PN rendszerrel. A HTS-CDBK rendszer maradékjelén (c) látszik, hogy ez lényegesen több információt tartalmaz az impulzus gerjesztésnél (a), valamint az MGC maradékjel kódkönyv alapú módszerrel szintetizált beszédminta (d) is különbözik az alaprendszertől (b). Emellett természetes beszéddel összevetve (e és f ábra) észrevehető, hogy a HTS-CDBK rendszer hullámformái jobban közelítenek az eredeti beszédből származó mintához, mint a HTS-PN rendszeréi.

Az internetes meghallgatásos tesztben a tesztelőknek összesen 40 mondatot (2 rendszer · 20 mondat) kellett értékelniük páros összehasonlítás keretében természetesség szerint. A tesztelők a mintákat párosával hallgatták meg véletlen sorrendben, és minden mintapár után a következő CMOS-jellegű (*Comparative Mean Opinion Score*) kérdésre válaszoltak: „Kérlek jelöld be, hogy melyik mondatot hallod jobb minőségűnek!” („1 - első sokkal jobb” ... „5 - második sokkal jobb”).

A tesztet összesen 16 magyar anyanyelvű tesztelő végezte el. Az eredmények alapján egyikük véletlenszerűen válaszolt és az átlagosnál lényegesen rövidebb idő alatt fejezte be a tesztet, így az elemzésben 15 tesztelő válaszait összegeztük (12 férfi, 3 nő). A kísérleti alanyok egyike sem volt beszédkutató. Az átlagos életkor 32,00 év volt (szórás: 9,02 év). Öten fejhallgatóval végezték a kísérletet, tízen hangszórón hallgatták a mintákat. Átlagosan 5,08 percig tartott a teszt kitöltése (szórás: 1,47 perc).

A teszt eredményei

Az 5.3. ábra mutatja a CMOS kérdésre adott válaszok összesítését. Az ábrán a kérdésre adott válaszokat összevontuk: az „1” és „2” válaszok aránya a bal oldali „HTS-PN” részben, a „3” válaszok a középső „egyforma” részben, a „4” és „5” válaszok a jobb oldali „HTS-CDBK” részben láthatóak. A mintákon végzett statisztikai elemzés (egymintás t-teszt) szerint a CMOS értékek összesített átlaga szignifikánsan ($p < 0{,}0005$) különbözik a 3,0 értéktől (CMOS átlag: 3,23), azaz a teszt adatainak összesítése alapján a tesztelők a HTS-CDBK rendszert tartották jobb minőségűnek. A tesztelők válaszait mondatonként is megvizsgáltuk, mely szerint az egyik

mondatpár esetén a kísérleti alanyok a HTS-PN rendszert preferálták. A szintetizált beszédmintákat utólag elemezve azt vettük észre, hogy bizonyos esetekben a HTS-CDBK generált $gain$ paraméterében hirtelen változás fordult elő, ami zavaró, erős amplitúdó ugrást eredményezett.

A meghallgatásos teszt összesített eredménye szerint tehát a kísérleti alanyok az új, MGC maradékjel kódkönyv alapú gerjesztést használó szintetizált beszédmintákat jobb minőségűnek hallották az alaprendszerhez képest.

5.1.4. Irreguláris zönge kezelése az alaprendszerben

A statisztikai parametrikus beszédszintézis és a legtöbb ebben használt gerjesztési modell (így az 5.1.2. fejezetben ismertetett módszer is) ideális beszéd esetén működik megfelelően, és számos hibát eredményez nem-modális zöngeképzés, például irreguláris fonáció esetén (ld. 1.4. fejezet). A glottalizált beszédszakaszokon (általában a mondatok utolsó szótagjában) az $F0$-mérő algoritmus nem megfelelően méri az $F0$-t és zöngétlennek ítéli a keretet. Ezt a mintázatot a gépi tanulás is megtanulja, és az irreguláris fonációt a HTS-CDBK rendszer a zöngétlen beszédhez hasonlóan modellezi. Ez kellemetlen, rossz minőségű hangzást okoz, és nem megfelelő modellje a glottalizációnak.

A továbbiakban az 5.1.2. fejezetben ismertetett HTS-CDBK rendszert használjuk alaprendszernek és ezt egészítjük ki irreguláris zönge szintézisére alkalmas modellekkel.

5.2. Az új gerjesztési modell felhasználása irreguláris beszéd gépi előállítására

Az irreguláris zönge (azaz glottalizáció) szintézise az 1.4.3. fejezet irodalmi áttekintése szerint egy új kutatási terület, mellyel kevesen foglalkoztak eddig. Statisztikai parametrikus beszédszintézisben nem történtek még széles körű vizsgálatok a glottalizáció megfelelő modellezéséről. Ebben a fejezetben bemutatunk két alternatív irreguláris zönge modellt, amelyeket rejtett Markov-modell alapú beszédszintézisbe illesztünk. A módszerek eredményességét meghallgatásos és akusztikai tesztekkel vizsgáljuk.

5.2.1. Szabály alapú irreguláris zönge modell kidolgozása

Az első megközelítés esetén olyan heurisztikákat alkalmazunk, melyek segítségével az irreguláris zönge tulajdonságai egyszerű szabályokkal modellezhetőek. A természetes beszéd során az irreguláris zöngeképzés több tulajdonságban különbözik a reguláris fonációtól ([5, 68], 1.4. fejezet) :

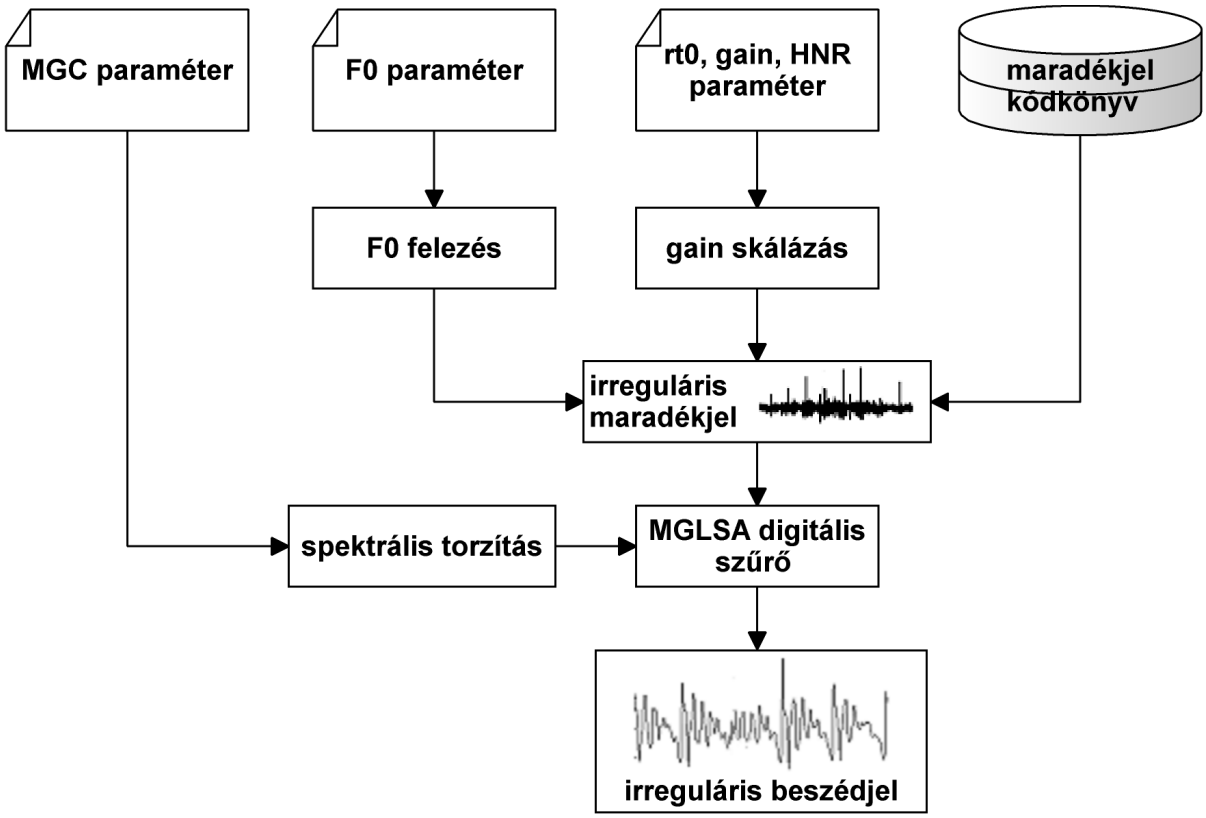
- az egymást követő glottális impulzusok között eltelő idő hosszabb és szabálytalanabb, amely alacsonyabb $F0$-t és magasabb jittert eredményez,

- az összesített intenzitás szint alacsonyabb,
- az irreguláris zöngeperiódusok amplitúdójában hirtelen változások fordulnak elő,
- a nyitott hányad (a glottális ciklus azon szakasza, amikor a hangrés nyitva van) alacsonyabb,
- az első formáns sávszélessége nagyobb a hangrésben jelentkező nagyobb mértékű akusztikai veszteségek miatt,
- a hangszalagok záródása kevésbé szabályos, azaz a spektrális lejtés meredekebb.

Bőhm korábban létrehozott egy reguláris-irreguláris transzformációs módszert, amely a fenti tulajdonságok alapján az egyes glottális ciklusok amplitúdóját skálázza ([5, 62], 1.4.2. fejezet). A módszer alkalmas irreguláris zönge előállítására, de manuális vagy félautomatikus működésű, ezért eredeti formájában nem alkalmas beszédszintézisbe integrálásra. A reguláris-irreguláris transzformációs eljárás egyes ötleteit felhasználtuk, és ez alapján egy automatikus módszer készült, amely beszédszintézisben is használható.

Az új, szabály alapú irreguláris zönge szintézisére alkalmas rendszerben az analízis és a tanítási lépések a HTS-CDBK rendszervével egyezőek (5.1.2. fejezet), az új rendszer csak a szintézis fázisban különbözik. A kiegészített rendszert HTS-CDBK+Irreg-Rule-nak nevezzük. A HTS-CDBK rendszerhez hasonlóan a glottalizáció helyére nincs külön előrejelző eljárás, hanem azt a generált $F0$ paraméterfolyamból állapítjuk meg. Amennyiben legalább 5 egymás utáni magánhangzó keretben ($1 \cdot 25ms$ az első keret + $4 \cdot 5ms$ eltolás, összesen $45ms$) nulla az $F0$ értéke, alkalmazzuk az irreguláris zönge modellt az adott magánhangzóra. Ezekben az esetekben az $F0$-menetet lineárisan interpoláljuk a környező zöngés részeknek megfelelően, vagy amennyiben nincs ilyen, akkor enyhén ereszkedő $F0$-menetet állítunk be. A HTS-CDBK+Irreg-Rule rendszer három heurisztikát használ az irreguláris zönge modellezésére az akusztikai jellemzők fenti felsorolásának megfelelően: 1) $F0$ felezés 2) zöngeszinkron maradékjel amplitúdó skálázás véletlen számokkal és 3) spektrális torzítás. Az 5.4. ábra bemutatja a módszer működését.

1) A szintézis során a modális zöngés és zöngétlen részeken a HTS-CDBK rendszer által generált maradékjelet használjuk. Azokban a szakaszokban, amelyeknek szintézise irreguláris módon történik, az interpolált $F0$ értékek felét használjuk fel („$F0$ felezés” az 5.4. ábran). A glottalizációt gyakran extrém alacsony alapfrekvencia kíséri, a kódkönyvben viszont kevés az ilyen $F0$-lal rendelkező elem. Emiatt a maradékjel periódusokat nullákkal töltjük ki az átlapolt összeadás előtt. Az $F0$ felezés és nullákkal kitöltés eredménye olyan, mintha minden második periódust törölnénk, és ez percepciós szempontból hasonló, mint az alacsonyabb nyitott hányad [62].



5.4. ábra. A szabály alapú irreguláris zönge modell szintézis része (HTS-CDBK+Irreg-Rule).

2) A maradékjel szintézisben a kiválasztott kódkönyv elemeken amplitúdó skálázást végzünk: a módszer minden zönge periódust megszoroz egy $\{0 \ldots 1\}$ közötti egyenletes eloszlású véletlen értékkel („$gain$ skálázás” az 5.4. ábran). A heurisztika alkalmazását az motiválta, hogy irreguláris zönge esetén az egymás utáni periódusok amplitúdója sokszor ingadozó a kváziperiodikus rezgéssel szemben (ld. 1.5. ábra). Ez a lépés hasonlít Bőhm korábbi módszeréhez [62], de jelen eljárásban automatikusan meghatározott, véletlen számokkal skálázzuk a periódusokat a manuálisan beállított értékek helyett. A skálázott maradékjel periódusokból az eredeti HTS-CDBK szintézis szerint átlapolt összeadással állítjuk elő a teljes maradékjelet.

3) Végül spektrális torzítást alkalmazunk. Korábbi kutatásban észrevettük, hogy az irreguláris szakaszokon mért MGC paraméterfolyam kevésbé sima a reguláris beszédhez képest (4.2. fejezet). Emiatt az irreguláris zönge modellben az MGC értékeket torzítjuk: $\{0{,}995 \ldots 1{,}005\}$ közötti véletlen számokkal szorozzuk a paramétereket, ami várhatóan az irreguláris zöngéhez hasonló hatást eredményez („spektrális torzítás” az 5.4. ábran). Az MGC ábrázolás a spektrum interpolálható reprezentációja, ezért az MGC értékek módosítása nem okoz instabilitást. A szintetizált beszédet a maradékjelből a korábbiakhoz hasonlóan MGLSA szűréssel, az MGC paramétereket felhasználva kapjuk vissza.

Az 5.5. ábra egy példát mutat a HTS-CDBK (a és b) és a HTS-CDBK+Irreg-Rule (c és d) rendszerek által generált szóra (vízszintes nyíl jelöli az irreguláris szakaszt). A „Mihály” szó „á” hangjában alkalmaztuk az irreguláris zönge modellt. A nullákkal kitöltés eredményeként a zöngeperiódusok elkülönülnek, míg az amplitúdó skálázás a negyedik periódus erős lecsökkenését eredményezte. Az ábrán látható, hogy a szabály alapú irreguláris zönge modell eredménye jobban hasonlít az eredeti glottalizált beszédre (4.7. a és 4.7. b ábra), mint az alaprendszer.

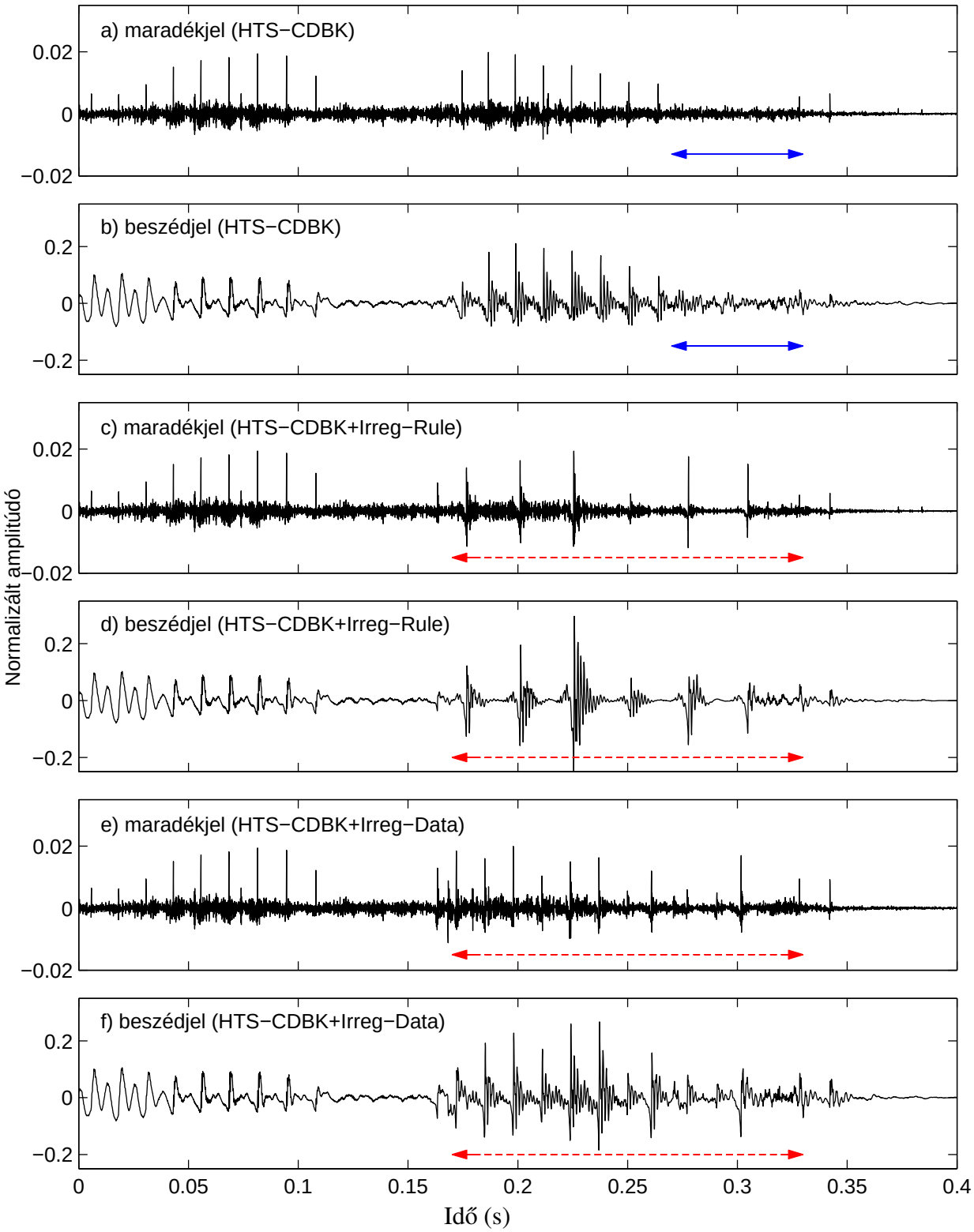
5.2.2. Meghallgatásos teszt a szabály alapú modell vizsgálatára

A szabály alapú irreguláris zönge modell eredményének vizsgálatára percepciós tesztet végeztünk. A módszer felhasználhatóságát nagyban befolyásolja, hogy az emberek számára megfelelő-e a modell által szintetizált beszéd. A meghallgatásos teszt egyik célja az volt, hogy megvizsgáljuk, mennyire kellemes az új módszer eredménye a HTS-CDBK alaprendszerhez képest. A második cél az eredeti beszélőhöz való hasonlóság vizsgálata volt.

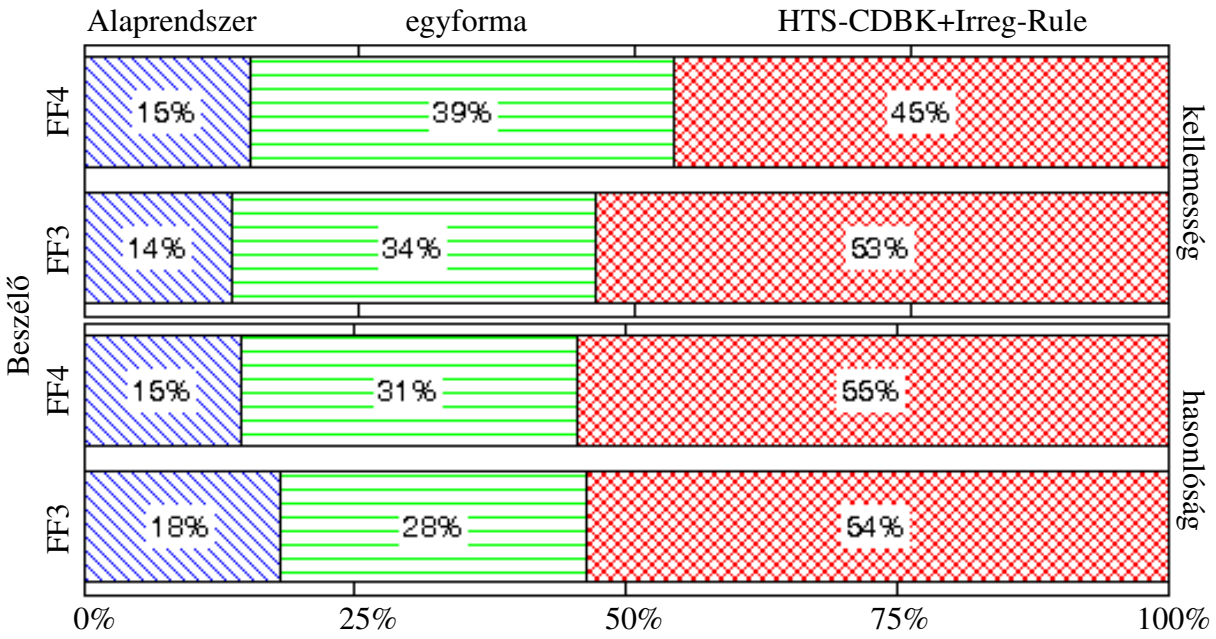
Hanganyag és módszer, kísérleti személyek

A PPBA adatbázis két férfi beszélőjének (FF3 és FF4) hangja alapján tanítást végeztünk a HTS-CDBK alaprendszerrel és a HTS-CDBK+Irreg-Rule kiegészített rendszerekkel. Mivel az irreguláris zönge modell csak a szintézis lépésben különbözik az alaprendszertől, ezért ugyanazokat a beszédadatbázisokat használtuk fel a tanítás során. A maradékjel kódkönyv FF3 beszélő esetén 3394 elemet, FF4 beszélő esetén 2218 elemet tartalmazott, melyeket a beszédadatbázis első 150 mondatából, kb. 10 percnyi hanganyag alapján készítettünk. 130-130 mondatot szintetizáltunk, majd ebből 10-10 olyan mondatot választottunk, amelyben előfordult irreguláris fonáció. A mondatok utolsó, irreguláris magánhangzót tartalmazó, legalább két szótagú szavát kivágtuk és ezeket használtuk fel a meghallgatásos tesztben. Az 5.5. b) és d) ábra egy példát mutat a tesztben szereplő beszédmintákra.

A tesztben minden szónak két változata szerepelt, így összesen 40 mintát kellett meghallgatniuk a tesztelőknek páros összehasonlítás keretében (2 beszélő $\cdot$ 10 szó $\cdot$ 2 változat). Internetes meghallgatásos tesztet készítettünk két CMOS-jellegű kérdéssel. A teszt megkezdése előtt megkértük a tesztelőket, hogy hallgassanak meg egy beszédmintát FF3 beszélőtől. A teszt első felében a preferenciát vizsgáltuk (kérdés: „Melyik változatot hallgatnád szívesebben?” válaszok: „1 - elsőt sokkal szívesebben” ... „5 - másodikat sokkal szívesebben”). A teszt második része azt vizsgálta, hogy melyik változat áll közelebb az eredeti beszélőhöz. Ehhez a mintapár meghallgatása előtt egy referencia mintát játszottunk le az adott beszélőtől (kérdés: „Melyik változat hasonlít jobban az eredeti beszélőre?” válaszok: „1 - első jobban hasonlít” ... „3 - második jobban hasonlít”). A meghallgatandó mintapárok sorrendje véletlen volt minden tesztelő esetén.



5.5. ábra. A „Mihály” szó szintetizált változatai (egy hosszabb mondatból kivágva):
a) maradékjel b) beszédjel a HTS-CDBK alaprendszerrel
c) maradékjel d) beszédjel a HTS-CDBK+Irreg-Rule modellel
e) maradékjel f) beszédjel a HTS-CDBK+Irreg-Data modellel.
Az irreguláris zöngeképzésű szakaszokat vízszintes nyilak jelölik.



5.6. ábra. A HTS-CDBK alaprendszerrel és HTS-CDBK+Irreg-Rule szabály alapú irreguláris zönge modellekkel szintetizált szavak szubjektív összehasonlításának eredménye.

A tesztet összesen 11 magyar anyanyelvű tesztelő végezte el (9 férfi, 2 nő). A kísérleti alanyok egyike sem volt beszédkutató. Az átlagos életkor 23,81 év volt (szórás: 4,31 év). Tízen fejhallgatóval végezték a kísérletet, egy tesztelő hangszórón hallgatta a mintákat. Átlagosan 9,03 percig tartott a teszt kitöltése (szórás: 2,09 perc).

A teszt eredményei

A meghallgatásos teszt eredményeit az 5.6. ábra mutatja. Az ábra a két beszélőre külön összehasonlítja a HTS-CDBK alaprendszer (bal oldal) és a kiegészített HTS-CDBK+Irreg-Rule rendszer (jobb oldal) szintetizált mintáira adott értékeléseket. Látható, hogy a preferencia kérdés esetén mindkét beszélőnél a válaszok eredménye magasabb, mint az 50%-os átlag (CMOS = 3,0), vagyis a kiegészített rendszert preferálták a tesztelők (átlagos CMOS = 3,36). A hasonlósági kérdésre adott válaszok összesített értéke is magasabb az 50%-nál (CMOS = 2,0), amennyiben FF3 és FF4 beszélőt együtt vizsgáljuk (átlagos CMOS = 2,38). A tesztelők értékelését t-teszttel is összehasonlítottuk. A statisztikai elemzés szerint a szabály alapú irreguláris modell szignifikánsan preferáltabb ($p < 0{,}0005$) és szignifikánsan jobban hasonlít az eredeti beszélőre ($p < 0{,}0005$), mint az alaprendszer. A tesztbeli mintapárokat egyesével is megvizsgáltuk, amely szerint a hasonlóság kérdés esetén mindig az új rendszer kapott magasabb értékelést, míg a preferencia kérdésnél a 20-ból 18 esetben érezték jobbnak az irreguláris zöngével kiegészített módszert a tesztelők.

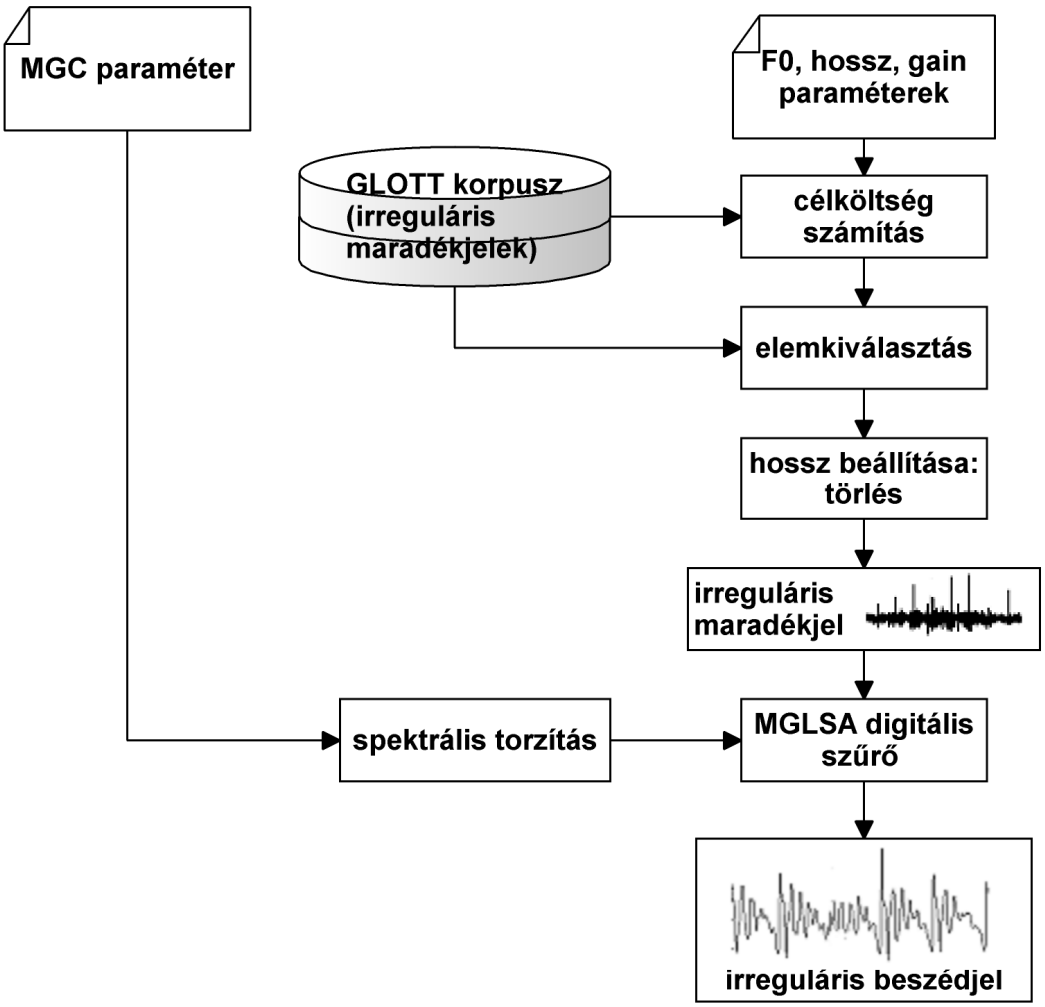
A meghallgatásos teszt alapján azt a következtetést vonhatjuk le, hogy a HTS-CDBK+Irreg-Rule rendszer növeli a szintetizált beszédminták természetességét a HTS-CDBK rendszerhez képest és az irreguláris zönge modellel az eredeti beszélőre jobban emlékeztető beszéd hozható létre.

5.2.3. Adatvezérelt irreguláris zönge modell kidolgozása

Az irreguláris zönge szintézisbe illesztésére egy másik, adatvezérelt modellt is létrehoztunk, amely maradékjel elemkiválasztáson alapul. A kiegészített rendszert HTS-CDBK+Irreg-Datanak nevezzük. Az új modellben az analízis és a tanítási lépések a HTS-CDBK rendszerével egyezőek (5.1.2. fejezet), az új rendszer csak a szintézis fázisban különbözik. Emellett az új modellben egy irreguláris beszédszakaszok maradékjeléből álló korpuszt is építünk az analízis után. Feltételezéseink szerint az adatvezérelt modell által szintetizált beszéd jobb minőségű lesz, mint a korábbi rendszerek.

Az analízis az alaprendszerhez hasonlóan történik, azaz kinyerjük az öt paraméterfolyamot minden beszédmintából. Az analízis elvégzése után a beszédadatbázis irreguláris szakaszainak maradékjeléből glottalizációs korpuszt építünk („GLOTT” korpusz). Az irreguláris szakaszok megtalálásához egy magas találati arányú glottalizáció detektort alkalmazunk („creak_detect”, [67]). Azokat a maradékjel szakaszokat vesszük be a GLOTT korpuszba, amelyek esetén a detektor a magánhangzó kereteinek legalább felében „irreguláris” bináris döntést hozott. Az adatvezérelt módszernél teljes, magánhangzó-hosszúságú maradékjel szakaszokat tárolunk a korpuszban a korábbi zönge-szinkron maradékjel periódusokkal szemben.

A szintézis során (5.7. ábra) a modális maradékjel szakaszok a HTS-CDBK módszerével készülnek. A HTS-CDBK és HTS-CDBK+Irreg-Rule rendszerekhez hasonlóan a glottalizáció helyére nincs külön előrejelző eljárás, hanem azt a generált $F0$ paraméterfolyamból állapítjuk meg. Az irreguláris részekhez a glottalizációs korpuszból keresünk illeszkedő elemet. A módszer jelen változatában azt feltételezzük, hogy csak egy magánhangzót kell irreguláris módon szintetizálni, így nem foglalkozunk az elemek közötti összefűzéssel. Az elemkiválasztáshoz csak célköltséget használunk, ami három rész-költségből áll: 1) a paraméterfolyamból származó és a kódkönyv elemek közötti átlagos $F0$ különbség 2) átlagos hossz különbség valamint 3) a maradékjel szakasz hangkörnyezete. Olyan elemeket keresünk, amelyek a szintetizálandó szakasznál hosszabbak. A cél maradékjel hangkörnyezetét azért használjuk költségként, mert így a szintetizálandó szakasznak megfelelő maradékjel darabot találhatunk a GLOTT korpuszban. Miután a cél maradékjelet megtaláltuk a célköltség minimalizálásával, a kiválasztott maradékjel utolsó mintáit levágjuk, így beállítva a jeldarab hosszát. Az irreguláris maradékjel energiáját a $gain$ paraméterek átlaga alapján skálázzuk, de a jel más tulajdonságát nem módosítjuk. Ezáltal azt feltételezhetjük, hogy a szintetizált beszédszakasz az irreguláris fonációnak megfelelő tulaj-



5.7. ábra. Az adatvezérelt irreguláris zönge modell szintézis része (HTS-CDBK+Irreg-Data).

donságú lesz. A HTS-CDBK+Irreg-Rule modellhez hasonlóan spektrális torzítást alkalmazunk, és a végül az MGC paramétereket felhasználó MGLSA szűréssel állítjuk elő a szintetizált beszédet az összefűzött modális és irreguláris maradékjel szakaszokból.

Az 5.5. ábra egy példát mutat az adatvezérelt irreguláris fonáció modell eredményére (e és f). Az alap HTS-CDBK rendszerhez (a és b) hasonlóan a HTS-CDBK+Irreg-Data maradékjele is csak az utolsó magánhangzó egy részében tartalmaz hirtelen amplitúdó ingadozást. Ha ezt összehasonlítjuk az eredeti irreguláris beszédmintával (4.7. a és 4.7. b ábra), az látható, hogy a szintetizált maradékjel is másodlagos impulzusokat tartalmaz a periódusokon belül, az eredeti beszédből származtatott maradékjelhez hasonlóan.

5.2.4. Meghallgatásos teszt az adatvezérelt modell vizsgálatára

A korábbiakhoz hasonlóan percepciós teszttel ellenőriztük az adatvezérelt irreguláris zönge modell eredményét a PPBA adatbázis FF3 és FF4 beszélőjének mintái alapján. Az új HTS-CDBK+Irreg-Data adatvezérelt modellt a HTS-CDBK alaprendszerrel és a HTS-CDBK-Irreg-Rule szabály alapú irreguláris zönge modellel hasonlítottuk össze kellemesség és az eredeti beszélőhöz való hasonlóság szempontjából.

Hanganyag és módszer, kísérleti személyek

A HTS-CDBK+Irreg-Data analízis során FF3 beszélő teljes anyaga alapján 1116 maradékjel szakasz, FF4 beszélő beszédadatbázisa esetén 1822 maradékjel szakasz került a GLOTT korpuszba. 130-130 mondatot szintetizáltunk a HTS-CDBK alaprendszerrel és a HTS-CDBK+Irreg-Rule valamint HTS-CDBK+Irreg-Data rendszerekkel, majd 10-10 mondatot kiválasztottunk, amelyek tartalmaztak irregulárisan szintetizált magánhangzót, és ezeket a szavakat kivágtuk (példa: 5.5. ábra).

Az internetes meghallgatásos teszt körülményei és kérdései az előző teszthez hasonlóak voltak (5.2.2. fejezet). A tesztelők a szavak különböző változatait értékelték páros összehasonlítás keretében, összesen 80 mintapárt meghallgatva véletlenszerű sorrendben (párok: alaprendszer vs. adatvezérelt, illetve szabály alapú vs. adatvezérelt).

A tesztet összesen 17 magyar anyanyelvű tesztelő végezte el (13 férfi, 4 nő). Közülük heten beszédkutatók voltak. Az átlagos életkor 31,76 év volt (szórás: 11,15 év). 13-an fejhallgatóval végezték a kísérletet, négyen hangszórón hallgatták a mintákat. Átlagosan 17,11 percig tartott a teszt kitöltése (szórás: 7,01 perc).

A teszt eredményei

Az 5.8. ábra mutatja az alaprendszer (bal oldal) vs. HTS-CDBK+Irreg-Data (jobb oldal) összehasonlítását, míg az 5.9. ábrán láthatóak a szabály alapú (bal oldal) vs. adatvezérelt (jobb oldal) irreguláris zönge modellek összehasonlításának eredményei. Mindkét ábrán a középső rész jelöli az egyforma értékeléseket.

Az alaprendszer és az új, adatvezérelt modell összehasonlításában az látható beszélőnként és kérdésenként külön-külön az 5.8. ábrán, hogy a tesztelők az új rendszert részesítették előnyben. Az első kérdésre az átlagos CMOS = 3,36, ami a t-teszt szerint szignifikánsan különbözik 3,0-tól ($p < 0{,}0005$), míg a második kérdésre az átlagos CMOS = 2,28, ami szignifikánsan különbözik 2,0-tól ($p < 0{,}0005$). Az eredmények hasonlóan szignifikánsak, ha a két beszélőt külön vizsgáljuk.

A két alternatív irreguláris zönge modell összehasonlítását az 5.9. ábra mutatja. Preferencia szempontjából nincs szignifikáns különbség a két módszer által szintetizált minták között (átlagos CMOS = 3,07, nem szignifikáns a különbség 3,0-tól, $p = 0{,}16$). A tesztelők hasonló módon nem éreztek szignifikáns különbséget a két modellben az eredeti beszélőre való hasonlóság során (átlagos CMOS = 1,95, nincs szignifikáns különbség 2,0-tól, $p = 0{,}23$). A beszélőnkénti eredmények szerint a tesztelők úgy találták, hogy FF3 beszélő esetén a szabály alapú irreguláris zönge modell kicsit közelebb áll az eredeti beszélőhöz, míg FF4 beszélőnél ez nem áll fenn.

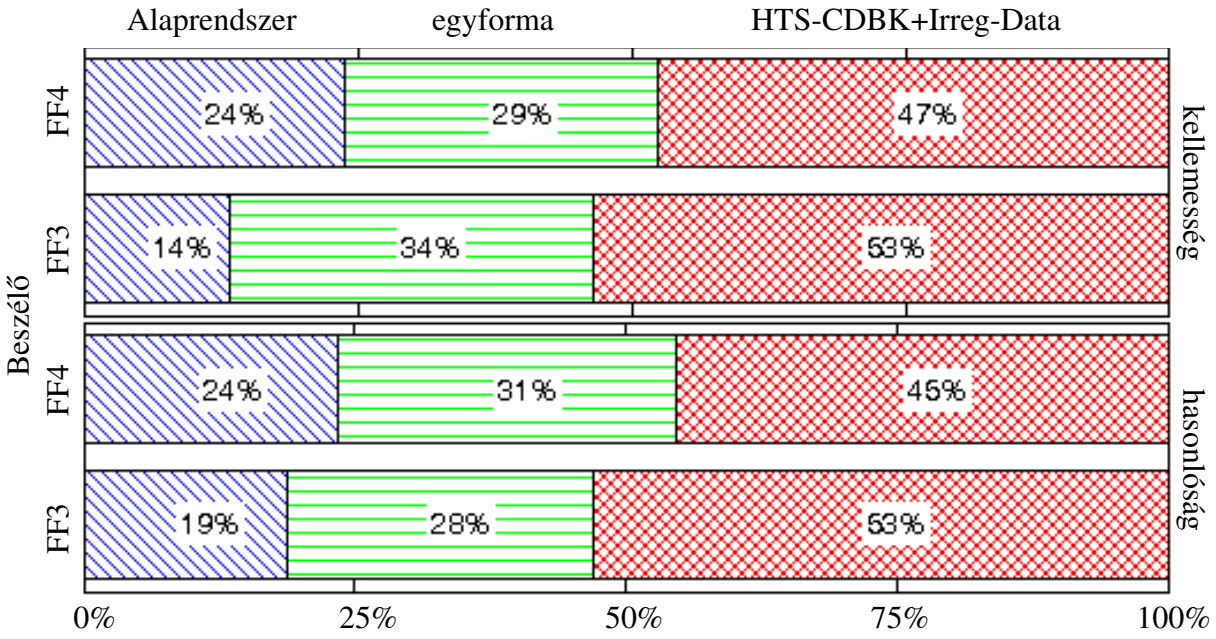
A 2. meghallgatásos teszt néhány tesztelője megjegyezte, hogy a válaszlehetőségek között hasznos lett volna egy „egyik sem hasonlít” az eredeti beszélőre válasz. Ezekben az esetekben a mintapárt egyformának értékelték.

Az egyik tesztelő megjegyezte, hogy bizonyos esetekben a mintákat túlságosan rekedtesnek találta, ami szerinte biztosan nem fordul elő természetes beszédben. A beszédmintákat megvizsgálva azt találtuk, hogy ez a megfigyelés annak köszönhető, hogy a szabály alapú irreguláris zönge modell helyenként túlságosan éles amplitúdó ingadozást eredményez. Más beszélők azonban ezt nem tartották zavarónak.

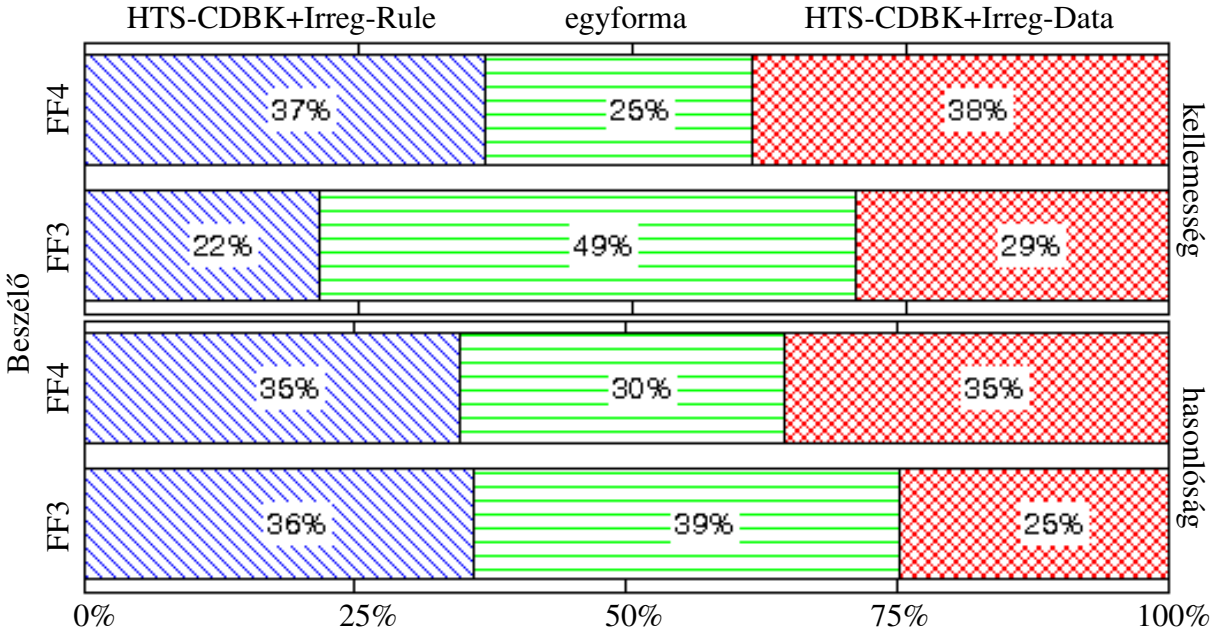
A 2. meghallgatásos tesztből azt a következtetést vonhatjuk le, hogy 1) a kísérleti alanyok az adatvezérelt modellt részesítették előnyben az alaprendszerrel szemben kellemesség és az eredeti beszélőre való hasonlóság szempontjából, 2) a HTS-CDBK+Irreg-Rule és HTS-CDBK+Irreg-Data rendszerek által szintetizált beszéd nem különbözik egymástól sem kellemesség, sem hasonlóság szerint.

5.2.5. Akusztikus elemzés

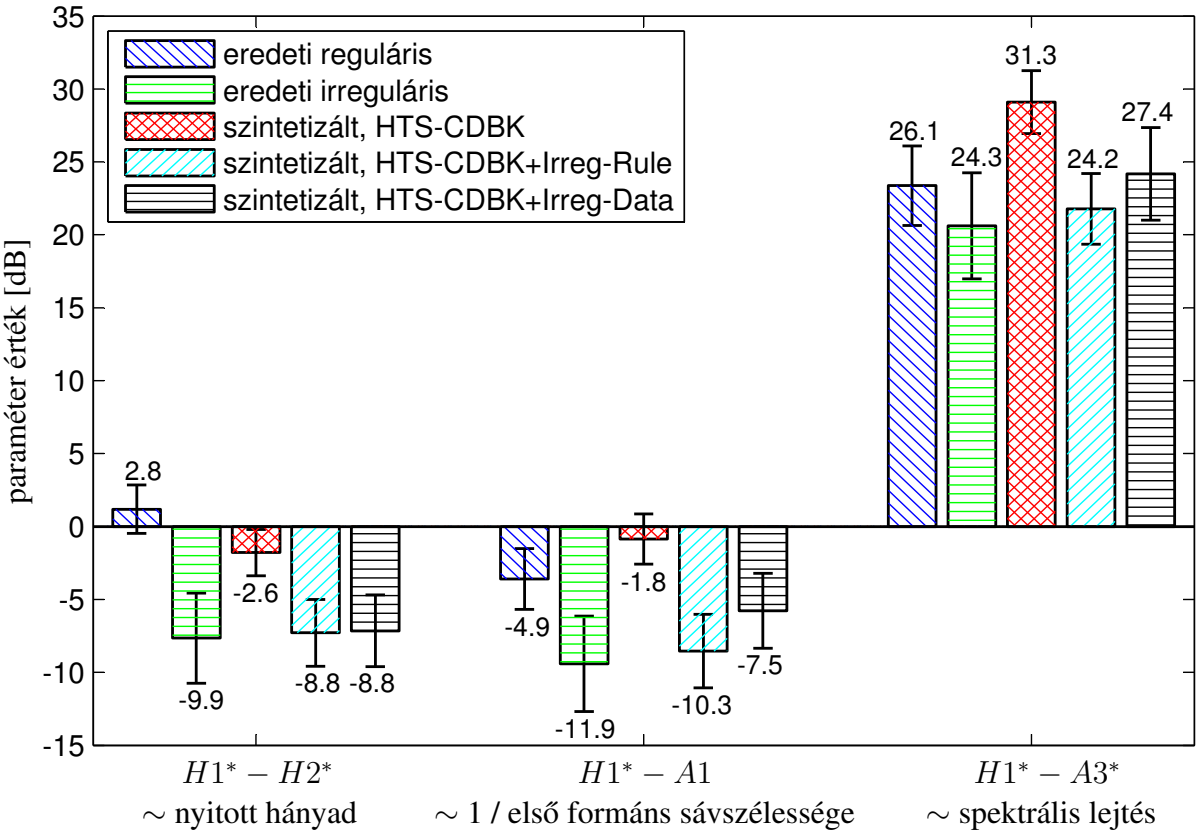
Az 1. és 2. meghallgatásos teszthez (5.2.2. és 5.2.4. fejezet) kiválasztott beszédmintákon akusztikus elemzést is végeztünk, a 4.2.3. fejezethez hasonló módon. Az elemzés célja az volt, hogy megállapítsuk, az irreguláris zönge modellekkel szintetizált beszédminták néhány releváns akusztikai jegy szempontjából közel vannak-e az eredeti glottalizált beszédhez. A szakirodalom



5.8. ábra. A HTS-CDBK alaprendszerrel és HTS-CDBK+Irreg-Data adatvezérelt irreguláris zönge modellekkel szintetizált szavak szubjektív összehasonlításának eredménye.



5.9. ábra. A HTS-CDBK-Irreg-Rule és HTS-CDBK+Irreg-Data irreguláris zönge modellekkel szintetizált szavak szubjektív összehasonlításának eredménye.



5.10. ábra. Az irreguláris zönge modellekkel szintetizált szavak akusztikus elemzésének eredménye. A függőleges fekete vonalak a 95%-os konfidenciaintervallumot jelölik.

alapján felhasználtunk három olyan akusztikai jegyet, amelyeket korábban irreguláris és reguláris beszéd megkülönböztetésére használtak [5, 62, 68]. Ezek alapján irreguláris beszédben a nyitott hányad (OQ) alacsonyabb; az első formáns sávszélessége ($B1$) nagyobb; a spektrum lejtése (TL) meredekebb.

Az elemzéseket a két beszélő 10-10 szintetizált szaván (HTS-CDBK alaprendszer, HTS-CDBK+Irreg-Rule és HTS-CDBK+Irreg-Data kiegészített rendszerek), és 10-10 másik, eredeti reguláris és eredeti irreguláris felvételen végeztük, összesen 100 szót vizsgálva. Az OQ helyett a $H1^* - H2^*$-ot mértük, az $1/B1$-et $H1^* - A1$ elemzésével vizsgáltuk, a TL akusztikai jegyet pedig $H1^* - A3^*$ alapján mértük a 4.2.3. fejezetben leírt és a 4.9. ábrán látható módon a magánhangzók három-három pontján. A HTS-CDBK alaprendszerben a vizsgált magánhangzók több esetben zöngétlen szakaszokat is tartalmaztak. Itt a $H1$ és $H2$ értéket a többi esethez hasonlóan az első két spektrális csúcsként mértük.

A három paraméter és öt beszédminta típus összehasonlítása az 5.10. ábrán látható. Az egytényezős ANOVA analízis szerint a beszédminta típusának szignifikáns hatása volt mindhárom paraméterre ($F(4, 295) = 11{,}89; 7{,}70; 4{,}49$; sorban; $p < 0{,}005$). A beszédminta típusok átlagos paramétereinek összehasonlítására Tukey-HSD post-hoc tesztet végeztünk.

A $H1^* - H2^*$ hasonló volt az alaprendszer és az eredeti reguláris rendszer esetén (nincs szignifikáns különbség, $p = 0{,}37$). Az érték közel azonos volt az eredeti irreguláris és a szintetizált irreguláris mintákban (nincs szignifikáns különbség, $p = 0{,}99$). Az ábrán is látható, hogy a reguláris változatok viszont szignifikánsan eltérnek az irreguláris változatoktól ($p < 0{,}0005$). Ez azt jelenti, hogy a nyitott hányad szempontjából az irreguláris modellekkel szintetizált minták közel állnak az eredeti glottalizált beszédmintákhoz és távol vannak az eredeti modális mintáktól. $H1^* - A1$ tekintetében a homogén rész-csoportok: eredeti reguláris és szintetizált alaprendszer vs. eredeti irreguláris és szintetizált szabály alapú irreguláris ($p < 0{,}05$). Az ábrán észrevehetőek az irreguláris zönge modellek által eredményezett trendek. Az első formáns sávszélessége, azaz $H1^* - A1$ szempontjából a szabály alapú irreguláris minták nagyon közel vannak az eredeti glottalizált változatokhoz. Az adatvezérelt modell $B1$ paraméterei az eredeti reguláris és irreguláris magánhangzók közöttiek. Ebben a kísérletben a $H1^* - A3^*$ érték nem segítette a beszédminták elkülönítését. Egyedül a szintetizált HTS-CDBK minták különböznek a többitől ($p < 0{,}05$). Ez annak az eredménye lehet, hogy a $H1$-et nem tudtuk pontosan mérni a zöngétlen szakaszokban. Az 5.10. ábra szerint a beszédmintákon mért spektrális lejtés paraméter nem mutat egyértelmű tendenciát.

Az akusztikus elemzésből azt a következtetést vonhatjuk le, hogy a vizsgált három jellemző közül kettő esetén a szintetizált változatok közel vannak az eredeti glottalizált beszédhez. Az adatvezérelt modell és az eredeti irreguláris minták között mért nagyobb akusztikai paraméter különbség magyarázata az lehet, hogy a HTS-CDBK+Irreg-Data mintákban csak a szintetizált magánhangzók kisebb része tartalmaz irreguláris részt (ld. 5.5. e és 5.5. f ábra), így a mérések a magánhangzó elején, közepén vagy végén nem mutatták meg a glottalizáció akusztikai korreláltjait.

5.3. Összegzés

A jelen fejezetben bemutatott új eredmények tézisszerű összefoglalása a 7. fejezetben található (*II. téziscsoport*).

Az 5.1. fejezetben ismertettük az I.1. tézisben kidolgozott újszerű, MGC maradékjel kódkönyv alapú gerjesztési modell statisztikai parametrikus beszédszintézisbe illesztését (*II.1. tézis*). Az új rendszert percepciós kísérletben vizsgálva igazoltuk, hogy a módszer az impulzus-zaj gerjesztéshez képest jobb minőségű beszéd szintézisére alkalmas. Az 5.2. fejezetben bemutattuk két új, alternatív irreguláris zönge modell kidolgozását. A szabály alapú és az adatvezérelt modellt is rejtett Markov-modell alapú beszédszintézisben alkalmaztuk, majd megmutattuk, hogy a módszerekkel szintetizált beszéd szignifikánsan kellemesebb és jobban hasonlít az eredeti beszélőre, mint az alaprendszer (*II.2. és II.3. tézisek*). Végül akusztikai elemzést végeztünk az irreguláris zönge szintézis módszerekkel, mely szerint mindkét modell közel áll az eredeti irreguláris beszédhez a vizsgált releváns akusztikai jegyek tekintetében (*II.4. tézis*).

Az 5. fejezetben ismertetett új módszerek és eredmények felhasználhatóak természetesebb, expresszív és személyre szabott beszédszintézis rendszerek kialakítására, azaz növelik a gépi beszédkeltés természetességét. Az alkalmazási lehetőségeket a 7. fejezet részletezi.

6. fejezet

Szubglottális rezonanciák elemzése a magyar beszédben

A beszédkeltés forrás-szűrő modellje [1], melyet a 4. fejezetben a gerjesztési modell kidolgozása során is alkalmaztunk, azt az egyszerűsítést használja, hogy a forrás és a szűrő tökéletesen szétválasztható. A valóságban azonban a forrás (gége) és a szűrő (artikulációs csatorna) között nemlineáris csatolás jöhet létre, melyet részben a szubglottális rendszer okoz. Emellett azt is kimutatták, hogy a szubglottális rezonanciák frekvenciájának környezete akusztikai szempontból előnytelen [6]. Az 1.5. fejezetben bemutattuk azt a feltételezést, mely szerint a beszéd képzése során próbáljuk elkerülni azokat az artikulációs helyzeteket, amikor a formánsok (az artikulációs csatorna rezonancia frekvenciái) és szubglottális rezonanciák (az alsó légúti tér rezonancia frekvenciái) között interakció léphetne fel. A formánsok értékei folyamatosan változnak beszéd közben, az SGR-ek azonban közel konstansak egy-egy beszélő esetén. A formánsok és a szubglottális rezonanciák között ugyan nincs közvetlen ok-okozati összefüggés, azonban a közöttük fennálló indirekt kapcsolat különböző magánhangzó csoportok elkülönüléséhez vezet, melyre a kvantális elmélet ad magyarázatot.

A kvantális elmélet [86] elvileg univerzálisan, nyelvektől függetlenül rendszert alkot a beszédhangok kategorizálására, azonban a gyakorlatban nem egyértelmű, hogy a szubglottális rezonanciák minden nyelven hozzájárulnak-e a beszédhangok elkülönítéséhez. A szubglottális rezonanciák és magánhangzó formánsok kapcsolatát korábban vizsgálták beszédprodukciós szempontból amerikai angol [7], spanyol [83], német [84] és koreai [85] nyelvre; magyarra azonban eddig nem voltak eredmények. Az eddigi eredmények szerint a fenti nyelvekben a szubglottális rezonanciák a formánsok szempontjából természetes elválasztó funkciót töltenek be és bizonyos megkülönböztető jegyeknek megfelelő kategóriákra osztják a magánhangzókat.

Ebben a fejezetben bemutatjuk a szubglottális rezonanciák vizsgálatára irányuló magyar nyelvre végzett elemzéseket és egy új, szubglottális rezonancia alapú magánhangzó osztályozó eljárást.

6.1. Kísérlet a szubglottális rezonanciák beszédre vonatkozó hatásának vizsgálatára

Első kísérletként magyar magánhangzókra vonatkozóan vizsgáltuk a szubglottális rendszer hatását[1]. Ehhez új logatom felvételeket rögzítettünk (3.2. fejezet), melyek alapján beszélőnként külön-külön és összevonva, normalizálással is végeztünk elemzéseket.

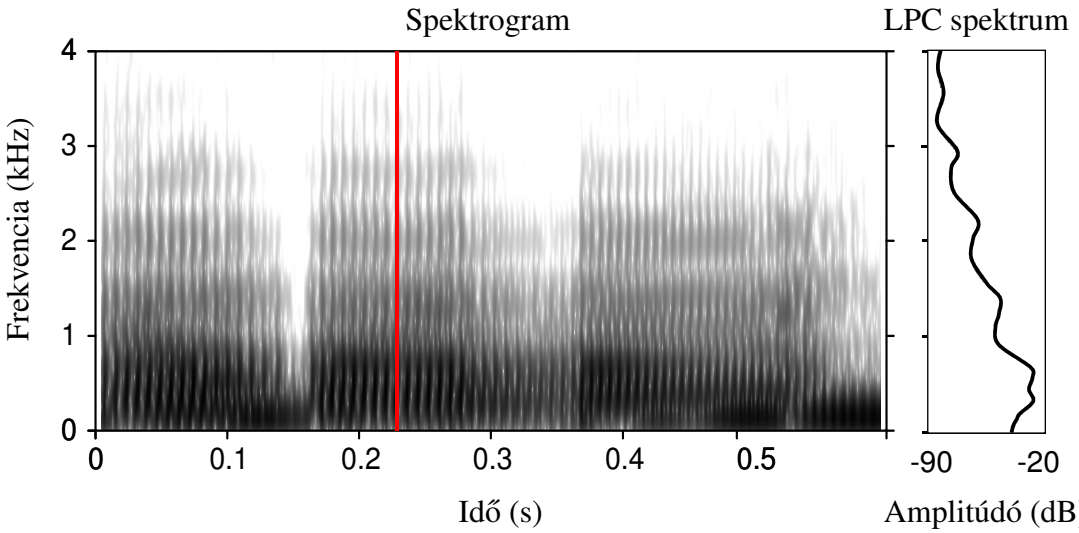
A négy beszélővel (Log_FF1, Log_FF2, Log_NO1 és Log_NO2) készült logatom felvételek mind a 14 magyar magánhangzót tartalmazták. Az első három formáns értékét ($F1$, $F2$ és $F3$) automatikusan mértük a beszédjelből a Praat programmal a vizsgálandó magánhangzók közepén, majd manuálisan javítottuk. A manuális javítások során a spektrogramról vizuális leolvasással határoztuk meg a formánsokat. Az első három szubglottális rezonanciát ($Sg1$, $Sg2$ és $Sg3$) manuálisan mértük a Wavesurfer programmal a gyorsulásmérő jelből minden beszélő és SGR esetén 25-25 ponton. Ez az eljárás hasonló a beszédjelből történő formáns méréshez, azaz a szubglottális rezonanciákat a gyorsulásmérő jel LPC spektrum burkolójának csúcsaiként mértük, amelyet a 6.1. ábra mutat be (részletes leírás a mérésről: [7, 88]). A mérés pontossága a szakirodalmi adatok szerint kb. 5 % [99]. Az ábra példáján az első három szubglottális rezonancia értéke kb. 590 Hz, 1350 Hz, 2150 Hz. Az is észrevehető, hogy az SGR értékek közel konstansak a teljes szóban.

A szubglottális rezonanciák mért értékeinek mediánjait használtuk fel a továbbiakban, melyet a 6.1. táblázat összegez. Eszerint a beszélők szubglottális rezonanciái a szakirodalmi adatok szerinti elvárt tartományban vannak.

6.1. táblázat. Négy beszélő logatom felvételein mért szubglottális rezonancia értékek mediánjai (Hz-ben).

	$Sg1$	$Sg2$	$Sg3$
Log_FF1	617	1301	2015
Log_FF2	577	1235	1974
Log_NO1	607	1478	2268
Log_NO2	662	1543	2426

[1] A többes szám a kutatásban részt vevő többi személyre utal. Bárkányi Zsuzsanna a fonológiai megkülönböztető jegyekkel és a kvantális elmélet vizsgálatával foglalkozott valamint a manuális mérésekben vett részt. Gráczi Tekla Etelka a felvételek készítésében, a manuális mérések és kézi javítások végzésében vett részt valamint a formánsok és a szubglottális rezonanciák közötti kapcsolatot vizsgálta. Bőhm Tamás az ROC elemzést kezdeményezte. Steven M. Lulich az eredeti amerikai angol kísérletek magyarázatában segített. Saját magam a felvételek készítésében, az automatikus formánsmérések illetve a manuális mérések és kézi javítások végzésében, a szubglottális rezonancia alapú formáns normalizálásban, valamint a formánsok és szubglottális rezonanciák közötti kapcsolat magyar nyelvre történő vizsgálatában vettem részt.



6.1. ábra. Szubglottális jel spektrogramja és LPC spektruma Log_FF2 beszélő „adaba" szava alapján. A jobb oldali spektrumot a bal oldali spektrogram függőleges vonallal jelölt pontján számítottuk.

6.1.1. A magyar magánhangzók rendszere szubglottális rezonanciák szempontjából

A magyar magánhangzók rendszerében 7 rövid és 7 hosszú magánhangzó található. Fonológiai szempontból ezek párokat alkotnak, melyet a 6.2. táblázat mutat be [117, 47. oldal] alapján. Megjegyezzük, hogy a rövid [ɛ] párja a hosszú [eː].

Fonetikai szempontból az [ɛ - eː] és az [ɔ - aː] magánhangzó párok minőségben és hosszúságban is különböznek. A minőségbeli különbséget az eltérő artikuláció okozza, melyet a 6.3. táblázat foglal össze [117, 44. oldal] alapján. Ezen kategóriák szerint a rövid [ɛ] párja a hosszú [eː] és a rövid [ɔ] párja a hosszú [aː].

A továbbiakban az artikulációs vetületnek megfelelő fonetikai szempontú felbontást használjuk néhány kategória összevonásával. Ezeket a kategória definíciókat a más nyelven végzett szubglottális rezonanciákat elemző kutatások alapján választottuk. Alsó nyelvállású magánhangzónak tekintjük az [ɔ, aː, ɛ] hangokat, míg a többi ([i, iː, eː, y, yː, ø, øː, u, uː, o, oː]) a nem-alsó kategóriába tartozik. A hátul képzett magánhangzók közé a [ɔ, o, oː, u, uː] tartoznak, az elöl képzett hangok pedig a [i, iː, ɛ, eː, y, yː, ø, øː]. A későbbiekben bizonyos esetekben külön kezeljük az elöl képzett, ajakréses, nem-alsó [eː, i, iː] hangokat a többitől.

6.1.2. Modell a szubglottális rezonanciák beszédre vonatkozó hatására

A mérések alapján a szubglottális rezonanciák magyar beszédre vonatkozó hatására akusztikai alapú modellt dolgoztunk ki. Az amerikai angol nyelvre kidolgozott modellt [7] alkalmaztuk a magyar nyelvre, és megállapítottuk, hogy

6.2. táblázat. A magyar magánhangzók fonológiai osztályozása.
Forrás: [117, 47. oldal] alapján.

	elöl képzett				hátul képzett	
	ajakréses		ajakkerekítéses			
	rövid	hosszú	rövid	hosszú	rövid	hosszú
felső nyelvállású	i	iː	y	yː	u	uː
középső nyelvállású		eː	ø	øː	o	oː
alsó nyelvállású	ɛ				ɔ	aː

6.3. táblázat. A magyar magánhangzók artikulációs tulajdonságai.
Forrás: [117, 44. oldal] alapján.

	elöl képzett				centrális		hátul képzett	
	ajakréses		ajakkerekítéses		ajakréses		ajakkerekítéses	
	rövid	hosszú	rövid	hosszú	rövid	hosszú	rövid	hosszú
legfelső nyelvállású	i	iː	y	yː			u	uː
felső nyelvállású		eː	ø	øː			o	oː
alsó nyelvállású	ɛ						ɔ	
legalsó nyelvállású						aː		

1) az első szubglottális rezonancia ($Sg1$) az első formáns ($F1$) tartományában az alsó és a nem-alsó nyelvállású magánhangzók között található,

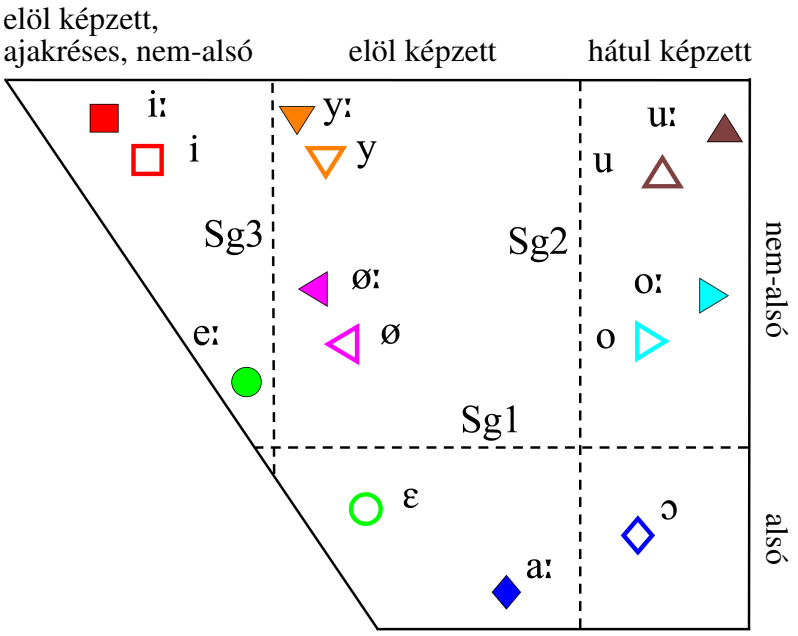
2) a második szubglottális rezonancia ($Sg2$) a második formáns ($F2$) tartományában az elöl és hátul képzett magánhangzók között található,

3) a harmadik szubglottális rezonancia ($Sg3$) a második formáns ($F2$) tartományában az elöl képzett, ajakréses, nem-alsó magánhangzókat választja el a többi elöl képzett magánhangzótól.

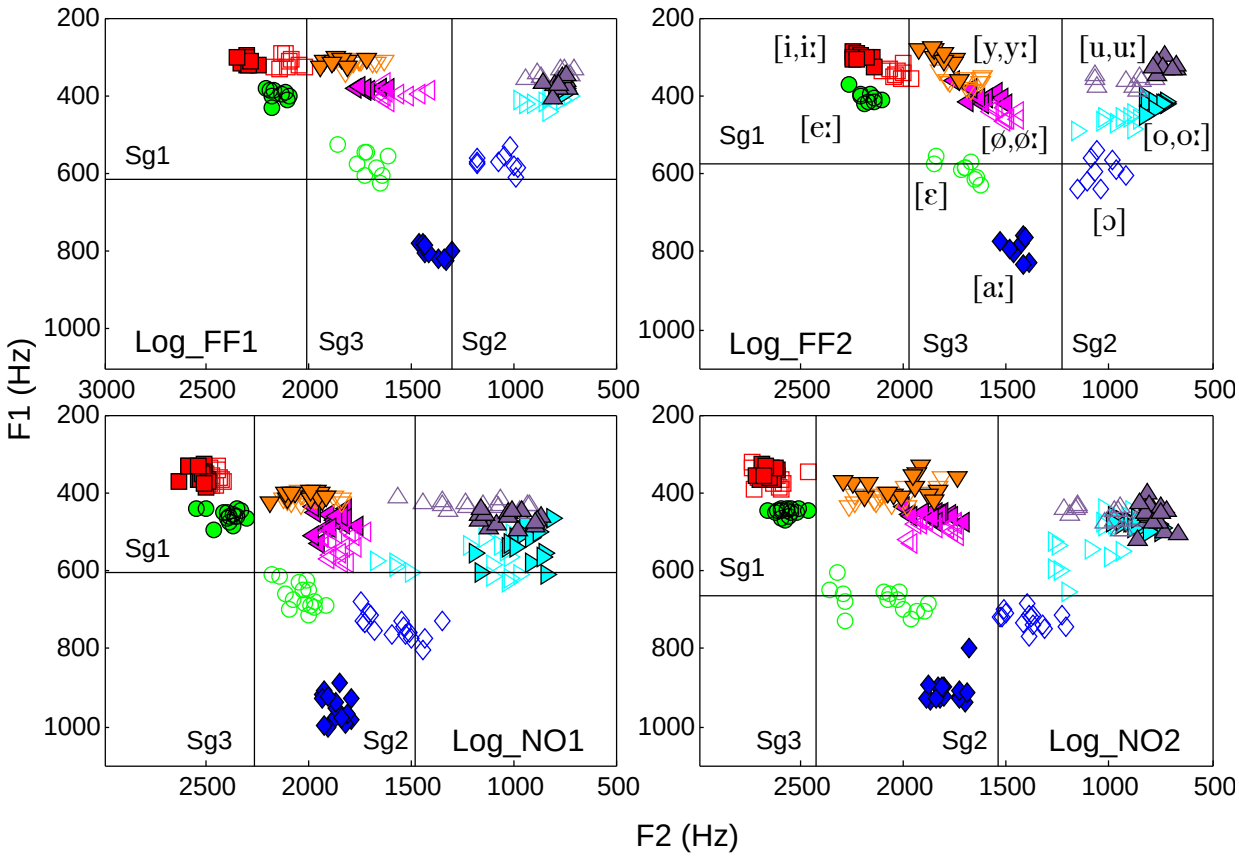
A modellben szereplő formánsok és szubglottális rezonanciák kapcsolatát a 6.2. ábra szemlélteti vizuálisan: a magánhangzók terében a függőleges irány az $F1$ változására, míg a vízszintes irány az $F2$ változására utal. Az ábra szerint az $Sg1$ az alsó [ɔ, aː, ɛ] magánhangzókat választja el a nem-alsóktól, az $Sg2$ a hátul képzett [ɔ, o, oː, u, uː] és az elöl képzett [i, iː, ɛ, eː, y, yː, ø, øː] magánhangzók között található, míg az $Sg3$ az [eː, i, iː] magánhangzókat választja el a többitől.

6.1.3. Beszélőnkénti elemzés

Az egyes beszélők magánhangzó formánstereit ($F1$ és $F2$ értékek magánhangzónként) a 6.3. ábra mutatja be az SGR értékeket is feltüntetve. Általánosságban elmondható, hogy a magánhangzó teret a szubglottális rezonanciák jól látható módon kategóriákra osztják, azaz teljesülnek a modell hipotézisei.



6.2. ábra. Magyar magánhangzók formánstere a szubglottális rezonanciákkal kiegészítve. A feltételezés szerint az SGR-ek különböző magánhangzó csoportokat választanak el egymástól.



6.3. ábra. Négy beszélő formánsainak és szubglottális rezonanciáinak kapcsolata logatom felvételek alapján. A vízszintes vonal az $Sg1$ értékét mutatja, a függőleges vonalak az $Sg2$ és $Sg3$ értékeit jelölik.

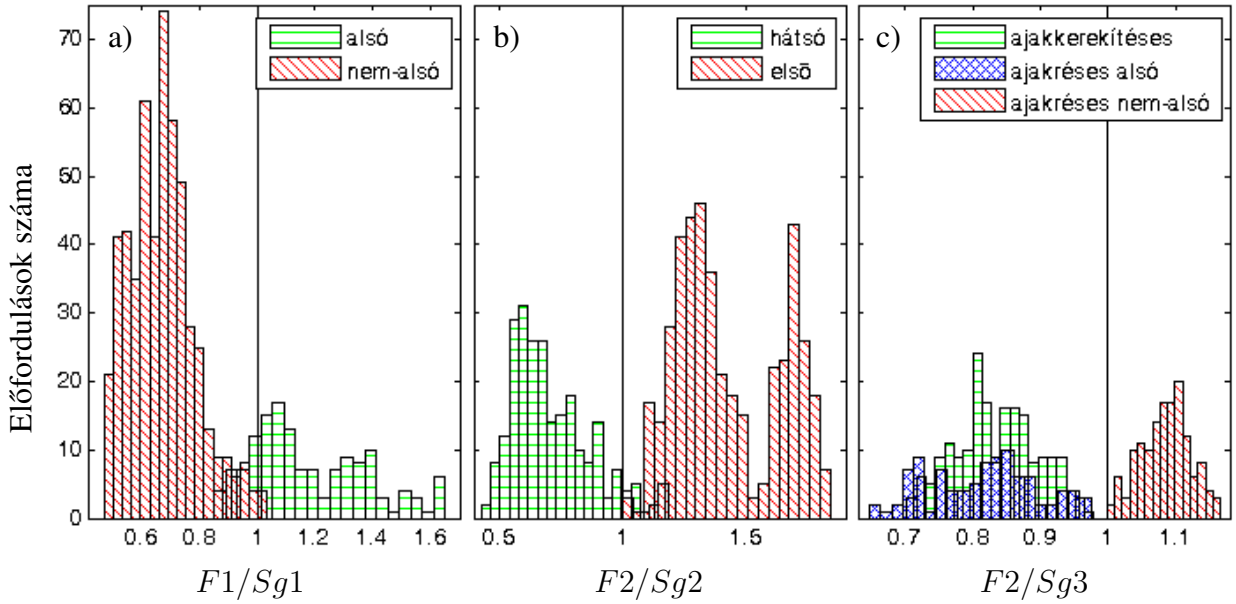
A formánstereket a 6.2. ábrával összehasonlítva észrevehető, hogy néhány kivétel előfordul. Log_FF1 beszélő esetén az [ɛ] és [ɔ] magánhangzók $F1$ értéke kisebb, mint $Sg1$, valamint Log_FF2 és Log_NO2 esetén az [ɛ] hang $F1$ formánsai nagyon közeliek $Sg1$-hez. Erre két lehetséges magyarázat van. Az első szerint a gyorsulásmérő jel rezonanciái nem annyira tiszták, mint a beszédben lévő formánsok. Ezt a szubglottális rendszer nagyobb csillapító hatása, valamint a nyakon lévő lágyszövetek aluláteresztő jellege okozza [6]. Mivel az első szubglottális rezonancia sokszor közel van az alapfrekvenciához, az $Sg1$ mérése az intenzív alsó harmonikusok miatt is nehézkes. Log_FF1 beszélő esetén valószínűsíthető, hogy az $Sg1$ mérése pontatlan volt a toldalékcső és a szubglottális rendszer közötti erős akusztikai csatolás miatt, mivel a mért $Sg1$ érték aránytalanul magas az $Sg2$ és $Sg3$ értékekhez képest. A második magyarázat az lehet, hogy még laboratóriumi beszédben is előfordul koartikuláció, így előfordulhat, hogy az [ɛ] és [ɔ] magánhangzók alacsony $F1$ értéke a mássalhangzókkal történő koartikuláció eredménye.

Az elöl és hátul képzett magánhangzókat vizsgálva az vehető észre a 6.3. ábrán, hogy az [ɔ] magánhangzó $F2$ értéke alacsonyabb $Sg2$-nél Log_FF1, Log_FF2, és Log_NO2 esetében, de Log_NO1-nél nem. Madsack és társai korábban azt találták, hogy a német nyelvben az alsó [a] magánhangzó a beszélőtől függően vagy kategorikusan az $Sg2$ alatt, vagy fölötte volt [84]. Koreai nyelven végzett kutatás alapján Jung és társai azt vették észre, hogy az [a] hang relatív $F2$ értéke a szomszédos mássalhangzó képzési helyétől függ. Előfordulhat, hogy az alsó magánhangzók formánsai hasonlóan változnak más nyelvekben, így a magyarban is.

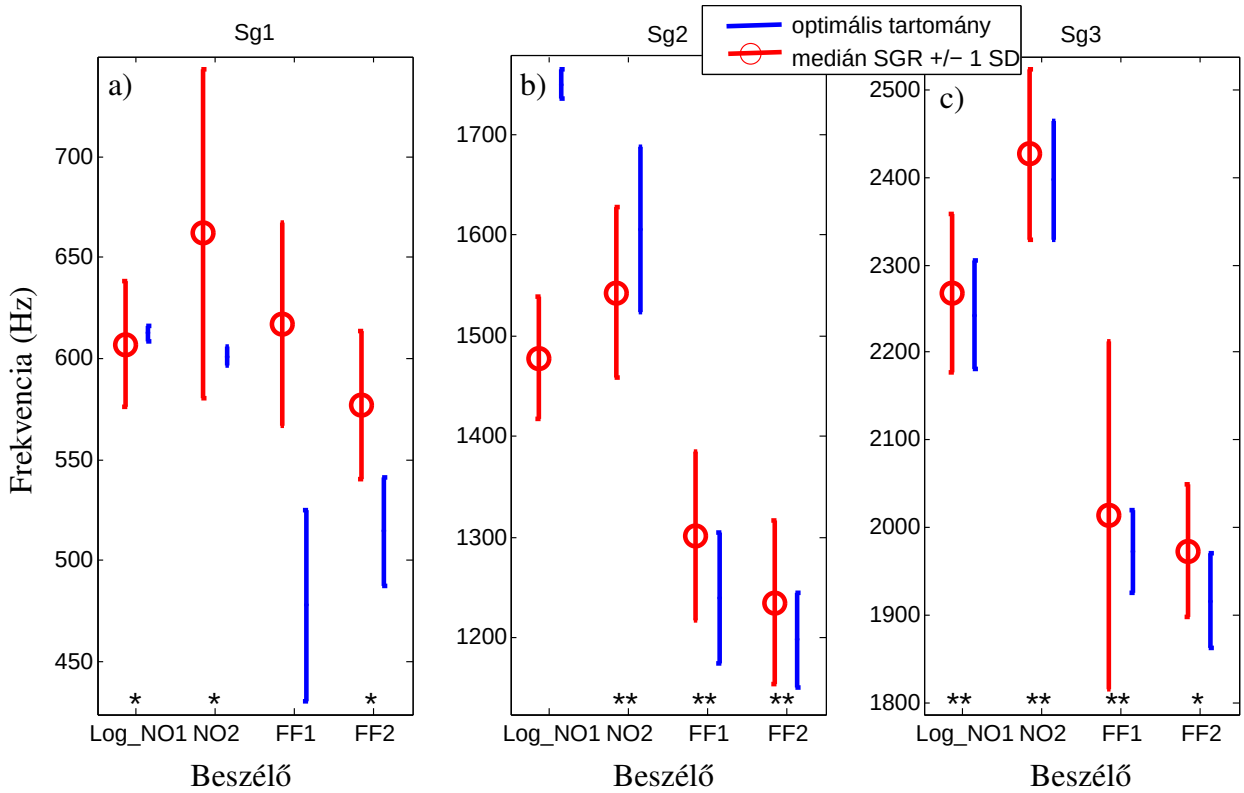
Az $Sg3$ elválasztó szerepe az ábra alapján egyértelmű, és nem láthatóak beszélőnkénti eltérések.

6.1.4. Normalizált elemzés

A fenti modellt mérnöki módszerekkel is igazoltuk: a magánhangzó formánsok normalizálásával az egyes formáns értékeket a beszélő megfelelő szubglottális rezonanciájával elosztottuk ($F1/Sg1$, $F2/Sg2$ és $F2/Sg3$), majd a beszélőnkénti adatokat összevontuk. A formáns normalizálásra más elvű (pl. matematikai statisztikai alapú [118, 119]) megoldás is létezik, azonban ezek nem veszik figyelembe a szubglottális rendszer hatását. A 6.4. ábra az SGR-normalizált formáns hisztogramokat mutatja: az a) ábra szerint az $Sg1$ függőleges vonal elválasztja az alsó és nem-alsó magánhangzókat. A b) ábrán az vehető észre, hogy az $Sg2$ az elöl képzett és a hátul képzett magánhangzókat közel optimálisan választja el. A c) ábrán $F2$ formáns $Sg3$ szerinti normalizált értékei láthatóak, amely alapján az elöl képzett, ajakréses, nem-alsó [eː, i, iː] magánhangzók elkülönülnek az ajakkerekítéses párjaiktól ([y, yː, ø, øː]).



6.4. ábra. Szubglottális rezonanciák szerint normalizált formáns hisztogramok logatom beszéd alapján: a) $F1/Sg1$, b) $F2/Sg2$, c) $F2/Sg3$ értékek összevonva az összes beszélőre. A függőleges vonal a normalizált $Sg1$, $Sg2$, $Sg3$ értéket jelöli.



6.5. ábra. ROC elemzés eredménye a szubglottális rezonanciák magánhangzó csoportokra elkülönítésének vizsgálatára: a) $Sg1$, b) $Sg2$, c) $Sg3$. A világos vonalak az SGR értékeket és egységnyi szórásukat mutatják, a sötét vonalak az optimális elválasztó tartományt jelölik. ** jelöli, ha az SGR az optimális elválasztási tartományon belül van. * jelöli, ha az SGR az optimális elválasztási tartomány egységnyi szórásán belül van.

6.1.5. Optimális kategóriahatárok vizsgálata

Az eddigi vizsgálatok szerint a fentiek nem teljesülnek minden beszélő és minden kategória esetén. A kategóriák optimális elválasztásának részletes vizsgálatára ROC (*Receiver Operating Characteristics*) elemzést végeztünk külön-külön minden SGR-re és beszélőre, amelynek eredménye a 6.5. ábrán látható. Az ROC analízis mindegyik esetben meghatározta azokat a frekvencia tartományokat, amelyek optimálisan elválasztják a különböző kategóriákat (az ábrán ezt sötét függőleges vonal jelöli). Az SGR-ek medián értékei (az ábrán világos kör) nagyrészt egységnyi szóráson belül vannak az optimális elválasztási tartományhoz képest. Az elemzés megmutatta, hogy a 12-ből 6 esetben az SGR az optimális elválasztási tartományon belül van (** az ábrán), 4 további esetben egységnyi szóráson belül található (* az ábrán), míg a maradék 2 esetben távolabb van. Ez utóbbi két eset a korábban [ɛ] és [ɔ] magánhangzóra ismertetett kivételek miatt jelentkezik az ROC elemzésben is.

Összefoglalva az eredményeket, a szubglottális rezonanciák közel optimálisan választják el egymástól az alsó vs. nem-alsó nyelvállású, elöl képzett vs. hátul képzett, illetve elöl képzett, ajakréses, nem-alsó nyelvállású vs. egyéb elöl képzett magánhangzókat a magyar nyelvben.

6.2. Automatikus, szubglottális rezonancia-normalizáció alapú magánhangzó osztályozó kidolgozása

Az előző fejezetben bemutatott formális, kvantitatív modell alkalmas gépi implementációra. Annak tesztelésére, hogy a szubglottális rezonanciák ismerete segítheti-e a beszédfeldolgozást, egy kísérletet terveztünk, amelyben megvizsgáljuk, hogy az SGR-ek felhasználásával pontosabb osztályozást tudunk-e végezni magánhangzókra, mint anélkül[2].

A kísérlet során a 3.2. fejezetben ismertetett spontán beszédanyagot használtuk fel. A magánhangzó-formánsokat ($F1$ és $F2$) automatikusan mértük a Praat programmal a magánhangzó közepén, majd vizuális elemzés alapján kézzel javítottuk a spektrogram alapján. Az $Sg1$, $Sg2$ és $Sg3$ értékeket külön olvasott beszéd felvételek alapján kézzel mértük (azaz a gyorsulásmérő által szolgáltatott jel elemzésével, az érték leolvasásával kaptuk) a Wavesurfer programban minden beszélő és SGR esetén 20-20 ponton, majd a mediánjaikat használtuk fel, melyet a 6.4. táblázat összegez.

[2]A többes szám a kutatásban részt vevő többi személyre utal. Gráczi Tekla Etelka és Bárkányi Zsuzsanna a manuális mérésekben és a kézi javításokban segített, valamint a formánsok és szubglottális rezonanciák közti kapcsolatot elemezte a spontán beszédanyag esetén. Beke András az osztályozó eljárás ötletével járult hozzá a kutatáshoz. Steven M. Lulich az amerikai angol kísérletek ismertetésével segített. Saját magam a manuális mérésekben, az automatikus formánsmérés kidolgozásában és a kézi javításokban vettem részt, valamint a szubglottális rezonanciákat használó, formáns normalizáláson alapuló osztályozó eljárást dolgoztam ki illetve hasonlítottam össze a referencia osztályozóval

6.4. táblázat. Hat beszélő olvasott beszéd felvételein mért szubglottális rezonancia értékek mediánjai (Hz-ben).

	$Sg1$	$Sg2$	$Sg3$
Spo_FF1	556	1392	2273
Spo_FF2	587	1326	2096
Spo_FF3	567	1326	2192
Spo_FF4	521	1402	2420
Spo_FF5	545	1299	2193
Spo_NO1	558	1532	2354

Az automatikus osztályozások során tanító és tesztelő adatként a fenti hat beszélő spontán beszéd felvételeiből származó 5948 magánhangzó formánsait használtuk fel, a beszélőket összevonva.

6.2.1. Döntési fa alapú referencia osztályozó

A kísérlet során referencia osztályozónak J4.8 típusú döntési fákat használtunk a Weka programban [108]. A döntési fára azért esett a választás, mert ez a C4.5 típusú, széles körben használt döntési fa továbbfejlesztett változata, és a legtöbb esetben közel optimális osztályozást eredményez. A három szubglottális rezonanciának és a 6.1. fejezetben ismertetett modellnek megfelelően három osztályozót készítettem, melyek bemenetei a magánhangzónkénti tiszta formáns értékek, kimenetei pedig a modell kategóriái:

a) bemenet: $F1$, kimenet: alsó – nem-alsó

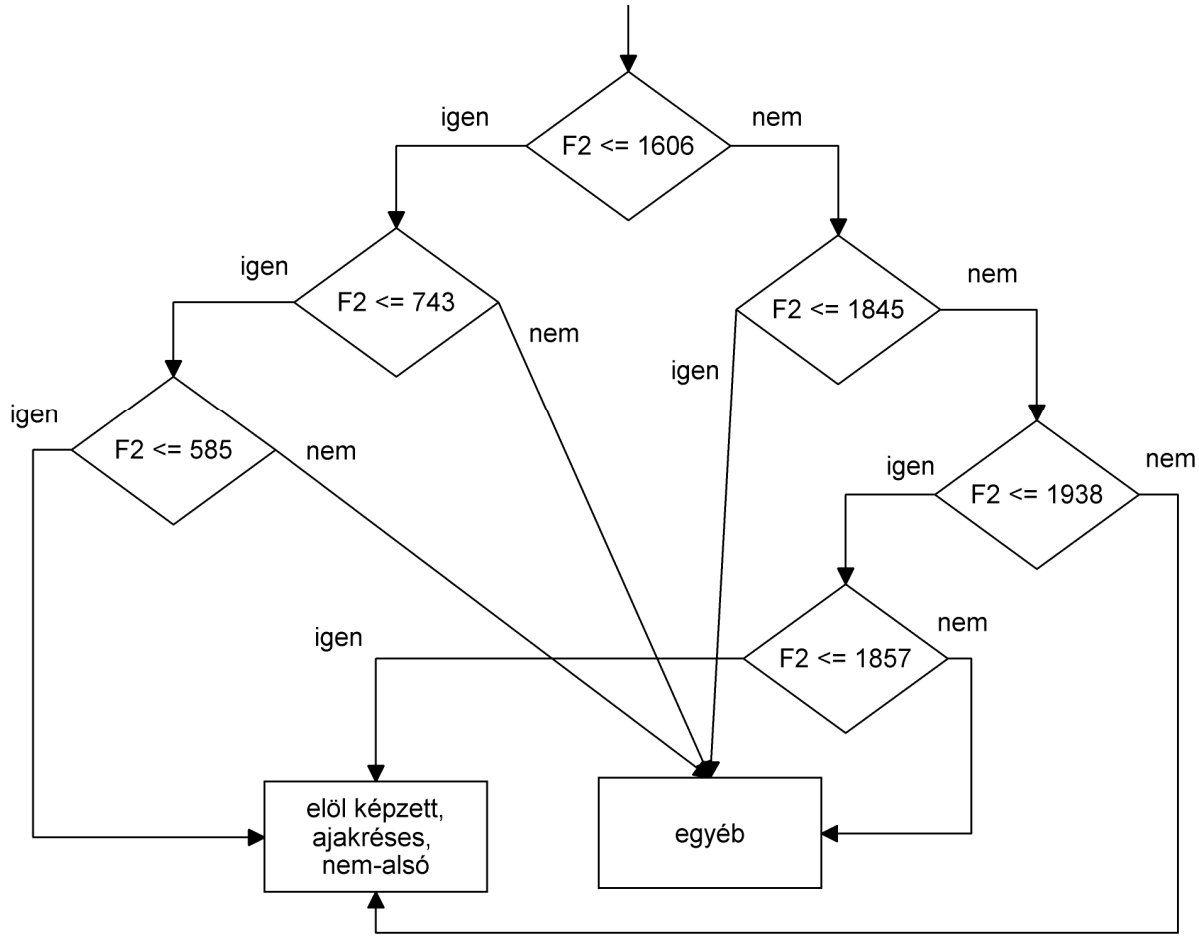
b) bemenet: $F2$, kimenet: elöl képzett – hátul képzett

c) bemenet: $F2$, kimenet: elöl képzett, ajakréses, nem-alsó – egyéb

A döntési fák építése során a később bemutatott módon a tanítóadat mennyiségét fokozatosan növeltük, aminek függvényében a fák felépítése is változott. A J4.8 típusú döntési fa egyik c) esetét a 6.6. ábra mutatja be. Az ábra szerint például egy [i] magánhangóra, aminek $F2$ értéke 2543 Hz, a kimenet a helyes elöl képzett, ajakréses, nem-alsó kategória. Ezzel szemben az $F2$ = 1901 Hz értékű [eː] hangra a döntési fa a helytelen egyéb kategória kimenetet adja.

6.2.2. Szubglottális rezonancia-normalizálás alapú osztályozó

A három szubglottális rezonanciának és a 6.1. fejezetben ismertetett modell három kategóriájának megfelelően SGR alapú formáns normalizálást használó osztályozókat készítettünk. Az osztályozók bemenete a magánhangzó $F1$ vagy $F2$ formánsának normalizált értéke, azaz a formánsfrekvencia elosztva a megfelelő szubglottális rezonancia frekvenciájával ($Sg1$, $Sg2$ vagy $Sg3$). Az osztályozók kimenetei a modellben ismertetett magánhangzó kategóriák:



6.6. ábra. Példa a formáns alapú döntési fa c) esetére. Az $F2$ érték ismeretében fentről elindulva a döntési fában lévő kérdések alapján eljuthatunk a lenti kategóriákig. A kérdéseknél a bal oldali nyíl az „igen” válasz, a jobb nyíl a „nem” válasz.

a) bemenet: $Fn1 = F1/Sg1$, kimenet: alsó – nem-alsó

b) bemenet: $Fn2 = F2/Sg2$, kimenet: elöl képzett – hátul képzett

c) bemenet: $Fn3 = F2/Sg3$, kimenet: elöl képzett, ajakréses, nem-alsó – egyéb

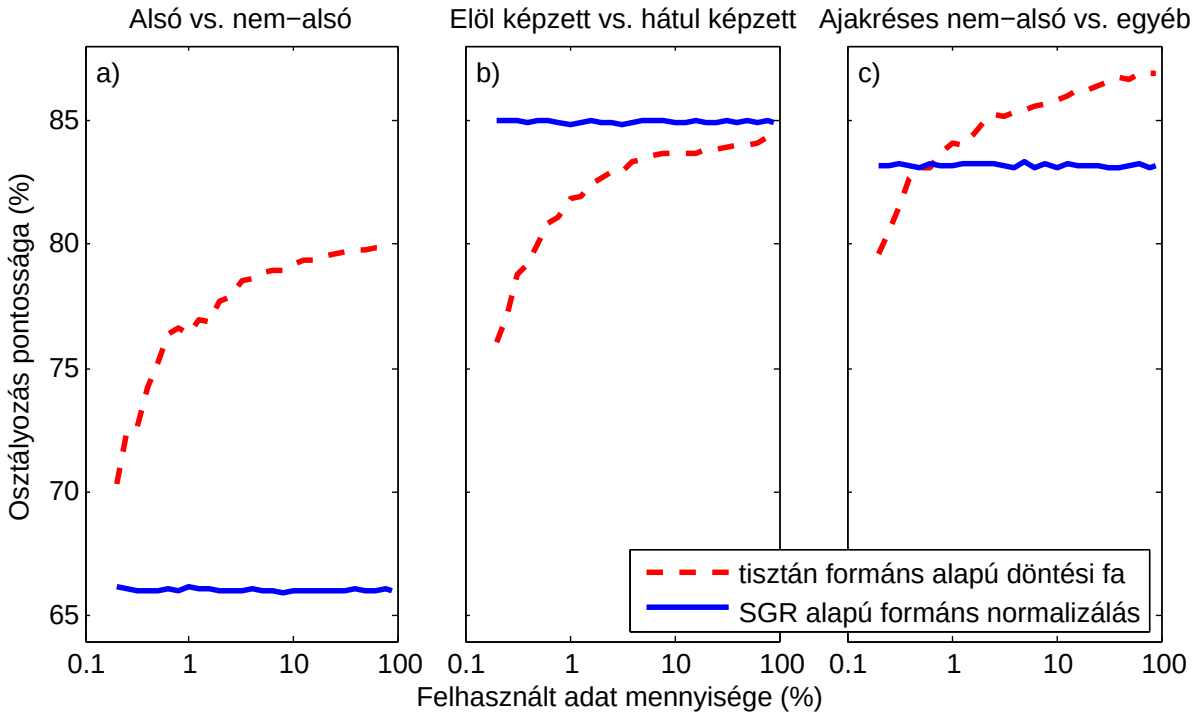
Az osztályozó egyszerű küszöbérték alapján működik: például a b) esetben amennyiben a bemeneti magánhangzóra vonatkozó $Fn2 \geq 1{,}0$ (pl. Spo_FF1 beszélő egyik [y:] magánhangzója esetén $F2$ = 1670 Hz és $Sg2$ = 1392 Hz), akkor elöl képzett kategóriára dönt, ha $Fn2 < 1{,}0$, akkor pedig hátul képzett kategóriára dönt az osztályozó. Az a) és c) esetben is az $Fn1$ és $Fn3$-ra vonatkozó küszöbértéknek 1,0-t választottam a modellnek megfelelően.

6.2.3. A két osztályozó összehasonlítása

A tiszta formáns bemenetet használó (SGR-ek ismerete nélküli) osztályozókat összehasonlítottuk a szubglottális rezonancia-normalizálás alapú osztályozók eredményével. A kísérletben azt vizsgáltuk, hogy a felhasznált adat mennyiségének függvényében melyik osztályozó teljesít jobban. A döntési fa alapú osztályozó esetén a tanítóadatot a teljes adat 0,2. . . 90 %-a között (12 – 5353 adatpont) változtattuk, és a maradék adatmennyiséget használtuk tesztelésre. Az SGR alapú osztályozó pontossága nem függ a tanító adat mennyiségétől, amennyiben a szubglottális rezonancia értékek meghatározásra kerültek. A szubglottális rezonancia alapú osztályozás esetén az adathalmaz 50 %-án végeztük a teszteket. Minden mérést 100 véletlen csoporton ismételtünk és az eredményeket átlagoltuk.

A kísérlet eredményeit a 6.7. ábra mutatja. Az a) esetben az $Sg1$ ismerete nem segítette az osztályozást. Ezt valószínűleg az okozta, hogy az $Sg1$ mérése sokszor nehézkes a gyorsulásmérő felvételből, mert az intenzív alsó harmonikusok torzíthatják a méréseket. Az $Sg1$ a korábbi logatomos kísérletben is számos kivételt okozott (6.1.3. fejezet), itt pedig olvasott beszéd során rögzített szubglottális jelből mértük az SGR-eket, ami nehezebb feladat. A b) ábrán az elöl képzett és hátul képzett magánhangzók elkülönítésének eredménye látható. Ebben az esetben az $Sg2$ ismerete egyértelműen javította az osztályozást: kevés adat esetén közel 20 %-kal, míg az adathalmaz jelentős részét felhasználva is 1 %-kal pontosabb a szubglottális rezonancia alapú osztályozó a tisztán formáns alapú döntési fához képest. Ez az eredmény megfelel a szakirodalom alapján elvártnak, mert a kutatások szerint a szubglottális rezonanciák közül az $Sg2$ hatása a legjelentősebb a magánhangzók kategóriákra osztásában [7]. A c) esetben az $Sg3$ alapú osztályozás pontosabb, amennyiben átlagosan az adathalmaz kevesebb, mint 1 %-át (50 magánhangzó) ismerjük.

A kísérlet alapján az $Sg2$ alapú osztályozás mindig, míg az $Sg3$ kevés tanító adat rendelkezésre állása esetén (50 magánhangzónál kevesebb adat) jobb eredményre vezet a döntési fa alapú referencia osztályozónál. Az SGR-normalizálás alapú osztályozás esetén elegendő körülbelül 10-20 magánhangzó tanító adatnak, melyek a szubglottális rezonanciák méréséhez szüksé-



6.7. ábra. A tisztán formáns alapú döntési fa és SGR-normalizált formáns alapú automatikus osztályozók pontosságának összehasonlítása a tanításhoz felhasznált adat mennyiségének függvényében: a) $Sg1$, b) $Sg2$, c) $Sg3$.

gesek. Az SGR alapú módszer tehát egyszerű, gyorsan adaptálódik a beszélőhöz, és elméletileg megalapozott, mivel a 6.1. fejezet modellje alapján működik. A tisztán formáns alapú módszer viszont érzékeny a tanítóminták jellegére és mennyiségére.

A fentiek során a beszédhangok formánsainak és a szubglottális rezonanciáknak az összefüggését vizsgáltuk beszédprodukcióban és automatikus osztályozás során, magyar nyelvre. Az elemzések és kísérletek szerint a szubglottális rezonanciák magyar nyelven is segítik a magánhangzók fonológiai megkülönböztető jegyek szerinti elkülönülését, így hozzájárulva a kvantális elmélet [86] szubglottális rezonanciákra vonatkozó kiegészítéséhez [7].

6.3. Összegzés

Ebben a fejezetben bemutattuk az alsó légúti rendszer rezonanciáinak hatását a magyar beszédre. Az új eredmények tézisszerű megfogalmazása a 7. fejezetben található (*III. téziscsoport*).

A 6.1. fejezetben kidolgoztunk egy modellt az első három szubglottális rezonancia és a magyar magánhangzók formánsainak kapcsolatára, melynek célja a szubglottális rendszer hatásának igazolása volt. Ezután elemzések során bemutattuk, hogy ezek a rezonanciák felhasználhatóak magyar beszédben magánhangzó osztályok formánsok szerinti elkülönítéséhez (*III.1. té-*

zis). Az elméleti modellt gépi implementációban is alkalmaztuk egy automatikus osztályozót megvalósítva, amely az osztályozási körülmények függvényében pontosabb lehet egy referencia osztályozónál (*III.2. tézis*).

A 6. fejezet új eredményei egyrészt hozzájárulnak az emberi beszédkeltés működésének megértéséhez: a gégében lévő forrásjel és az artikulációs csatorna mellett a szubglottális rendszer is hozzájárul a beszédhangok alakításához. Másrészt a modelleket felhasználva lehetőség van például beszédadatbázis javítására: amennyiben valamely realizált magánhangzó formánsa nem a modell szerinti kapcsolatban van a szubglottális rezonanciákkal, az várhatóan percepciós szempontból kevésbé előnyös (azaz nehezebben érthető beszédet jelent). Az ilyen beszédhangok tanítóadatbázisból történő kihagyása közvetve természetesebbé teheti a gépi beszédkeltést (részletesen ld. 7. fejezet).

7. fejezet

Összefoglalás és tézisek

Kutatói munkámat három fő részre osztottam, melyek egyrészt természetes beszéd analízisével és szintézisével, másrészt gépi beszédkeltéssel, harmadrészt az alsó légúti rendszer vizsgálatával foglalkoztak. Doktori értekezésem új eredményeit ennek megfelelően három téziscsoportban foglalom össze[1].

I. téziscsoport: Új, MGC maradékjel kódkönyv alapú gerjesztési modell kidolgozása és felhasználása irreguláris zöngeképzés javítására

A 4. fejezetben ismertettem egy új, MGC maradékjel kódkönyv alapú eljárást, amely a természetes beszéd paraméterekre bontására és abból történő visszaállítására alkalmas (4.1. fejezet). A paraméter reprezentáció lehetővé tette, hogy a természetes beszéd zöngeminőségét változtassam. Ezt felhasználva bemutattam egy irreguláris-reguláris transzformációs eljárást, amely a rekedtes, érdes, irreguláris zöngével képzett beszédet tudja javítani (4.2. fejezet).

I.1. tézis: [C3] Új, maradékjel kódkönyv elemkiválasztás alapú nyelvfüggetlen gerjesztési modellt dolgoztam ki, amely a beszédjel paraméterekre bontására (analízis) és visszaállítására (szintézis) alkalmas. A módszerben beszéd maradékjel halmaz alapján zöngeszinkron periódusokból álló kódkönyv készül, melyekből szintézis során automatikus elemkiválasztás határozza meg az összeillesztendő elemeket, célköltséget és összefűzési költséget felhasználva.

A kidolgozott új gerjesztési modell célja az volt, hogy egy olyan, beszédet paraméterekre bontó eljárást készítsek, amely gépi tanulásra alkalmas a rejtett Markov-modell alapú beszédszintetizátorban. Ehhez először tanulmányoztam a szakirodalomban rendelkezésre álló beszédkódoló algoritmusokat és gerjesztési modelleket. Megállapítottam, hogy a beszédkódolók legkorszerűbb technológiája, a CELP alapú kódoló ugyan jó minőségű szintetizált beszéd létrehozását eredményezné, de a kódoló belső paraméter reprezentációja közvetlenül nem alkalmas

[1] Ebben a fejezetben egyes szám első személyt használok saját munkám elkülönítésére.

a hagyományos HMM-ekkel történő modellezésre. Létrehoztam egy új, forrás-szűrő szétválasztáson alapuló gerjesztési modellt, amely a beszéd maradékjelének analízisén és szintézisén alapul és felhasználható a HMM alapú gépi tanításban.

A módszer az analízis lépésben (4.1.1. fejezet) először $F0$-detekciót és zöngehatár detekciót alkalmaz, majd inverz szűréssel előállítja a beszéd maradékjelét. A zöngés részeken a maradékjelet zöngeszinkron módon ablakozza, majd az ablakozott jelre három paramétert számít: $gain$, HNR és $rt0$. A $gain$ a keret RMS energiája, míg a HNR érték a keret harmonikus-zaj aránya. Az $rt0$ paraméter egy új reprezentáció az ablakozott maradékjel alakjának leírása a kiugró csúcsok egymástól való távolságának számításával. A két periódus hosszú ablakozott maradékjel elemekből egy kódkönyvet is építek. A maradékjel zöngétlen részein csak a $gain$ paramétert számítom.

A módszer a szintézis lépésben (4.1.2. fejezet) minden zöngés kerethez keres egy illeszkedő maradékjel elemet a kódkönyvből. Az elemkiválasztás során célköltséget és összefűzési költséget is számít. A célköltség a keret paraméterei és a kódkönyvbeli elemek paraméterei között számított RMS különbség. Az összefűzési költség a maradékjel kódkönyvben található elemek normalizált változatának RMSE távolsága. Az elemkiválasztás eredményeként kapott kódkönyv elemeket átlapolt módon összeadom. A zöngétlen kereteket véletlen zajként állítom elő. A maradékjel keretek energiáját a $gain$ paraméterrel szorozva skálázom, majd spektrális szűréssel visszaállítom a beszédet.

Az eljárás nem tartalmaz nyelvfüggő elemeket, ezért tetszőleges nyelvű beszéd analízisére és szintézisére alkalmas. A korábbi módszerektől az elemkiválasztás megvalósításában és az alkalmazott paraméter reprezentációban különbözik. A DSM eljárás is maradékjel kódkönyv alapú, azonban ez nem alkalmaz összefűzési költséget az elemkiválasztás során [52]. A GlottHMM rendszerben alkalmaznak ugyan célköltséget és összefűzési költséget is, de ez glottális forrásjel szintjén történik [39]. Az általam bemutatott $rt0$ paramétert egyik korábbi gerjesztési modellben sem használták.

I.2. tézis: [C1, C5] Nyelvfüggetlen eljárást dolgoztam ki irreguláris zöngével képzett beszéd regulárissá alakítására az I.1. tézisben kidolgozott modell felhasználásával. Percepciós teszttel kimutattam magyar mintákon, hogy az irreguláris-reguláris transzformáció után a beszéd szignifikánsan kevésbé érdes, mint az eredeti irreguláris beszéd.

Az I.1. tézis új gerjesztési modelljének segítségével a természetes beszéd bizonyos tulajdonságait meg lehet változtatni. Az analízis lépésben kapott paraméterek módosításával a szintézis lépésben előállított beszédminta jellegzetességei is változnak. Ezt kihasználva kidolgoztam egy új transzformációs eljárást, amelynek célja az irreguláris zöngével képzett beszéd regulárissá alakítása. Az irreguláris zöngeképzés (glottalizáció) egy természetes jelenség, amely az egymás

utáni zöngeperiódusok hirtelen amplitúdó változását eredményezheti. Az irreguláris zöngével képzett beszéd mindennapi kommunikációban gyakran előfordul és nem zavaró, azonban a gépi beszédfeldolgozó algoritmusok működését negatívan befolyásolhatja.

A transzformációs módszer (4.2.1. fejezet) analízis lépése megegyezik az I.1. tézis gerjesztési modellével. A maradékjel paramétereinek számítása után $F0$ interpolációt, majd kézi alapfrekvencia javítást, $gain$ simítást és spektrális simítást is végez a módszer. Az $F0$ javításra azért van szükség, mert az alapfrekvenciát a glottalizált szakaszokban a detektorok sok esetben hibásan mérik. A $gain$ és a spektrális paraméterekben az irreguláris fonáció kis perturbációkat okoz, melyek a simítással eltüntethetőek. A megváltoztatott paraméterekből az I.1. tézis szintézis lépése állítja vissza a beszédet.

Az irreguláris-reguláris transzformáció működését percepciós tesztben ellenőriztem magyar mintákon (4.2.2. fejezet). Eszerint négy beszélő transzformált mintáit a kísérleti alanyok szignifikánsan kevésbé érezték rekedtesnek, mint az eredeti glottalizált mintákat. Az eredeti beszéd természetességét két beszélő esetén tudta megtartani a transzformáció, aminek oka az lehet, hogy a glottalizáció különböző megjelenési formáit nem egyformán kezeli az algoritmus.

Az eljárás nem tartalmaz nyelvfüggő elemeket, ezért tetszőleges nyelvű irreguláris beszéd regulárissá transzformálására alkalmas. Reguláris-irreguláris transzformációra számos módszert készítettek már (pl. [19, 62, 68]), de a glottalizált beszéd javítására nem találtam korábbi megoldást a szakirodalomban.

I.3. tézis: [C1, C5] Kísérleti úton igazoltam magyar mintákon, hogy az I.2. tézis eljárása a beszéd több releváns akusztikai paraméterét (nyitott hányad, első formáns sávszélessége, spektrális lejtés) az irreguláris-reguláris transzformáció során a reguláris zöngeképzésre jellemző értékek irányába módosítja.

Az I.2. tézis eredményeit egy akusztikai kísérletben is vizsgáltam (4.2.3. fejezet). Korábban kimutatták, hogy az irreguláris és a reguláris beszéd jól megkülönböztethető néhány akusztikai paraméter vizsgálatával. Glottalizált beszédben a nyitott hányad nagyobb, az első formáns sávszélessége alacsonyabb, a spektrum lejtése pedig meredekebb, mint modális beszéd esetén [5, 62, 68].

Négy magyar anyanyelvű beszélő eredeti reguláris, eredeti irreguláris és irregulárisból regulárissá transzformált mintáit elemeztem a fenti három paraméter szerint. Az eredmények alapján a transzformált minták szignifikánsan különböznek az eredeti irreguláris beszédtől, és nem különböznek az eredeti reguláris beszédtől. A transzformációs eljárás a vizsgált akusztikai paraméterek tekintetében tehát modális beszédre jellemző irányba módosította a beszédmintákat.

II. téziscsoport: Az új gerjesztési modell illesztése gépi szövegfelolvasóhoz és felhasználása irreguláris beszéd szintézisére

A következő szakaszban (5. fejezet) először az analízis-szintézis gerjesztési modellt statisztikai parametrikus beszédszintézisbe illesztettem (5.1. fejezet). Az irreguláris fonáció modellezésére kidolgoztam két alternatív glottalizáció modellt rejtett Markov-modell alapú beszédszintézisben, melyek szintén a fenti analízis-szintézis paraméter reprezentáción alapulnak (5.2. fejezet).

II.1. tézis: [J2] Rejtett Markov-modell alapú gépi szövegfelolvasó rendszerhez illesztettem az I.1. tézisben ismertetett nyelvfüggetlen gerjesztési modellt. Percepciós teszttel igazoltam magyar mintákon, hogy a módszerrel előállítható beszéd szignifikánsan jobb minőségű az impulzus-zaj gerjesztésű gépi szövegfelolvasóhoz képest.

Az I.1. tézisbeli gerjesztési modell kidolgozásának célja többek között az volt, hogy ezt a rejtett Markov-modell alapú szövegfelolvasóhoz illesztve javítani lehessen a gépi beszéd természetességét.

Az eljárás HMM-TTS-be illesztése során a paramétereket a gépi tanulás igényeihez alakítottam (HTS-CDBK, 5.1.2. fejezet). A paramétereket logaritmizáltam, majd a $gain$ és spektrális paramétereket hagyományos HMM-ekkel, míg az $F0$, HNR és $rt0$ paraméterfolyamokat MSD-HMM-ekkel modelleztem. A szintézis lépést kiegészítettem fehérzaj hozzáadásával a felsőbb frekvencia komponensekben.

A szintézis eredményét percepciós kísérletben vizsgáltam (5.1.3. fejezet). Egy magyar anyanyelvű beszélő beszédkorpuszából előállítottam a fenti paramétereket, majd a sikeres gépi tanítás után mintamondatokat szintetizáltam. A HTS-CDBK rendszert a referencia impulzus-zaj gerjesztésű rendszerrel hasonlítottam össze, mely szerint az általam javasolt módszer mintái szignifikánsan jobb minőségűek az impulzus-zaj gerjesztési modellhez képest.

A HTS-CDBK rendszer jelen változata magyar nyelvű szövegfelolvasásra alkalmas, de a gerjesztési modellt könnyen lehet illeszteni más nyelvű HMM-TTS-hez is, mivel a gerjesztés nyelvfüggetlen.

II.2. tézis: [C2, J1] Kidolgoztam egy nyelvfüggetlen szabály alapú irreguláris zöngeképzés modellt és illesztettem ezt a II.1. tézisben ismertetett gépi szövegfelolvasóhoz. A modell alapfrekvencia felezést, maradékjel periódus amplitúdó skálázást és spektrális torzítást alkalmaz. Percepciós teszttel igazoltam magyar mintákon, hogy a kiegészített rendszerrel szintetizált beszéd szignifikánsan preferáltabb és jobban emlékeztet az eredeti beszélőre, mint a II.1. tézis rendszere.

A II.1. tézis szintézis lépését kiegészítettem a paraméterek automatikus változtatásával, hogy az eljárás az irregulárisra emlékeztető beszéd szintézisére alkalmas legyen (HTS-CDBK+Irreg-Rule, 5.2.1. fejezet). Ehhez három fő lépést alkalmazok: $F0$ felezés, maradékjel periódus amplitúdó skálázás és spektrális torzítás. Az $F0$ felezés célja a glottalizáció egyik jellegzetességének, az extrém alacsony alapfrekvenciának modellezése. A periódusok amplitúdó skálázása és a spektrális torzítás az eredeti glottalizált beszédhez hasonló amplitúdó ingadozásokat eredményeznek a szintetizált beszédben.

A szabály alapú irreguláris zönge modell eredményét percepciós tesztben vizsgáltam (5.2.2. fejezet). A kísérlet során két magyar beszélő mintáit értékelték a tesztelők természetesség és az eredeti beszélőre való hasonlóság szerint. A HTS-CDBK+Irreg-Rule rendszer mindkét szempontból szignifikánsan preferáltabb (azaz közelebb áll az eredeti irreguláris beszédhez), mint a HTS-CDBK alaprendszer.

II.3. tézis: [J1] Kidolgoztam egy nyelvfüggetlen adatvezérelt irreguláris zöngeképzés modellt és illesztettem ezt a II.1. tézisben ismertetett gépi szövegfelolvasóhoz. A modell irreguláris beszédrészletek maradékjeléből épített korpuszból elemkiválasztással keresi meg a szintézis során a megfelelő elemeket. Percepciós teszttel igazoltam magyar mintákon, hogy a kiegészített rendszerrel szintetizált beszéd szignifikánsan preferáltabb és jobban emlékeztet az eredeti beszélőre, mint a II.1. tézis rendszere.

A II.1. tézis módszerét kiegészítettem egy olyan irreguláris maradékjelekből álló korpusszal, amelyből irregulárisra emlékeztető beszéd szintetizálható (HTS-CDBK+Irreg-Data, 5.2.3. fejezet). Az analízis lépés során a paraméterek számítása mellett összegyűjtöm egy korpuszba azokat a magánhangzó-hosszúságú maradékjel szakaszokat, amelyek glottalizált módon jöttek létre. Az irreguláris zönge szintézisekor a maradékjelet ebből a korpuszból választja az automatikus elemkiválasztó eljárás, csak célköltséget felhasználva. A kiválasztott szakaszt a szintetizált maradékjel többi részéhez illesztem.

Meghallgatásos tesztet készítettem az adatvezérelt irreguláris zönge modell eredményének vizsgálatára (5.2.4. fejezet). A kísérleti alanyok magyar anyanyelvű mintákon a kiegészített HTS-CDBK+Irreg-Data rendszert szignifikánsan kellemesebbnek és az eredeti beszélőre jobban hasonlítónak ítélték meg, mint a HTS-CDBK alaprendszert.

A II.2 és II.3. tézisek irreguláris zönge modelljei nyelvfüggetlenek, a szabályok valamint a glottalizált korpusz készítés könnyen alkalmazhatóak más nyelvekre is. Kutatásom kezdete óta néhány más módszer is foglalkozik irreguláris beszéd szintézisével a rejtett Markov-modell alapú rendszerben [70, 71, 78, 79]. A különböző megoldások összehasonlító elemzése eddig nem történt meg.

II.4. tézis: [J1] Kísérleti úton igazoltam magyar mintákon, hogy a II.2 és II.3. tézisek eljárásai beszédszintézis során a beszéd több releváns akusztikai paraméterét (nyitott hányad: II.2 és II.3, első formáns sávszélessége: II.2) az irreguláris zöngeképzésre jellemző módon modellezik.

A II.2. és II.3. tézisek eredményeit egy akusztikai kísérletben vizsgáltam (5.2.5. fejezet). Az I.3. tézisben is alkalmazott nyitott hányad, első formáns sávszélesség és spektrális lejtés paramétereket elemeztem.

A két magyar beszélő eredeti reguláris, eredeti irreguláris és szintetizált beszédmintáin az akusztikus elemzés szignifikáns eredményeket mutatott ki a fenti paraméterekben. Az eredeti és szintetizált irreguláris minták nem különböznek szignifikánsan a nyitott hányad szempontjából. Az első formáns sávszélessége szerint a szabály alapú irreguláris zönge modell eredménye közel áll a glottalizált beszédhez. Az akusztikus elemzés tehát azt mutatja, hogy a HTS-CDBK+Irreg-Rule és HTS-CDBK+Irreg-Data modellek jól modellezik az irreguláris zöngét.

III. téziscsoport: Szubglottális rezonanciák elemzése a magyar beszédben

Az I. és II. téziscsoportokban kidolgozott és alkalmazott új gerjesztési modell a forrás-szűrő szétválasztáson alapul, amely nem veszi figyelembe a szubglottális rendszer hatását. A 6. fejezetben az emberi beszéd működésének vizsgálata során elemeztem az artikulációs csatorna és a szubglottális rendszer kölcsönhatását. Modellt dolgoztam ki a magyar magánhangzók formánsai (az artikulációs csatorna rezonanciái) és a szubglottális rezonanciák (a szubglottális rendszer rezonanciái) indirekt kapcsolatának jellemzésére, melyet egy automatikus gépi osztályozóban alkalmaztam.

III.1. tézis: [C4, J4] Modellt dolgoztam ki az alsó légúti (szubglottális) rendszer rezonanciáinak magyar beszédre vonatkozó hatására. Kimutattam, hogy a szubglottális rezonanciák (az alsó légúti rendszer első három rezonanciafrekvenciája) magyar beszédben felhasználhatóak magánhangzó osztályok formánsok szerinti elkülönítéséhez a szubglottális rezonanciák és formánsok közti indirekt kapcsolatot kihasználva.

A magyar magánhangzók első két formánsa ($F1$ és $F2$) és az első három szubglottális rezonancia ($Sg1$, $Sg2$ és $Sg3$) alapján a következő összefüggéseket alkottam meg:

1) az $Sg1$ az $F1$ tartományában az alsó és a nem-alsó nyelvállású magánhangzók között van,

2) az $Sg2$ az $F2$ tartományában az elöl és hátul képzett magánhangzók között található,

3) az $Sg3$ az $F2$ tartományában az elöl képzett, ajakréses, nem-alsó magánhangzókat választja el a többi elöl képzett magánhangzótól.

A modell működését vizsgáltam beszélőnként külön-külön, az adatokat normalizálással összevonva és ROC elemzés keretében is. A kísérletek szerint a szubglottális rezonanciák közel optimálisan választják el egymástól az alsó vs. nem-alsó nyelvállású, elöl képzett vs. hátul képzett, illetve elöl képzett, ajakréses, nem-alsó nyelvállású vs. egyéb elöl képzett magánhangzókat a magyar nyelvben.

A magyar nyelvre megalkotott modell hasonlít a korábban amerikai angolra és koreaira elkészítettre. Lulich azt találta, hogy angolban az $Sg2$ az elöl és hátul képzett magánhangzók között található [7], míg Jung az $Sg1$ és alsó vs. nem-alsó magánhangzók közti összefüggést vizsgálta angolban és az $Sg2$ hatását elemezte koreaiban [85]. Néhány más nyelvben is vizsgálták a szuglottális rezonanciák hatását (pl. német [84], spanyol [83]), azokban ezen kezdeti eredmények alapján nem hoztak létre egyértelmű modelleket. Az $Sg3$ szerinti magánhangzó kategorizálást korábban nem foglalták modellbe a fenti módon.

III.2. tézis: [J4] Automatikus osztályozót készítettem, mely egy beszélő magánhangzó formánsainak és szubglottális rezonanciáinak indirekt kapcsolatán alapulva normalizálásával a magánhangzókat a III.1. tézisben ismertetett kategóriákba sorolja. Megmutattam, hogy a vizsgált mintákon az Sg2 alapú módszer mindig pontosabb, az Sg3 alapú módszer kis tanítóadatmennyiség esetén pontosabb, míg az Sg1 alapú módszer nem pontosabb mint egy tisztán formánsokat felhasználó döntési fa alapú referencia osztályozó.

A III.1. tézis modelljét felhasználva egy automatikus magánhangzó osztályozót készítettem, amely a beszédhang formánsait a beszélő szubglottális rezonanciáival normalizálva ($F1/Sg1$, $F2/Sg2$ és $F2/Sg3$) a bemeneti beszédhangot a modellnek megfelelő kategóriába sorolja. Az $Sg1$ alapú módszer nem pontosabb, az $Sg2$ alapú módszer minden vizsgált esetben pontosabb a referencia döntési fa alapú osztályozónál. Az $Sg3$ alapú osztályozó kevés adat rendelkezésre állása esetén hasznos; egy-egy beszélőtől származó néhány magánhangzót használva magasabb pontosságot tudtam elérni a referenciához képest.

A szubglottális rezonanciákat korábban már sikerrel alkalmazták automatikus beszélő normalizálásban [83] és a beszélő magasságának becslésére is [99]. Lulich és Chen készített egy olyan osztályozó eljárást, amely az $F2$ és az $Sg2$ viszonya alapján mássalhangzó-magánhangzó hangkapcsolatokat tudott kategorizálni [93, 94]. Az általam bemutatotthoz hasonló, magánhangzókat artikulációs hely szerinti kategóriákba soroló szubglottális rezonancia alapú osztályozót nem találtam a szakirodalomban.

7.1. Az eredmények alkalmazhatósága

Kutatásom eredményei számos beszédtechnológiai alkalmazásban felhasználhatóak, amelyek egyrészt hozzájárulhatnak a természetesebb ember-gép kommunikációhoz, másrészt segíthetnek megérteni az emberi beszédképzés működését. Eljárásaimat magyar nyelvű mintákon teszteltem. Az I. és II. téziscsoportban alkalmazott módszerek nyelvfüggetlenek, így a modellek kiterjeszthetőek más nyelvekre is. Az alábbiakban téziscsoportonként bemutatok néhány alkalmazási lehetőséget.

Az I.1. tézisben ismertetett maradékjelen alapuló analízis-szintézis gerjesztési modell alkalmas különböző zöngeminőségek gépi előállítására és transzformációjára. Előzetes kísérleteim szerint levegősből modális beszéd átalakítására is megfelelő lehet a módszer. Az I.2. tézis glottalizáció javító eljárását ki lehet terjeszteni hosszabb beszédszakaszokra is, amivel rekedtes, patologikus hangokat várhatóan szebbé, kellemesebbé lehet tenni (pl. színészek, bemondók hangja). Az irreguláris-reguláris átalakító eljárás automatikussá kiegészített változatával beszédadatbázisokból el lehetne tüntetni az irreguláris zöngéjű szakaszokat, ezáltal ideálisabbá téve a beszédet a további feldolgozás céljából.

A II.1. tézisben bemutatott beszédszintetizátor rendszer javíthatja a korlátozott erőforrású eszközökben (pl. okostelefon) alkalmazott gépi szövegfelolvasás minőségét. A kevés erőforrás miatt bonyolultabb gerjesztési modellek nehézkesen kezelhetőek, viszont a tézis modellje várhatóan bizonyos korlátozott erőforrású eszközökön képes valós idejű működésre. A II.2. és II.3. tézisek irreguláris zönge modelljei hozzájárulhatnak a természetesebb, expresszív és személyre szabott beszédszintézishez. A természetességen és személyre szabhatóságon itt azt értem, hogy az eredeti beszédadatbázisban előforduló glottalizált eseteknek megfelelő arányú irreguláris hangot tudunk képezni szintetizált beszédben is. Mivel az irreguláris zönge gyakran előfordul mindennapi kommunikációban, ezért ennek alkalmazása beszédszintézisben a természeteshez közelebbi gépi beszédet eredményez. Korábban kimutatták, hogy bizonyos érzelmeket a beszélők a zöngeminőség módosításával is jeleznek: például magyarban a szomorú [77], japánban az ingerült [120] és angolban az unott [121] érzelem esetén használtak glottalizációt a beszélők. Így az irreguláris zönge modell javíthatja az érzelmes, expresszív beszédszintézist. A beszédsérülteket segítő kommunikációs eszközökben hasznos lehet, ha a rendszer az eredeti beszélőre emlékeztető hangon szólal meg, de a legtöbb rendszerben ez nem megvalósított [122]. A személyre szabott szövegfelolvasó jó példája lehet az „idős hang” létrehozása, amely esetén gyakran előfordul a glottalizáció.

A szubglottális rezonanciák vizsgálata, így a III.1. tézis hozzájárul a kvantális elmélet szerinti fonológiai megkülönböztető jegyek működésének megértéséhez. A feltételezések szerint a percepció során a beszédhangok formánsait részben a szubglottális rezonanciákhoz viszonyítjuk (normalizáljuk), ezáltal megkönnyítve egymás beszédének megértését, hiszen az egyes egyének akusztikai produktumában nagy eltérések mutatkoznak. Ez a tulajdonság kihasználha-

tó a beszédtechnológiában is: a szubglottális rezonanciákat már sikerrel alkalmazták automatikus beszélő normalizálásban, melynek során egy beszédfelismerő rendszer minőségét javították gyermek beszéd esetén [83]. A szubglottális rezonanciákon alapuló módszerek praktikus alkalmazhatóságát ugyan csökkenti, hogy a rezonanciafrekvenciák meghatározásához szükséges a beszélő nyakára erősített gyorsulásmérő berendezés jelének rögzítése is, de Arsikere és társai kimutatták, hogy az SGR értékek közvetlenül a beszédjelből mért paraméterek alapján is származtathatóak [99].

Egy előzetes percepciós teszt során megfigyeltük, hogy a formánsok és szubglottális rezonanciák aránya kapcsolatba hozható az észlelt magánhangzó minőségével [J4]. A kísérlet az elöl és hátul képzett magánhangzók illetve az $Sg2$ viszonyát vizsgálta egy beszélő beszédén. Az eredmények alapján, amennyiben az $F2$ és $Sg2$ aránya nem a III.1. tézis szerinti volt (azaz nem teljesült, hogy elöl képzett magánhangzóra $F2 > Sg2$ és hátul képzett magánhangzóra $F2 < Sg2$), akkor a tesztelők nehezebben ismerték fel a magánhangzót. Várhatóan a nem megfelelő $F2 - Sg2$ arány a beszéd során percepciós szempontból előnytelen, és nehezíti a beszéd megértését. Ez alapján készíthető egy olyan eljárás, amely beszédszintetizátor adatbázisából kitisztítja a formáns - szubglottális rezonancia szempontjából nem megfelelő beszédrészleteket, ezzel hozzájárulva a szintetizált beszéd érthetőbbé tételéhez. A III.2. tézisben ismertetett osztályozó kiegészíthető hosszabb hangkapcsolatok (pl. CV vagy VC kapcsolat) artikuláció szerinti osztályozására is, melyre amerikai angol nyelvű mintákkal készült már kísérlet [93]. Az SGR-eket gépi beszédkeltés környezetben eddig csak kezdeti kutatásokban, elsősorban artikulációs beszédszintézisben vizsgálták [100, 101]. Amennyiben a rejtett Markov-modell alapú beszédszintetizátorban a forrás-szűrő modellt sikerül kiegészíteni a szubglottális rezonanciák modellezésével, az tovább javíthatja a gépi beszéd természetességét.

Köszönetnyilvánítás

Ezúton mondok köszönetet konzulensemnek, Dr. Németh Gézának témavezetéséért, a munkám során nyújtott folyamatos segítségéért és támogatásáért, hasznos tanácsaiért és észrevételeiért. Köszönöm neki, hogy munkájával megalapozta tudományos szemléletemet.

Köszönettel tartozom a Beszédtechnológiai Laboratórium jelenlegi és volt munkatársainak. Bartalis Mátyás baráti beszélgetésekkel, Dr. Bőhm Tamás kutatási és módszertani irányelvekkel, Dr. Fék Márk beszédkódolással kapcsolatos ismereteivel, Kiss Géza programozási segítséggel, Dr. Olaszy Gábor nagymértékű tapasztalatával, Tóth Bálint a statisztikai parametrikus beszédszintézis megismertetésével, Dr. Zainkó Csaba jelfeldolgozási ismereteivel segítette munkámat és járult hozzá a disszertáció létrejöttéhez. Emellett köszönöm Fegyó Tibor, Kiss Gábor, Dr. Mihajlik Péter, Nagy Péter, Dr. Szaszák György, Sztahó Dávid, Tarján Balázs és Dr. Vicsi Klára segítségét.

Köszönöm Dr. Steven M. Lulichnak (Indiana University, Bloomington, USA), hogy megismertette velem szubglottális rezonanciákkal foglalkozó kutatásait és támogatta kísérleteimet ebben a témában. Köszönettel tartozom Dr. Gráczi Tekla Etelkának (MTA Nyelvtudományi Intézet), Dr. Bárkányi Zsuzsannának (MTA Nyelvtudományi Intézet) és Beke Andrásnak (MTA Nyelvtudományi Intézet) a kutatási együttműködésért és látóköröm szélesítéséért.

Köszönöm továbbá Dr. Henk Tamás és Dr. Magyar Gábor tanszékvezető uraknak, hogy vezetésük alatt a tanszéken végezhettem doktori munkámat.

Köszönöm minden társszerzőmnek a közös cikkek írásának lehetőségét és a csapatmunkában történő kutatás örömét.

A PPBA, BEA adatbázisok és a III. téziscsoport beszélőinek köszönöm, hogy a kísérleteimhez felhasználhattam a hangjukat. A percepciós tesztekben résztvevőknek köszönöm, hogy meghallgatták és értékelték a hanganyagokat, valamint hasznos megjegyzéseikkel a kutatási irányok távlati meghatározásában is segítettek.

Köszönöm Dr. Gósy Máriának és Dr. Olaszi Péternek, hogy értékes észrevételeikkel és hasznos javaslataikkal segítették a disszertáció jobbá tételét.

Külön köszönöm családomnak: feleségemnek Berninek, kislányomnak Lilinek, kisfiamnak Ábelnek, édesanyámnak Édinek, édesapámnak Istvánnak és bátyámnak Krisztiánnak, hogy doktori tanulmányaim alatt folyamatosan támogattak és megteremtették számomra a kutatáshoz szükséges nyugodt légkört.

A kutatást a NAP (OMFB-00736/2005), az Enhances (NKFP 2/034/2004), a Teleauto (OM-00102/2007), a BelAmi (ALAP2-00004/2005), az ETOCOM (TÁMOP-4.2.2-08/1/KMR-2008-0007), a Kutatóegyetem (TÁMOP-4.2.1/B-09/1/KMR-2010-0002), a CESAR (Grant No. 271022), a Paelife (Grant No. AAL-08-1-2011-0001) és az EITKIC_12-1-2012-001 projektek támogatták.

Irodalomjegyzék

[1] G. Fant, *Acoustic theory of speech production*. The Hague: Mouton, 1960.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, and A. Black, „The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, (Bonn, Germany), pp. 294–299, 2007.

[3] H. Zen, K. Tokuda, and A. W. Black, „Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, Nov. 2009.

[4] A. Hunt and A. Black, „Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, vol. 1, (Atlanta, Georgia, USA), pp. 373–376, 1996.

[5] T. Bőhm, *Analysis and modeling of speech produced with irregular phonation*. PhD disszertáció, BME TMIT, 2009. `http://www.omikk.bme.hu/collections/phd/Villamosmernoki_es_Informatikai_Kar/2010/Bohm_Tamas_Mihaly/ertekezes.pdf`.

[6] K. N. Stevens, *Acoustic Phonetics*. Cambridge: Cambridge University Press, 1998.

[7] S. M. Lulich, „Subglottal resonances and distinctive features," *Journal of Phonetics*, vol. 38, no. 1, pp. 20–32, 2010.

[8] G. Németh and G. Olaszy, eds., *A MAGYAR BESZÉD; Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek*. Budapest: Akadémiai Kiadó, 2010.

[9] M. Gósy, *Fonetika, a beszéd tudománya*. Budapest, Hungary: Osiris Kiadó, 2004.

[10] G. Olaszy, M. Kovács, P. Nikléczy, and M. Gósy, *Magyar nyelvi beszédtechnológiai alapismeretek. (600 oldal CD-ROM-on)*. Budapest: Nikol Kiadó, 2002. `http://alpha.tmit.bme.hu/pub/beszinf/start.html`.

[11] M. Fék, P. Pesti, G. Németh, and C. Zainkó, „Generációváltás a beszédszintézisben," *Híradástechnika*, vol. LXI, no. 3, pp. 21–30, 2006.

[12] F. Jelinek, „Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[13] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, „Speech Synthesis Based on Hidden Markov Models," *Proceedings of the IEEE*, vol. 101, pp. 1234–1252, May 2013.

[14] B. Tóth, *Rejtett Markov-modell alapú gépi beszédkeltés*. PhD disszertáció, BME TMIT, 2013.

[15] B. Tóth and G. Németh, „Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis," *Acta Cybernetica*, vol. 19, no. 4, pp. 715–731, 2010.

[16] M. Airaksinen, *Analysis/Synthesis Comparison of Vocoders Utilized in Statistical Parametric Speech Synthesis*. MS diplomaterv, Aalto University, Finland, 2012. `https://aaltodoc.aalto.fi/handle/123456789/7268`.

[17] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, „An experimental comparison of multiple vocoder types," in *Proc. ISCA SSW8*, pp. 155–160, 2013.

[18] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. Hoboken, New Jersey: John Wiley & Sons, 2003.

[19] A. McCree and T. Barnwell, „A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 242–250, July 1995.

[20] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs: Prentice Hall, 1978.

[21] T. Yoshimura and K. Tokuda, „Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, (Aalborg, Denmark), pp. 2263–2266, 2001.

[22] S.-j. Kim and M. Hahn, „Two-Band Excitation for HMM-Based Speech Synthesis," *IEICE Transactions on Information and Systems*, vol. E90-D, pp. 378–381, Jan. 2007.

[23] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, „Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[24] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, „Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 1, pp. 325–333, 2007.

[25] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, „An excitation model for HMM-based speech synthesis based on residual modeling," in *Proc. ISCA SSW6*, (Bonn, Germany), pp. 131–136, 2007.

[26] R. Maia, T. Toda, and K. Tokuda, „On the state definition for a trainable excitation model in HMM-based speech synthesis," in *Proc. ICASSP*, (Las Vegas, USA), pp. 3965–3968, 2008.

[27] R. Maia, M. Akamine, and M. J. Gales, „Complex cepstrum for statistical parametric speech synthesis," *Speech Communication*, vol. 55, pp. 606–618, Feb. 2013.

[28] G. Fant, J. Liljencrants, and Q. Lin, „A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[29] J. P. Cabral, *HMM-based Speech Synthesis Using an Acoustic Glottal Source Model*. PhD disszertáció, University of Edinburgh, United Kingdom, 2010. http://www.era.lib.ed.ac.uk/bitstream/1842/4877/1/Cabral2011.pdf.

[30] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, „Towards an Improved Modeling of the Glottal Source in Statistical Parametric Speech Synthesis," in *Proc. ISCA SSW6*, (Bonn, Germany), pp. 113–118, 2007.

[31] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, „Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, (Brisbane, Australia), pp. 1829–1832, 2008.

[32] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, „HMM-based speech synthesiser using the LF-model of the glottal source," in *Proc. ICASSP*, (Prague, Czech Republic), pp. 4704–4707, 2011.

[33] P. Alku, „Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, pp. 109–118, June 1992.

[34] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, „HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, (Brisbane, Australia), pp. 1881–1884, 2008.

[35] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, „HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 153–165, Jan. 2011.

[36] T. Raitio, A. Suni, and H. Pulakka, „Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *Proc. ICASSP*, (Prague, Czech Republic), pp. 4564–4567, 2011.

[37] A. Suni, T. Raitio, M. Vainio, and P. Alku, „The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation," in *Blizzard Challenge 2011*, (Turin, Italy), 2011. http://festvox.org/blizzard/bc2011/HELSINKI_Blizzard2011.pdf.

[38] A. Suni, T. Raitio, M. Vainio, and P. Alku, „The GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach," in *Blizzard Challenge 2012*, (Portland, Oregon, USA), 2012. http://festvox.org/blizzard/bc2012/HELSINKI_Blizzard2012.pdf.

[39] T. Raitio, A. Suni, M. Vainio, and P. Alku, „Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. ICASSP*, (Vancouver, Canada), pp. 7830–7834, 2013.

[40] P. Lanchantin, G. Degottex, and X. Rodet, „A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," in *Proc. ICASSP*, (Dallas, Texas, USA), pp. 4630–4633, 2010.

[41] G. Degottex, *Glottal source and vocal-tract separation*. PhD disszertáció, Ircam, France, 2010. http://hal.archives-ouvertes.fr/docs/00/64/22/93/PDF/Degottex2010_PhD_v4_Final.pdf.

[42] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, „Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, pp. 278–294, Feb. 2013.

[43] D. Erro, A. Moreno, and A. Bonafonte, „Flexible Harmonic/Stochastic Speech Synthesis," in *Proc. ISCA SSW6*, (Bonn, Germany), pp. 194–199, 2007.

[44] D. Erro, I. n. Sainz, E. Navas, and I. Hernáez, „Improved HNM-based Vocoder for Statistical Synthesizers," in *Proc. Interspeech*, (Florence, Italy), pp. 1809–1812, 2011.

[45] Z. Wen and J. Tao, „An excitation model based on inverse filtering for speech analysis and synthesis," in *IEEE MLSP*, (Beijing, China), 2011.

[46] Z. Wen and J. Tao, „Inverse Filtering Based Harmonic plus Noise Excitation Model for HMM-based Speech Synthesis," in *Proc. Interspeech*, (Florence, Italy), pp. 1805–1808, 2011.

[47] Z. Wen, H. Kawahara, and J. Tao, „Pitch-Scaled Analysis based Residual Reconstruction for Speech Analysis and Synthesis," in *Proc. Interspeech*, (Portland, Oregon, USA), pp. 374–377, 2012.

[48] Z. Wen and J. Tao, „Amplitude spectrum based Excitation model for HMM-based Speech Synthesis," in *Proc. Interspeech*, (Portland, Oregon, USA), pp. 1428–1431, 2012.

[49] J. S. Sung, D. H. Hong, K. Oh, and N. Kim, „Excitation modeling based on waveform interpolation for HMM-based speech synthesis," in *Proc. Interspeech*, (Makuhari, Japan), pp. 813–816, 2010.

[50] C.-s. Jung, Y.-s. Joo, and H.-g. Kang, „Waveform Interpolation-Based Speech Analysis/Synthesis for HMM-Based TTS Systems," *IEEE Signal Processing Letters*, vol. 19, pp. 809–812, Dec. 2012.

[51] J. S. Sung, D. H. Hong, H. W. Koo, and N. S. Kim, „Statistical Approaches to Excitation Modeling in HMM-Based Speech Synthesis," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 2, pp. 379–382, 2013.

[52] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, „Using a Pitch-Synchronous Residual Codebook for Hybrid HMM/frame Selection Speech Synthesis," in *Proc. ICASSP*, (Taipei, Taiwan), pp. 3793 – 3796, 2009.

[53] T. Drugman, G. Wilfart, and T. Dutoit, „A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, (Brighton, UK), pp. 1779–1782, 2009.

[54] T. Drugman, *Advances in Glottal Analysis and its Applications*. PhD disszertáció, University of Mons, Belgium, 2011. `http://tcts.fpms.ac.be/~drugman/files/DrugmanPhDThesis.pdf`.

[55] T. Drugman and T. Dutoit, „The Deterministic Plus Stochastic Model of the Residual Signal and its Applications," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 968–981, Mar. 2012.

[56] J. Nurminen, H. Silén, E. Helander, and M. Gabbouj, „Evaluation of detailed modeling of the LP residual in statistical speech synthesis," in *Proc. ISCAS*, pp. 313–316, 2013.

[57] L. Redi and S. Shattuck-Hufnagel, „Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, no. 4, pp. 407–429, 2001.

[58] A. Markó, „A glottalizáció határjelző szerepe a felolvasásban," *Beszédkutatás 2011*, pp. 31–45, 2011.

[59] A. Markó, „Az irreguláris zönge szerepe a magánhangzók határának jelölésében V(#)V kapcsolatokban," *Beszédkutatás 2012*, pp. 5–29, 2012.

[60] A. Markó, „Boundary marking in Hungarian V(#)V clusters with special regard to the role of irregular phonation," *The Phonetician*, no. 105-106, pp. 7–26, 2012.

[61] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, „Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 103, pp. 2649–2658, May 1998.

[62] T. Bőhm, N. Audibert, S. Shattuck-Hufnagel, G. Németh, and V. Aubergé, „Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles," in *Acoustics'08*, (Paris, France), pp. 6141–6146, 2008.

[63] T. Bőhm, Z. Both, and G. Németh, „Automatic Classification of Regular vs. Irregular Phonation Types," in *NOLISP*, (Vic, Spain), pp. 43–50, 2009.

[64] K. Surana, *Classification of vocal fold vibration as regular or irregular in normal voiced speech*. MEng diplomaterv, MIT, USA, 2006. `http://dspace.mit.edu/handle/1721.1/37104`.

[65] C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, „A Method for Automatic Detection of Vocal Fry," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 47–56, Jan. 2008.

[66] A. Beke and E. Heltovics, „A glottalizált magánhangzók automatikus osztályozása spontán magyar beszédben," *Beszédkutatás 2010*, pp. 253–263, 2010.

[67] J. Kane, T. Drugman, and C. Gobl, „Improved automatic detection of creak," *Computer Speech & Language*, vol. 27, pp. 1028–1047, June 2013.

[68] D. H. Klatt and L. C. Klatt, „Analysis, synthesis, and perception of voice quality variations among female and male talkers.," *The Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, Feb. 1990.

[69] H. Silén, E. Helander, K. Koppinen, and M. Gabbouj, „Building a Finnish unit selection TTS system," in *Proc. ISCA SSW6*, (Bonn, Germany), pp. 310–315, 2007.

[70] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, „Parameterization of vocal fry in HMM-based speech synthesis," in *Proc. Interspeech*, (Brighton, UK), pp. 1775–1778, 2009.

[71] T. Drugman, J. Kane, and C. Gobl, „Modeling the Creaky Excitation for Parametric Speech Synthesis," in *Proc. Interspeech*, (Portland, Oregon, USA), pp. 1424–1427, 2012.

[72] J. Slifka, „Irregular phonation and its preferred role as a cue to silence in phonological systems," in *ICPhS*, (Saarbrücken, Germany), pp. 229–232, 2007.

[73] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, „Glottalization of word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, vol. 24, pp. 423–444, Oct. 1996.

[74] C. Gobl and A. N. Chasaide, „The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, Apr. 2003.

[75] N. Malyska, *Analysis of nonmodal glottal event patterns with application to automatic speaker recognition*. PhD disszertáció, MIT, USA, 2008. `http://dspace.mit.edu/handle/1721.1/43804`.

[76] E. Moulines and F. Charpentier, „Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, Dec. 1990.

[77] C. Zainkó, M. Fék, and G. Németh, „Expressive Speech Synthesis Using Emotion-Specific Speech Inventories," *Lecture Notes in Computer Science*, no. 5042, pp. 225–234, 2008.

[78] T. Drugman, J. Kane, T. Raitio, and C. Gobl, „Prediction of Creaky Voice from Contextual Factors," in *Proc. ICASSP*, (Vancouver, Canada), pp. 7967–7971, 2013.

[79] T. Raitio, J. Kane, T. Drugman, and C. Gobl, „HMM-based synthesis of creaky voice," in *Proc. Interspeech*, pp. 2316–2320, 2013.

[80] H. Gray, *Anatomy of the human body*. 1918. `http://www.bartleby.com/107/illus961.html`.

[81] I. Titze, T. Riede, and P. Popolo, „Nonlinear source-filter coupling in phonation: vocal exercises," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1902–1915, Apr. 2008.

[82] M. S. Howe and R. S. McGowan, „Analysis of Flow-Structure Coupling in a Mechanical Model of the Vocal Folds and the Subglottal System," *Journal of Fluids and Structures*, vol. 25, pp. 1299–1317, Nov. 2009.

[83] S. Wang, S. M. Lulich, and A. Alwan, „Automatic detection of the second subglottal resonance and its application to speaker normalization," *The Journal of the Acoustical Society of America*, vol. 126, pp. 3268–3277, Dec. 2009.

[84] A. Madsack, S. M. Lulich, W. Wokurek, and G. Dogil, „Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs," in *Proc. LabPhon*, (Wellington, New Zealand), pp. 91–92, 2008.

[85] Y. Jung, *Acoustic Articulatory Evidence for Quantal Vowel Categories: The Features [low] and [back]*. PhD disszertáció, MIT, USA, 2009. http://dspace.mit.edu/handle/1721.1/54630.

[86] K. N. Stevens, „On the quantal nature of speech," *Journal of Phonetics*, vol. 17, pp. 3–45, 1989.

[87] S. M. Lulich, *The Role of Lower Airway Resonances in Defining Vowel Feature Contrasts*. PhD disszertáció, MIT, USA, 2006. http://dspace.mit.edu/handle/1721.1/38248.

[88] X. Chi and M. Sonderegger, „Subglottal coupling and its influence on vowel formants," *The Journal of the Acoustical Society of America*, vol. 122, pp. 1735–1745, Sept. 2007.

[89] S. M. Lulich, J. R. Morton, H. Arsikere, M. S. Sommers, G. K. F. Leung, and A. Alwan, „Subglottal resonances of adult male and female native speakers of American English," *The Journal of the Acoustical Society of America*, vol. 132, pp. 2592–2602, Oct. 2012.

[90] S. M. Lulich, A. Bachrach, and N. Malyska, „A role for the second subglottal resonance in lexical access," *The Journal of the Acoustical Society of America*, vol. 122, pp. 2320–2327, Oct. 2007.

[91] S. Wang, A. Alwan, and S. Lulich, „Speaker normalization based on subglottal resonances," in *Proc. ICASSP*, (Las Vegas, Nevada, USA), pp. 4277–4280, 2008.

[92] S. Wang, S. Lulich, and A. Alwan, „A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation," in *Proc. Interspeech*, (Brisbane, Australia), pp. 1717–1720, 2008.

[93] S. M. Lulich and N. Chen, „Automatic classification of consonant-vowel transitions based on subglottal resonances and second formant frequencies," in *Proceedings of Meetings on Acoustics*, vol. 6, pp. 060005 (1–8), 2009.

[94] S. M. Lulich, „On the relation between locus equations and subglottal resonances," in *Proceedings of Meetings on Acoustics*, vol. 5, pp. 060003 (1–10), 2009.

[95] H. Arsikere, S. M. Lulich, and A. Alwan, „Automatic estimation of the second subglottal resonance from natural speech," in *Proc. ICASSP*, (Prague, Czech Republic), pp. 4616–4619, 2011.

[96] H. Arsikere, S. M. Lulich, and A. Alwan, „Automatic estimation of the first subglottal resonance," *The Journal of the Acoustical Society of America, Express Letters*, vol. 129, pp. EL197–203, May 2011.

[97] H. Arsikere, G. K. Leung, S. M. Lulich, and A. Alwan, „Automatic estimation of the first two subglottal resonances in children's speech with application to speaker normalization in limited-data conditions," in *Proc. Interspeech*, (Portland, Oregon, USA), pp. 1267–1270, 2012.

[98] H. Arsikere, G. K. Leung, S. M. Lulich, and A. Alwan, „Automatic height estimation using the second subglottal resonance," in *Proc. ICASSP*, (Kyoto, Japan), pp. 3989 – 3992, 2012.

[99] H. Arsikere, G. K. Leung, S. M. Lulich, and A. Alwan, „Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation," *Speech Communication*, vol. 55, pp. 51–70, Jan. 2013.

[100] K. S. Gorbunov and I. S. Makarov, „The subglottic region in articulator synthesizers," *Journal of Communications Technology and Electronics*, vol. 56, pp. 1504–1509, Dec. 2011.

[101] S. Hiroya, N. Miki, and T. Mochida, „Multi-closure-interval Linear Prediction Analysis Based on Phase Equalization," in *Proc. APSIPA*, (Xian, China), 2011.

[102] G. Olaszy, „Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai," *Beszédkutatás 2013*, pp. 261–270, 2013.

[103] M. Gósy, „Magyar spontánbeszéd-adatbázis - BEA," *Beszédkutatás 2008*, pp. 194–207, 2008.

[104] P. Mihajlik, T. Révész, and P. Tatai, „Phonetic transcription in automatic speech recognition," *Acta Linguistica Hungarica*, vol. 49, no. 3-4, pp. 407–425, 2002.

[105] T. Drugman and T. Dutoit, „Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech*, (Brighton, UK), pp. 2891–2894, 2009.

[106] P. Boersma and D. Weenink, „Praat: doing phonetics by computer [Computer program]. Version 5.1.20," 2009. `http://www.praat.org`.

[107] K. Sjölander and J. Beskow, „Wavesurfer [Computer program], Version 1.8.5." `http://www.speech.kth.se/wavesurfer/`.

[108] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, 2nd ed., 2005.

[109] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, „Statistical Analysis of the Blizzard Challenge 2007 Listening Test Results," in *Blizzard Challenge 2007*, (Bonn, Germany), 2007. `http://festvox.org/blizzard/bc2007/blizzard_2007/full_papers/blz3_003.pdf`.

[110] D. Talkin, „A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 495–518, Elsevier, 1995.

[111] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, „Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, (Yokohama, Japan), pp. 1043–1046, 1994.

[112] S. Imai, K. Sumita, and C. Furuichi, „Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[113] T. Drugman and M. Thomas, „Detection of glottal closure instants from speech signals: a quantitative review," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 994–1006, Mar. 2012.

[114] G. de Krom, „A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech and Hearing Research*, vol. 36, pp. 254–266, Apr. 1993.

[115] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. C. Guiod, and S. L. Goldman, „Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech and Hearing Research*, vol. 38, pp. 1212–1223, Dec. 1995.

[116] M. Iseli and A. Alwan, „An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proc. ICASSP*, (Montreal, Quebec, Canada), pp. 669–672, 2004.

[117] A. Nádasdy and P. Siptár, „A magánhangzók," in *Strukturális magyar nyelvtan 2. - Fonológia* (F. Kiefer, ed.), pp. 42–181, Budapest: Akadémiai Kiadó, 1994.

[118] K. Abari and G. Olaszy, „A formánsmenetek rendszere CVC kapcsolatok magánhangzóiban a C képzési helyének függvényében," *Beszédkutatás 2012*, pp. 70–93, 2012.

[119] B. M. Lobanov, „Classification of Russian Vowels Spoken by Different Speakers," *The Journal of the Acoustical Society of America*, vol. 49, pp. 606–608, Feb. 1971.

[120] T. Sadanobu, „A natural history of Japanese pressed voice," *Journal of the Phonetic Society of Japan*, vol. 8, no. 1, pp. 29–44, 2004.

[121] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press, 1980.

[122] C. Jreige, R. Patel, and H. T. Bunnell, „VocaliD: personalizing text-to-speech synthesis for individuals with severe speech impairment," in *ASSETS*, (Pittsburgh, Pennsylvania, USA), pp. 259–260, 2009.

Az internetes források ellenőrzésének utolsó dátuma: 2013. július 23.

A szerző tudományos közleményei

A tézispontokhoz kapcsolódó tudományos közlemények

Folyóiratcikkek

[J1] Tamás Gábor Csapó, Géza Németh, „Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation," *IEEE Journal on Selected Topics in Signal Processing*, elfogadva, 2013.
(BME-PA pontszám: 100% · 6p = 6p.) Scopus / Web of Science, IF: 3.297.

[J2] Tamás Gábor Csapó, Géza Németh, „Statistical parametric speech synthesis with a novel codebook-based excitation model," *Intelligent Decision Technologies*, elfogadva, 2013.
(BME-PA pontszám: 100% · 6p = 6p.) Scopus.

[J3] Tamás Gábor Csapó, „Increasing the naturalness of synthesized speech (PhD summary)," *The Phonetician*, No. 104–105, pp. 88–97, 2012.
(BME-PA pontszám: 100% · 0p = 0p.) (ismeretterjesztő cikk)

[J4] Tamás Gábor Csapó, Tekla Etelka Gráczi, Zsuzsanna Bárkányi, András Beke, Steven M. Lulich, „Patterns of Hungarian vowel production and perception with regard to subglottal resonances," *The Phonetician*, No. 99–100, pp. 7–28, 2011.
(BME-PA pontszám: 50% · 6p = 3p.)

Konferenciacikkek

[C1] Tamás Gábor Csapó, Géza Németh, „Transformation of irregular voice to modal voice by residual analysis and synthesis," *IEEE Signal Processing Letters*, elkészítés alatt, 2013.
(BME-PA pontszám: 0p · 100% = 0p.)

[C2] Tamás Gábor Csapó, Géza Németh, „A novel irregular voice model for HMM-based speech synthesis," *Proc. ISCA SSW8 - 8th Speech Synthesis Workshop*, (Barcelona, Spanyolország), pp. 229-234., 2013.
(BME-PA pontszám: 100% · 3p = 3p.)

[C3] Tamás Gábor Csapó, Géza Németh, „A novel codebook-based excitation model for use in speech synthesis," *IEEE CogInfoCom 2012*, (Kassa, Szlovákia), pp. 661–665, 2012.
(BME-PA pontszám: 100% · 3p = 3p.)

[C4] Tamás Gábor Csapó, Zsuzsanna Bárkányi, Tekla Etelka Gráczi, Tamás Bőhm, Steven M. Lulich, „Relation of formants and subglottal resonances in Hungarian vowels," *Proc. Interspeech 2009*, (Brighton, Egyesült Királyság), pp. 484–487, 2009.
(BME-PA pontszám: 50% · 3p = 1.5p.)

Csak kivonatban megjelent konferencia-előadások

[C5] Csapó Tamás Gábor, Németh Géza, „Irreguláris beszéd regulárissá alakítása beszédkódoláson alapuló módszerrel," *Beszédkutatás*, (Budapest), 2013. november 14–15.
(BME-PA pontszám: 100% · 0p = 0p.)

[C6] Csapó Tamás Gábor, Bárkányi Zsuzsanna, Gráczi Tekla Etelka, Beke András, Bőhm Tamás, „A magánhangzó-formánsok és a szubglottális rezonanciák összefüggése a spontán beszédben," *Beszédkutatás*, (Budapest), 2009. október 16–17.
(BME-PA pontszám: 20% · 0p = 0p.)

A szerző további tudományos közleményei

Folyóiratcikkek

[J5] Tamás Gábor Csapó, Csaba Zainkó, Géza Németh, „A Study of Prosodic Variability Methods in a Corpus-Based Unit Selection Text-To-Speech System," *Infocommunications Journal*, LXV. évf., I. sz., pp. 32–37, 2010.
(BME-PA pontszám: 50% · 4p = 2p.)

[J6] Csapó Tamás Gábor, „Változatos prozódia megvalósítása szövegfelolvasó rendszerekben," *Akusztikai Szemle*, IX. évf., 3. sz., pp. 16–18, 2009.
(BME-PA pontszám: 100% · 2p = 2p.)

[J7] Csapó Tamás Gábor, Németh Géza, Fék Márk, „Szövegfelolvasó természetességének növelése," *Híradástechnika*, LXIII. évf., 5. sz., pp. 21–30, 2008.
(BME-PA pontszám: 50% · 2p = 1p.)

Konferenciacikkek

[C7] Éva Székely, Tamás Gábor Csapó, Bálint Tóth, Péter Mihajlik, Julie Carson-Berndsen „Synthesizing Expressive Speech from Amateur Audiobook Recordings," *SLT 2012*, (Miami, Florida, USA), pp. 297–302, 2012.
(BME-PA pontszám: 20% · 3p = 0.6p.)

[C8] Csapó Tamás Gábor, Németh Géza, „Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval," *Magyar Számítógépes Nyelvészeti Konferencia*, (Szeged), pp. 167–177, 2011.
(BME-PA pontszám: 100% · 1p = 1p.)

[C9] Tekla Etelka Gráczi, Steven M Lulich, Tamás Gábor Csapó, András Beke, „Context and speaker dependency in the relation of vowel formants and subglottal resonances - Evidence from Hungarian," *Proc. Interspeech 2011*, (Firenze, Olaszország), pp. 1901–1904, 2011.
(BME-PA pontszám: 25% · 3p = 0.75p.)

[C10] Géza Németh, Gábor Olaszy, Tamás Gábor Csapó, „Spemoticons: Text-To-Speech based emotional auditory cues," *ICAD 2011*, (Budapest), 2011.
(BME-PA pontszám: 50% · 2p = 1p.)

[C11] Csaba Zainkó, Tamás Gábor Csapó, Géza Németh, „Special Speech Synthesis for Social Network Websites," *Lecture Notes In Computer Science*, 6231: pp. 455–463, Paper 58, 2010.
(BME-PA pontszám: 50% · 6p = 3p.)

[C12] Csapó Tamás Gábor, Németh Géza, „Mássalhangzó-magánhangzó kapcsolatok automatikus osztályozása szubglottális rezonanciák alapján," *Magyar Számítógépes Nyelvészeti Konferencia*, (Szeged), 2009. december 3-4., pp. 226-237.
(BME-PA pontszám: 100% · 1p = 1p.)

[C13] Géza Németh, Márk Fék, Tamás Gábor Csapó, „Increasing Prosodic Variability of Text-To-Speech Synthesizers," *Proc. Interspeech 2007*, (Antwerpen, Belgium), pp. 474–477.
(BME-PA pontszám: 50% · 3p = 1.5p.)

Könyvfejezetek

[B1] Csapó Tamás Gábor, „Beszédfelismerők minősítése”, Németh, G., Olaszy, G. (szerk.), A magyar beszéd - beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Akadémiai Kiadó, Budapest, 2010, pp. 407–409.
(BME-PA pontszám: 100% · 0p = 0p.)

[B2] Csapó Tamás Gábor, „A beszéddallam változatosságának statisztikai modellezése”, Németh, G., Olaszy, G. (szerk.), A magyar beszéd - beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Akadémiai Kiadó, Budapest, 2010, pp. 446–449.
(BME-PA pontszám: 100% · 0p = 0p.)

[B3] Csapó Tamás Gábor, „VXML”, Németh, G., Olaszy, G. (szerk.), A magyar beszéd - beszédkutatás, beszédtechnológia, beszédinformációs rendszerek, Akadémiai Kiadó, Budapest, 2010, pp. 631–635.
(BME-PA pontszám: 100% · 0p = 0p.)

Csak kivonatban megjelent konferencia-előadások

[C14] Csapó Tamás Gábor, Zainkó Csaba, Németh Géza, „Szintetizált beszéd prozódiai változatosságának növelése spontán beszéd alapján,” *Beszédkutatás*, (Budapest), 2009. október 16-17.
(BME-PA pontszám: 50% · 0p = 0p.)

[C15] Géza Németh, Tamás Gábor Csapó, Bálint Tóth, „Improving the Quality of Unit Selection and HMM based Speech Synthesis,” *FuturICT*, (Budapest), 2009. június 29-30.
(BME-PA pontszám: 50% · 0p = 0p.)