PRECÍZIÓS, PÁRHUZAMOS, MAGYAR BESZÉDADATBÁZIS FEJLESZTÉSE ÉS SZOLGÁLTATÁSAI

Olaszy Gábor

Bevezetés

A beszédkutatásban világszerte egyre nagyobb teret kapnak az előre elkészített, annotált és szegmentált beszédadatbázisok. A gyarapodás oka, hogy a kutatásokban és a fejlesztésekben egyre inkább statisztikai eljárásokkal vizsgálják a beszédet, azaz nagy adattömeg vizsgálatával határozzák meg a beszédre jellemző paramétereket, azok változási tendenciáit. Mivel egy ilyen adatbázis nagyszámú adatot tartalmaz, az annotálását és szegmentálását csak gépi eljárások támogatásával lehet elvégezni. Ebből következik, hogy az adathalmazban lesznek hibás adatok is. A kutató számára viszont az lenne a kívánatos, hogy gyakorlatilag hibamentes adathalmazra építhesse a vizsgálatait. Az ideális megoldás tehát az lenne, ha az annotálási és címkézési hibák számát a minimumra lehetne csökkenteni a beszédadatbázisokban.

A jelen tanulmányban olyan beszédadatbázist mutatunk be, amely az ideális esetet közelíti, mivel a gépi hibázásokat félautomatikus támogatással feltártuk és manuális javítással megszüntettük. Ilyen részletességgel és pontossággal feldolgozott beszédadatbázis korábban még nem készült magyar nyelvre. Számos beszédadatbázist készítettek már az elmúlt évtizedekben Magyarországon, mindegyiket más-más céllal (Gósy et al. 2012).

Kiemelendő a legutóbbi fejlesztés, amely az MTA Nyelvtudományi Intézetében folyik, egy BEszélt nyelvi Adatbázis (BEA) létrehozása. Itt a kitűzött cél, hogy 500 beszélőtől gyűjtsenek felolvasott és spontán beszédet is. Jelenleg a kutatás a felénél tart, de már most is látszik, hogy számos fonetikai kutatás alapját képezi (Gósy 2012) a hatalmas adattár. A BEA hanghullámot és szöveg szintű átiratot tartalmaz.

Anyag, módszer, beszélő személyek

A most ismertetett beszédadatbázis nyelvi anyagának gerincét egy korábbi kutatáshoz gyűjtött szöveges mondatkorpusz képezi (Vicsi – Vígh 1998), amely a BABEL projekt keretében készült el (1992 mondat, 20873 szó 85424 hang, 33783 magánhangzó) és a magyar nyelvet jól reprezentálja hangstatisztikailag. Ezt egészítettük ki célzottan, egyrészről különböző hosszúságú és típusú kérdő mondatokkal (*Ő?, Én?, Még?, Baba?, Babával?*

stb.), valamint rövid kijelentő mondatokkal az egy hangos mondattól kezdve a háromszótagos egyszavas mondatig. Ez a szöveges anyag 522 mondatot, 1985 szót, 5534 hangot, ezen belül 1844 magánhangzót tartalmaz. A kibővítéssel az volt a célunk, hogy időszerkezetileg is és a beszéddallam vonatkozásában is minél széleskörűbben reprezentáljuk a magyar beszéd sajátosságait. Mindkét szöveges anyagot 10 beszélő, 5 nő (NO) és 5 férfi (FF) olvasta fel, életkoruk 30-65 év között oszlik el, mindannyian Budapestiek, a köznapi magyart beszélik. Foglalkozásuk: tanár, színész, adminisztrátor, kutató, mérnök, zenész, énekművész. A teljes beszédadatbázis tartalma mintegy 900 000 beszédhang, ebből közel 350 000 magánhangzó. A beszédadatbázis rövidített neve PPBA. A hangfelvételeket egységesen a BME Híradástechnikai Tanszék Rezgésakusztikai Laboratóriumának professzionális hangstúdiójában készítettük 44,1 kHz-es mintavételezéssel és 16 bites lineáris kvantálással. A hangrögzítés közvetlenül számítógépre történt. A hangfelvételeket annotáltuk és szegmentáltuk saját fejlesztésű kényszerített gépi beszédfelismerő programmal (Mihajlik et al. 2002). Ezután, szintén saját fejlesztésű hibadetektálóval (Olaszy – Bartalis 2008) ellenőriztük a kapott adatokat és a hibás hangkódokat, illetve hanghatárokat kézzel, vizuális és auditív ellenőrzés segítségével korrigáltuk. Az ellenőrzéshez és javításhoz a Praat program megjelenítési algoritmusait használtuk (Boersma–Weenink 2005). A kombinált annotálási és szegmentálási módszer végeredménye, hogy a beszédadatbázis pontos adatokat tartalmaz (ezért precíziós), tehát referenciaként használható. A hanghullám akusztikai tartalma és a hozzá rendelt hangszimbólumok és címkék pontosan megfelelnek egymásnak.

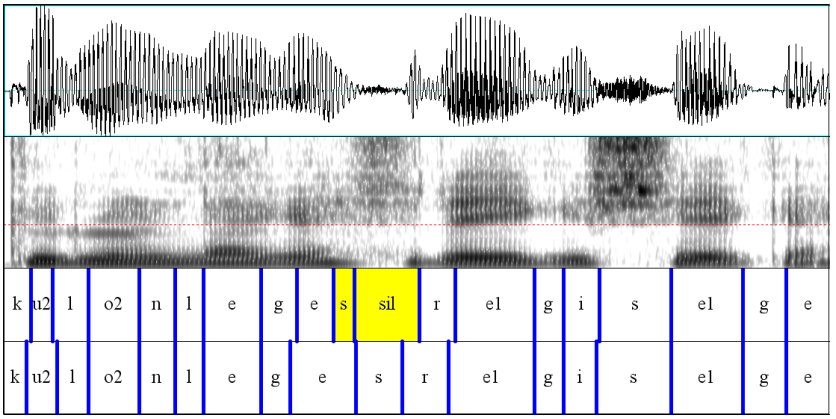
Az annotálás négy féle jelölési csoportot foglal magába: beszédhang szimbólumok, nem beszédhez tartozó részek jelölése „sil" jellel (szünet, krákogás, levegővétel stb.), ejtési eltérés jelölése (például glottalizáció) a hangszimbólum mellé tett csukó zárójellel. A hangjelöléseket mind Sampa, mind TMIT jelöléssel (Németh – Olaszy 2010 77) megadtuk. A beszédhullámmal párhuzamos szegmentálás (címkézés) feldolgozási szintjei a következők: zöngés hangperiódusok (pitch marks: PM) egyenkénti jelölése, hanghatárok és a „sil" jellel jelölt szakaszok határai, valamint szóhatár jelölések. Az annotálásra és címkézésre támaszkodva formánsadatok meghatározását is elvégeztük. Összességében mintegy 4,5 millió formánsadatról van szó. Itt is célunk volt a precíz adatmeghatározás. Ezért első lépésben a Praat programot használtuk, majd a kapott formánsértékeket saját algoritmussal ellenőriztük (Abari – Olaszy 2012) és a hibás értékeket manuálisan javítottuk. A mondat szerkezetéből adódó hangsúlyok jelölése is célunk volt. A beszédatbázis szövegkorpuszának kijelentő mondatait szó

szinten bináris (hangsúlyos/nem hangsúlyos) címkékkel láttuk el. A feldolgozáshoz gépi támogatásként a Profivox TTS rendszer hangsúlymeghatározó modulját használtuk, mivel más ilyen szoftver nem áll jelenleg rendelkezésre. A Profivox mintegy 72%-os pontossággal végzi a hangsúly besorolást (Tamm – Olaszy 2005). Manuálisan is ellenőriztük az eredményt, és ha kellett javítottuk. Ezzel létrehoztuk az első magyar hangsúlyjelölési szöveges adattárat.
A PPBA szerkezeti felépítése egységes, minden beszélő adatai ugyanolyan könyvtár szerkezettel rendelkeznek, tehát a párhuzamos adatkeresés és az esetleges összehasonlítás biztosított.

A precíziós feldolgozás és a tisztán gépi címkézés kérdéskörе

Vannak olyan kutatások, amelyek igénylik a precíz annotálást és címkézést. Ilyenek lehetnek a hangidőtartamokkal és a beszéd egyéb időszerkezeti elemeivel kapcsolatos statisztikai mérések. A beszéd spektrális felépítésével kapcsolatos statisztikai kutatások is ide tartoznak, például a formáns menetek modellezése, a zárfelpattanások vizsgálata stb. Ha ilyen vizsgálatokat egy beszédadatbázisból származó adathalmazra akar alapozni a kutató, akkor joggal várja el, hogy minden hanghatár címkéje korrekt időpontra legyen bejelölve, valamint azt, hogy a jelölt hangnak megfelelő akusztikai tartalom legyen a bejelölt beszédszakaszon.
A tisztán gépi annotálás egyik hátránya lehet, hogy a mondatkezdő zöngétlen zárhangok virtuális kezdetét a gépi címkézés általában nem jelöli. Ugyanez a helyzet a beszédszüneteket jelölő „sil” címkék utáni ilyen hangokra is. Az ilyen esetekben a hangsorkezdő zöngétlen zárhang kezdési pontját jelentő hanghatár címkét balra kell mozgatni, hogy a virtuális hangidőtartamot érzékeltessük, vagyis a nem látható néma fázist is a hanghoz jelöljük. Ha nem így van jelölve, akkor például hangidőtartam méréseknél hibásan mérjük ezen hangok időtartamát (túl rövidre).
A VV, VVV kapcsolatok határát csak vizuális ellenőrzéssel és meghallgatással lehet jó közelítéssel bejelölni.
A fonetikai átíró program is hibázhat, nemazt a hangot jelöli az adott helyen, ami elhangzik a hangsorban (*arccsont, technika, meggyújt*). A gép által automatikusan jelölt „sil” szakaszok esetében a beszédhangokon túli részek (összefoglalva: szünetek) gépi jelölésének kritériumrendszere nem kiforrott. Gyakran kell a kézi ellenőrzés során „sil”-t betenni, kivenni, a „sil”-hez jelölt határ helyét módosítani. A kényszerített felismerő gyakran jelöl a hanghullámban automatikusan olyan „sil”-szakaszt, ami valójában nem szünetet reprezentál, hanem egy hang része (1. ábra). Ennek ellenkezője is előfordult, hogy be kellett iktatni egy sil jelölést.

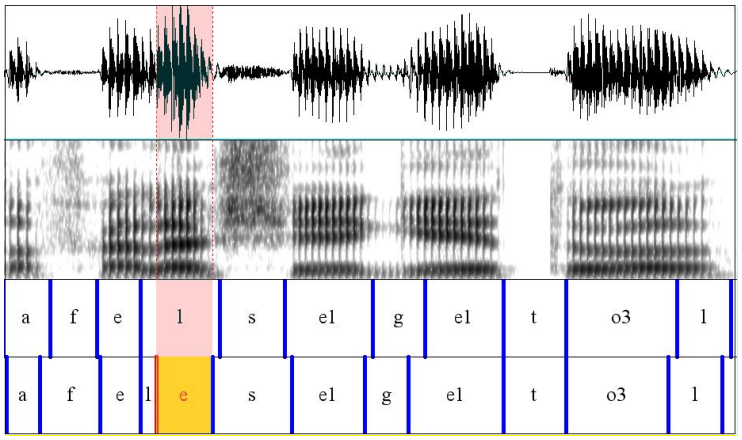
Emberi mulasztásból adódó hibák is keletkeznek egy nagyméretű beszédadatbázis készítése során. A felolvasó személy mást mond, mint a szöveg. Ez még akkor is előfordul, ha a szöveg teljesen helyesen van írva. A beszélő automatikusan átfogalmazza kissé a leírt mondatot és más szavakkal mondja azt (például beszúr egy névelőt és elhagy egy ragot, de a mondat teljesen értelmes marad). Előfordult, hogy egy mondatban a *fogok* helyett *fognak*-ot mondott, egy másikban a *valamilyen* helyett *milyen*-t ejtett a bemondó. Ezt is korrigálni kell.



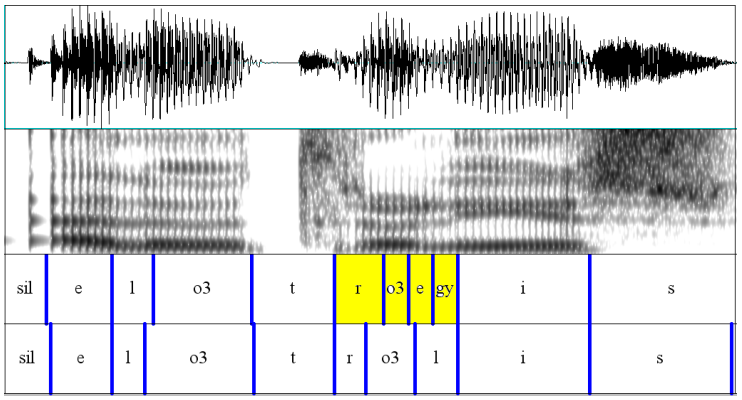
1. ábra
Példa egy felesleges szünetre (fent) és a kézi javítás utáni helyzetre (lent)

A szöveg is lehet hibás (például kimarad egy betű, elmarad egy rag stb.), de a felolvasó személy automatikusan korrigálja azt az ejtés során. Ilyenkor a kényszerített felismerő program a szöveg karaktereit veszi figyelembe és akár több hangon keresztül is hamis hanghatárokat jelöl meg, mivel a szöveg által megadott beszédhangra jellemző akusztikai tartalom nem egyezik az ejtett tartalommal. A 2. ábrán látható egy ilyen példa, a karakterhibás szöveg: *felségét*; a felolvasott szó helyesen: feleségét. Ezt a hibát például a gépi hibakereső segítő algoritmusunk a feltűnően hosszú hangidőtartam alapján találta gyanúsnak és a vizuális ellenőrzés során ki is derült, hogy kimaradt egy hangjelölés. Programhiba is előfordult a kényszerített beszédfelismerő működésében. Ilyenkor a helyes ortografikus szöveg fonetikai átírása nem korrekt és más hangokat tesz a hangsorba a gép, mint amilyenek a szövegből adódnának. A bemondó viszont helyesen olvassa fel a szöveget, tehát a hangazonosítások a spektrális tartalom és a hangjelölés alapján nem lesznek korrektek. Ilyenkor a felismerő csak több hang után tud újra visszatérni a

helyes döntésekhez (3. ábra). A példákból látható, hogy az annotálás és a címkézés utólagos ellenőrzése és javítása számos későbbi munkához szükségszerű és emeli a beszédadatbázis tudományos értékét.



2. ábra
A betűkimaradás rossz határjelölést eredményez (fent) A kézi javítás után áll helyre a helyes hang- és határjelölés (alsó hangsor)



3. ábra
A kényszerített beszéd felismerő hibázott (felső hangsor). A kézi javítás utáni helyes hangsor és határok alul láthatók

Az adatbázis szolgáltatásai

Az adatbázisban tárolt adatok a következők.

Szöveg: a mondat helyesírással megadott formája xx.txt fájlban.

Fonetikus átirat: a szöveges forma alapján készült el és a felolvasott mondat (xx.wav fájl) pontos tartalmát rögzíti a négyféle jelölési csoporttal. Az automatikus feldolgozásból eredő annotálási és címkézési adatokat az xx.TextGrid fájl tartalmazza, a kézi javítással módosított adatokat az xx_man.TextGrid fájl.

Hangnyomás-időfüggvény: a mondat felolvasott formája az xx.wav fájlban.

Zöngeszinkron jelek: Ennek megfelelően a zöngés szakaszok belsejében kiszámítható az alapfrekvencia, illetve az esetleges irregularitások is kereshetők. Az információt az xx.pit fájl tartalmazza.

Hanghatárok: a mondat beszédhangjainak és „sil" jelzéseinek kezdőpontjai (a pontos időkoordinátákkal). Az automatikus feldolgozásból eredő adatokat az xx.TextGrid fájl tartalmazza, a kézi javítással módosított adatokat az xx_man.TextGrid fájl. A hanghatárok pontossága 10 ms-on belüli.

Szóhatárok: A mondat szavainak végét külön címkével jelöljük. Tehát szó szintű vizsgálatok is végezhetők (szóhossz, szóhelyzet stb.). Ezt az információt az xx.ssw fájl tartalmazza. A jelölések kezelik azt az esetet is, amikor a szóhatár nem különíthető el hang szinten, hanem egyetlen hang képezi az előző szó végét és a következő szó elejét is.

Formánsfrekvenciák: a magánhangzók F1, F2, F3 formáns adatai minden magánhangzóra meghatározásra kerültek a hangon belüli 5 mérési pontban (10%, 25%, 50%, 75%, 90%). Az 5 pontos reprezentáció lehetővé teszi a jellemző formánsmenetek meghatározását is például a szűkebb és tágabb hangkörnyezet függvényében. A formánsadatok külön fájlban szerepelnek minden bemondóra vonatkoztatva.

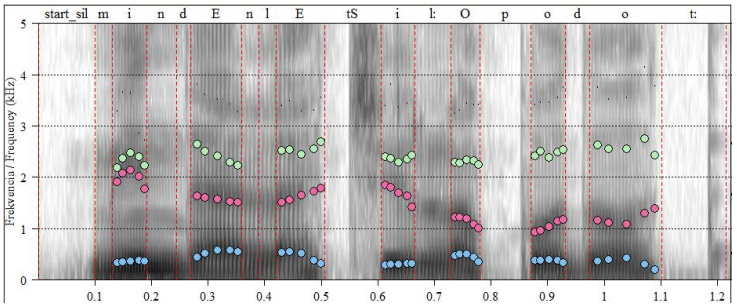
Hangsúlyjelölés: az adatázis szöveganyagát képező mondatkorpusz összes kijelentő mondatában, a mondat minden szavára elméleti hangsúlyozási címkét helyeztünk el bináris formában ([:W]=hangsúlyos/[:N]=nem hangsúlyos). Egy példamondat a feldolgozás bemutatására: [:W]*ehhez* [:N]*jön* [:N]*még* [:N]*a* [:W]*közjogi* [:N]*keret,* [:N]*és* [:N]*esetleg* [:N]*az* [:W]*új* [:W]*alkotmány*. A jelölések pontossága magas fokú, amit úgy kell érteni, hogy nincs benne jelölési hiba, vagyis ahol hangsúlyt jelöltünk, ott a hangsúlyos ejtés nem okoz megértési zavart és fordítva. Vannak olyan mondatok, amelyek több féle hangsúly kiosztással is ejthetők az értelmezés, illetve a közlési szándék szerint. Ezeknél a mondatoknál az egyik helyes formát tartalmazzák a jelölések.

A felsorolt nyolc szolgáltatás adatai szoros szinkronban vannak egymással. Ez biztosítja azt, hogy akár kereszt vizsgálatok is végezhetők (formáns-szóhossz; formáns-beszélő; szóhossz-beszélő stb.).

Az adatbázis felhasználása

A PPBA sokrétű adattár, amely hatékonyan felhasználható a beszédkutatásban, és a beszédtechnológiai fejlesztéseknél. Precíziós feldolgozottsága biztos alapot jelent a jövő kutatóinak. Nem kell a kutatónak adatgyűjtéssel, előzetes adatfeldolgozással foglalkozni, ezzel a kutatás hatékonysága növelhető.

Az adatbázis már most fontos szerepet tölt be a statisztikai parametrikus beszédszintézis fejlesztéseknél (Tóth 2013), valmint a magyar beszéd frománsterének újszerű modellezésénél és vizsgálatánál (Abari 2013). Egy külön erre a célra készített szoftver segítségével minden mondat formánsmenetei vizuálisan is tanulmányozhatók (4. ábra).



4. ábra

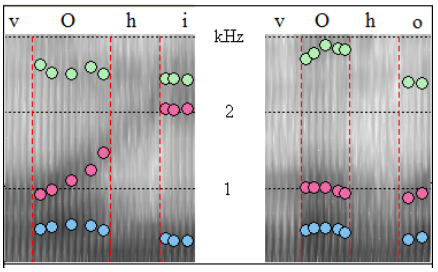
A *Minden lecsillapodott* ….férfi ejtésű hangsor formánsmenetei a mérési pontok értékeivel megjelenítve

Bármely formánsmeghatározó algoritmus hatásosságának vizsgálatát is alapozni lehet a formáns adatbázisra. A Praat-ra vonatkozó ilyen méréseket elvégeztük (1. táblázat), és az adatok azt mutatják, hogy beszélő függő a hibázás foka. A Praat automatikus mérő rendszere jól teljesített, 94,35%-99,22%-ban korrekten határozta meg a formánsadatokat. Ehhez jelentősen hozzájárult a precíziós címkézés is.

1. táblázat. A formánsjavítások száma %-ban megadva a teljes formánsadatbázisra vonatkoztatva a 10 bemondó szerinti elosztásban

FF1	FF2	FF3	FF4	FF5	NO1	NO2	NO3	NO4	NO5
2,50	5,36	0,78	2,37	1,77	1,35	2,42	5,72	4,47	6,65

Ugyancsak új eredmény, hogy a magánhangzó F2 formánsának mozgását nem csak a közvetlenül hozzá kapcsolódó hangok befolyásolhatják, hanem a tőle balra és jobbra elhelyezkedő -2. és +2. hangok is. Ez azt mutatja, hogy az artikulációs mozgások az egyes hangok, a képzési módok és helyek esetében már a közvetlenül csatlakozó hangon túl mutatva befolyásolják a magánhangzó képzését. Például más az [□] F2-jének a mozgása a [ϖ□ ɩ] hangsorban (..*szavahihető*…), mint a [ϖ□ o]-ban (..*tolvahozzá* …). Ezt a V utáni +2 hang befolyása okozza (5. ábra).



5. ábra
Példa arra, hogy a …*vah*… CVC szekvencia ejtése során a magánhangzó F2-jének menetét milyen erősen befolyásolja a magánhangzó utáni második [ɩ], illetve [o] beszédhang.

Az előbbiben felfelé mozog az F2, az utóbbiban lefelé. Ez a jelenség más hangkapcsolatoknál is kimutatható. A formánsadatbázis felhasználásával az ilyen jelenségek szisztematikusan felderíthetők és osztályozhatók. Ez hozzájárulhat a pontosabb formánsmozgások meghatározásához. A teljes rendszerezés után eljuthatunk oda, hogy a szöveg alapján a hangsor formánsmenetei jósolhatók lesznek sematikus formában, ami fontos lehet a beszédtechnológiai algoritmusok finomításához. Egy további felhasználási példa a glottalizált hangzást érinti. Az adatbázisban elhelyezett glottalizált hangra vonatkozó jelölések felhasználásával segítséget nyújtottunk egy új rendszerű glottalizált hangzást megszüntető javító algoritmus kidolgozásához.
Vizsgálatok folytathatók a gépi kényszerített beszédfelismerési algoritmusok hatékonyságának meghatározására is. Összehasonlítottuk a gépi feldolgozás eredményeit és a kézi javítások helyét és számát (xx.TextGrid és xx_man.TextGrid fájlok) ezzel képet kaptunk arról, hogy mely pontokon

gyengébb a gépi algoritmus teljesítőképessége, ez függ-e például a beszélő személy hangjától stb. (2. táblázat)

2. táblázat A kézzel javított hanghatárok száma a 10 beszélőre a kézi eltolások időintervallumaival

	10-19 ms	20-29 ms	30-39 ms	40-49 ms	50-59 ms	< 60ms
FF1	17128	5315	1657	552	185	169
FF2	9155	1869	760	383	203	262
FF3	23249	6285	1406	467	167	240
FF4	17850	4845	1240	428	198	272
FF5	9042	2491	941	525	331	558
NO1	10851	2450	1005	439	221	318
NO2	11062	2816	1003	432	190	456
NO3	11156	2603	924	435	227	418
NO4	13866	2321	658	228	91	93
NO5	9362	2401	878	442	253	375
Összesen	132721	33396	10472	4331	2066	3161

A táblázatból több dolog kiolvasható. Legtöbbször az egy-két zöngeperiódusra jellemző időtartam sávban kellett a hanghatárt a valós helyére tolni. Látható az is, hogy a gépi beszédfelismerő algoritmus hibázásai függnek a beszélő hangjától. Sok egyéb más mérés is végezhető lenne, azonban ennek a cikknek a terjedelme ezt nem teszi lehetővé.
Az adatbázis adataira támaszkodva egyedi számítási és keresési algoritmusokat bárki kidolgozhat és alkalmazhat. Ezzel szinte korlátlanul bővíthetjük a kutatási témák sorát. Csak néhány példát adunk: hangidőtartamok vizsgálata; fonológiailag hosszú hangok és kiejtésbeli megvalósulásaik felmérése; szünetek és osztályozásuk; szóhosszak kinyerése különböző szempontok alapján; bemondó-függő vizsgálatok; ritmikai mérések; hangstatisztika más paraméterek függvényében; formánsmenetek tulajdonságainak sokrétű vizsgálata; az elvi és gyakorlati hangsúlyozás közötti összefüggések elemzése; formánsmérő algoritmusok tesztelése; hangrealizációs összehasonlítások; oktatási adattárak kialakítása (hányféle akusztikai megvalósulásban jelennek meg ugyanazon hangok a folyamatos beszédben); betű-hang átalakító algoritmusok pontosságának vizsgálata.

Összefoglalás

A bemutatott beszédadatbázis sokoldalúan előkészített és feldolgozott adatok halmazait tartalmazza. Különlegessége a precíz adatfeldolgozás, ami jó alapot nyújt sokféle kutatáshoz. A PPBA 2012-ben a CESAR projekt részeként bekerült a META-SHARE nemzetközi nyelv és beszédtechnológiai adatbázis tárba, így része lett a nemzetközi kutatási beszédadatbázisoknak.

Köszönetnyilvánítás
A szerző köszönetet mond Laczkó Klárának az adatrendezés korrekt elvégzéséért, Bartalis Mátyásnak a kényszerített gépi felismerésben nyújtott segítségéért, valamint Tóth Bálintnak a hanghatárok szegmentálási adatain végzett összehasonlító mérés elvégzéséért. Mindhárman a BME TMIT munkatársai. Köszönet illeti Abari Kálmánt is, a Debreceni Egyetem kutatóját, aki a formánsok meghatározásában és a javításukhoz szükséges interaktív szoftvertámogatás készítésében végzett komoly munkát. A kutatást támogatta a CESAR project (Grant No. 271022).

Irodalom

Abari Kálmán – Olaszy Gábor 2012. Új módszer formánsadatok meghatározására és formánsadatbázis létrehozására nagyméretű beszédadatbázisokhoz. In: Gósy Mária (szerk) *Beszéd, adatbázis, kutatások.* Akadémiai Kiadó 2012. 251-268.

Abari Kálmán 2013. A formánsmozgások statisztikai vizsgálata és modellezése a magyar magánhangzókban. Doktori disszertáció, Debreceni Egyetem.

Boersma, Paul – Weenink, David 2005. Doing Phonetics by Computer. [Computer software], 2005. www.praat.org

Gósy Mária – Gyarmathy Dorottya – Horváth Viktória – Gráczi Tekla Etelka – Beke András – Neuberger Tilda – Nikléczy Péter 2012. BEA: beszélt nyelvi adatbázis. In: Gósy Mária (szerk.) *Beszéd, adatbázis, kutatások*. Akadémiai Kiadó 2012. 9-24.

Gósy Mária (szerk.) 2012. *Beszéd, adatbázis, kutatások.* Akadémiai Kiadó

Mihajlik, Péter – Révész, Tamás – Tatai, Péter 2002. Phonetic transcription in automatic speech recognition. *Acta Linguistica Hungarica 49/3–4*. 407–425.

Németh Géza – Olaszy Gábor (szerk.) 2010. *A Magyar beszéd – beszédkutatás, beszédtechnológia, beszédinformációs rendszerek.* Akadémiai Kiadó.

Olaszy Gábor – Bartalis Mátyás 2008. Jelfeldolgozási és fonetikai algoritmusok kombinációja a gépi hanghatárjelölés javítására. *Beszédkutatás 2008*. 208–220.

Tamm Anne – Olaszy Gábor 2005. Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához. In: *III. Magyar Számítógépes Nyelvészeti Konferencia.* Szerk.: Alexin Zoltán és Csendes Dóra. Szegedi Tudományegyetem Informatikai Tanszékcsoport. 2005 Szeged. 383-393.

Tóth Bálint 2013. Rejtett Markov-modell alapú gépi beszédkeltés. Budapesti Műszaki és Gazdaságtudományi Egyetem. Doktori értekezés.