Beszédadatbázis irodai számítógép-felhasználói környezetben

Vicsi Klára*, Kocsor András**, Teleki Csaba*, Tóth László**
*BME Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium
**MTA – SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport
E-mail: {vicsi, teleki}@tmit.bme.hu, {kocsor, tothl}@inf.u-szeged.hu

Összefoglalás: Az előadásban bemutatjuk az új, olvasott szöveget tartalmazó Magyar Referencia Beszédadatbázist, amelyet általános felhasználói környezetben, irodákban, laboratóriumokban, lakásokban rögzítettünk, összefoglaljuk az adatbázis létrehozásának akusztikai, nyelvi feldolgozási módszereit, ismertetjük az adatbázis pontos jellemzőit. Az adatbázis 332 beszélő olvasott szövegét tartalmazó hanganyag; beszélőnként 12 mondat és 12 szó speciális fonetikai elvárásoknak megfelelően összeállítva került bemondásra. A felvételek többféle mikrofonnal, hangkártyával és személyi számítógéppel készültek.

Cél: általános felhasználású, irodai, otthoni környezetben olvasott folyamatos szöveget tartalmazó beszédadatbázis létrehozása és akusztikai, valamint nyelvi feldolgozása, amely alkalmas PC-s beszédfelismerők betanítására, tesztelésére. A létrehozott beszédadatbázist *Magyar Referencia Beszédadatbázis*nak *(MRBA)* neveztük el.

Együttműködés: A BME TMIT Beszédakusztikai Laboratóriuma a szegedi SZTE Informatikai Tanszékcsoporttal együttműködve hozta létre ezt az adatbázist. A BME TMIT Beszédakusztikai Laboratóriuma végezte el a beszédadatbázis szöveganyagának megtervezését, az előfeldolgozást és annotálást, az automatikus betű – fonéma átírást; a szegedi SZTE Informatikai Tanszékcsoport a kézi szegmentálást végezte; a felvételek készítése megosztva történt.

A beszédadatbázis szöveganyagának megtervezése

A létrehozandó adatbázis számos különböző típusú beszédfelismerő betanítására és tesztelésére kell, hogy lehetőséget adjon. Ezek az izolált szavas rendszerek, szókereső és azonosító rendszerek, dialógus rendszerek, valamint folyamatos beszédet felismerő diktáló rendszerek, amelyeknél a felismerés szónál kisebb felismerési egységek modellezésén alapul. Ma már széles körben elterjedt, hogy a szófelismerésnél is szónál kisebb felismerési egységet használjanak, hiszen így nagyobb felismerési biztonság érhető el. A létrehozott adatbázist a mai követelményeknek megfelelően kellett előállítani.

Mivel a felismerés alapja ma már szavaknál kisebb egység, fonéma, difon, trifon, stb., olyan folyamatos szöveg összeállítására van szükség, ahol ezek az elemek elegendően sokszor fordulnak elő. A beszédfelismerési célokra rögzítendő beszédnek a lehető legjobban kell fednie a beszélt magyar nyelv sajátosságait. Mivel a felvett szöveg minden másodperce sok munkát von maga után, a szöveganyagnak minél rövidebbnek is kell lennie. Ezért alapos statisztikai vizsgálatokra van szükség a fonémák, difonok, trifonok szintjén egyaránt. Ennek alapján kell összeállítani olyan folyamatos szöveget, ahol a leggyakoribb di- és trifonok megfelelő mennyiségben állnak rendelkezésre. A folyamatos beszéd mellett fonetikailag gazdag szavakat is generálunk, egy sajátos algoritmus alapján, a folyamatos szöveg kiegészítéseként.

A mondatok összeállításához használt szöveg újságcikkeken alapult, kb. 1.6MB méretű és kb. 14000 mondatot tartalmazott. A szöveget fonémák sorozatára alakítottuk egy speciális algoritmus segítségével. Fonotipikus fonetikai átírást alkalmaztunk (Vicsi 2001, Fourcin, A.J. and Dolmazon, J-M.1991), vagyis a karakterek átírását a nyelv fonetikai szabályainak alapján végeztük el úgy, hogy a szövegkörnyezetet is figyelembe vettük (pl. koartikuláció, hasonulás).

Az ortografikus karakterek fonetikai átírásához SAMPA fonetikai jelölést használtunk (Vicsi, 2000). A fonémaátírás után következett a statisztikai vizsgálat. A fonéma, difon és trifon eloszlást vizsgáltuk az eredeti 1,6 MB méretű szövegadatbázison.

Egy bemondó 12 különböző mondatot és 12 különböző, a mondatoktól független szót olvas fel. A teljes szöveganyag 166 bemondásra készült el azért, hogy a szöveganyag 2-szer kerülhessen ismétlésre a 332 bemondás során. Így a teljes szöveganyag 12 x 166, azaz összesen 1992 db különböző mondatból és 12 x 166, azaz 1992 db különböző, a mondatoktól független szóból áll.

A mondatok kiválasztása a következő szempontok szerint történt: a fonémastatisztika alapján az 1%-nál gyakoribb fonémákat legalább egy példányban tartalmaznia kellett minden csoport valamelyik mondatának. Az összesen 69 darab fonémából ez 52-t jelentett. A mondatoknak legalább 30, legfeljebb 50 fonémából kellett állnia.

A mondatok kiválasztása a következő algoritmus alapján történt: Először elkészítettük a 166 db csoport helyét, amelyek először még üresek voltak. Minden csoporthoz tartozott egy fonéma-számláló, amely azt mutatta meg, hogy az adott csoportban melyik fonémából hány darab van. Először természetesen ezek is mind üresek voltak. Ezután a teljes szöveganyagon mondatonként végighaladva kiválasztottuk a soron következő még ki nem választott, és megfelelő hosszú mondatot, majd megvizsgáltuk, hogy van-e benne olyan fonéma, amelyik az aktuális csoportban a legritkább, és még nincs belőle egy sem. Ha megfelelt a mondat, akkor felvettük az adott csoport mondatai közé, és a saját fonémaszámláló táblázatát kiegészítettük az új mondat fonémáival. Ezután a feldolgozás a következő csoportnál folytatódott. Miután minden csoportba került 1 mondat, akkor következtek a 2. mondatok és így tovább. Ezzel a módszerrel egy olyan szöveganyagot kaptunk, ahol minden csoportban (azaz minden beszélőnél) van legalább egy előfordulása az 1%-nál gyakoribb fonémáknak (Táblázat 1.)

Táblázat 1. Fonémaelőfordulási statisztika első részlete az MRBA-ban, valamint az eredeti általános szöveganyagban.

Általános szöveganyag			***MRBA***		
	Fonéma előfordulási %	kumulatív %		Fonéma előfordulási %	kumulatív %
e	10.7122%	10.7122%	e	11.2489%	11.2489%
a	9.61378%	20.326%	a	9.75793%	21.0069%
t	6.70365%	27.0297%	t	6.63625%	27.6431%
l	5.3106%	32.3403%	l	5.5006%	33.1437%
k	4.91532%	37.2556%	k	4.76966%	37.9134%
i	4.71455%	41.9701%	o	4.40166%	42.315%
o	4.46948%	46.4396%	i	4.37887%	46.6939%
n	4.25112%	50.6907%	n	4.36537%	51.0593%
r	4.22586%	54.9166%	r	3.9577%	55.017%
á	3.84454%	58.7611%	m	3.76821%	58.7852%
é	3.56529%	62.3264%	á	3.52977%	62.3149%
m	3.44962%	65.776%	é	3.47955%	65.7945%
s	3.26176%	69.0378%	s	2.87944%	68.6739%
j	2.47277%	71.5106%	z	2.17551%	70.8494%
z	2.25925%	73.7698%	g	2.15061%	73.0001%
g	2.0971%	75.8669%	v	2.11305%	75.1131%
v	2.09018%	77.9571%	j	2.089%	77.2021%
d	1.82915%	79.7862%	d	1.94002%	79.1421%
sz	1.76941%	81.5557%	h	1.71213%	80.8543%
h	1.61713%	83.1728%	sz	1.71087%	82.5651%
b	1.56281%	84.7356%	b	1.57624%	84.1414%
p	1.17186%	85.9075%	ö	1.17786%	85.3192%
ö	1.14107%	87.0485%	gy	1.06982%	86.389%
gy	1.07891%	88.1274%	u	1.05378%	87.4428%
u	1.07072%	89.1982%	p	1.0196%	88.4624%
ó	1.02263%	90.2208%	f	0.979929%	89.4423%
f	0.995704%	91.2165%	ó	0.928864%	90.3712%
õ	0.956379%	92.1729%	õ	0.892571%	91.2638%
tt	0.647427%	92.8203%	tt	0.822094%	92.0859%
ny	0.641718%	93.462%	ny	0.675653%	92.7615%

í	0.631455%	94.0935%
ü	0.620268%	94.7137%
ll	0.536486%	95.2502%
n(k/g)	0.453971%	95.7042%
cs	0.420009%	96.1242%
c	0.391351%	96.5156%
ú	0.328326%	96.8439%
kk	0.296151%	97.14%
m(f/v)	0.258613%	97.3986%
mm	0.22638%	97.625%
ty	0.214502%	97.8395%
û	0.206314%	98.0458%
nn	0.197261%	98.2431%
zs	0.190572%	98.4337%

ü	0.674387%	93.4359%
í	0.573524%	94.0094%
cs	0.493341%	94.5028%
ll	0.483212%	94.986%
ú	0.423707%	95.4097%
nk	0.421175%	95.8309%
c	0.336349%	96.1672%
mm	0.320735%	96.488%
kk	0.296258%	96.7842%
nn	0.280221%	97.0644%
ty	0.261652%	97.3261%
bb	0.232955%	97.559%
û	0.232111%	97.7912%
mv	0.224514%	98.0157%

Ez a mondatkiválasztó algoritmus úgy készült, hogy ha valamelyik csoportból hiányzott volna egy fonéma, akkor azt jelezte volna.
A teljes szöveg és a válogatott szöveg difon statisztikáinak alapján a leggyakoribb 98.8% difon mindegyike szerepel mindkét anyagban, és csak itt kezdenek el hiányozni a válogatott szövegből a difonok. A következő táblázatokban a teljes szöveg, illetve a fonetikailag gazdag mondatok difon statisztikájának egy részlete látható (Táblázat 2.).

Táblázat 2. difon statisztika első részlete az MRBA-ban, valamint az eredeti általános szöveganyagban

Általános szöveganyag

Fonéma	előfordulási %	kumulatív %
el	1.39067%	1.39067%
ta	1.12385%	2.51452%
te	1.05964%	3.57416%
et	1.02421%	4.59837%
ak	0.970271%	5.56864%
al	0.95813%	6.52677%
at	0.917853%	7.44462%
le	0.91114%	8.35576%
ij	0.884755%	9.24052%
en	0.859655%	10.1002%
er	0.85697%	10.9571%
ek	0.855452%	11.8126%
az	0.805719%	12.6183%
ne	0.76118%	13.3795%
em	0.73894%	14.1184%
me	0.736605%	14.855%
na	0.725339%	15.5804%
an	0.702341%	16.2827%
ka	0.696678%	16.9794%
ke	0.688565%	17.668%
la	0.680509%	18.3485%
nt	0.677415%	19.0259%
am	0.62739%	19.6533%
in	0.625289%	20.2786%
re	0.618576%	20.8971%
ár	0.614431%	21.5116%
és	0.607135%	22.1187%
lt	0.585537%	22.7042%
or	0.578007%	23.2823%
mi	0.568959%	23.8512%
be	0.551914%	24.4031%
ze	0.534344%	24.9375%
ol	0.533877%	25.4713%
ve	0.53055%	26.0019%
tá	0.525646%	26.5275%
ás	0.518875%	27.0464%
ha	0.507959%	27.5544%
to	0.505508%	28.0599%
ik	0.505508%	28.5654%

MRBA

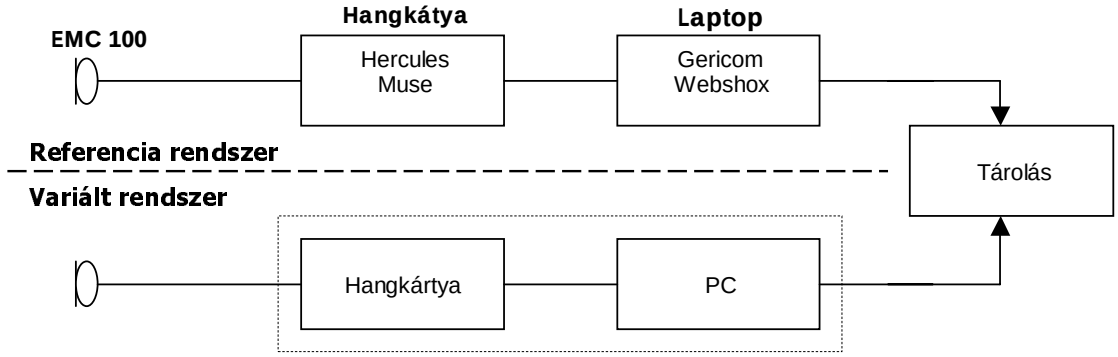
Fonéma	előfordulási %	kumulatív %
el	1.3626%	1.3626%
ta	1.13052%	2.49312%
te	1.02963%	3.52275%
le	1.01664%	4.53939%
et	0.996727%	5.53612%
em	0.969016%	6.50514%
ak	0.95949%	7.46463%
al	0.943903%	8.40853%
ne	0.904068%	9.3126%
ek	0.874625%	10.1872%
at	0.864234%	11.0515%
lt	0.823101%	11.8746%
en	0.819637%	12.6942%
er	0.807513%	13.5017%
az	0.804482%	14.3062%
an	0.756421%	15.0626%
ke	0.746896%	15.8095%
me	0.739968%	16.5495%
ol	0.706628%	17.2561%
ij	0.696669%	17.9528%
mi	0.690175%	18.6429%
ka	0.658567%	19.3015%
re	0.654237%	19.9557%
na	0.629124%	20.5849%
in	0.621763%	21.2066%
ha	0.604877%	21.8115%
am	0.590589%	22.4021%
nt	0.581496%	22.9836%
be	0.568074%	23.5517%
la	0.566775%	24.1184%
ez	0.543394%	24.6618%
ár	0.516982%	25.1788%
eg	0.508322%	25.6871%
va	0.507889%	26.195%
ve	0.500961%	26.696%
ze	0.499662%	27.1957%
rt	0.487106%	27.6828%
ar	0.485807%	28.1686%
ér	0.479745%	28.6483%

rt	0.491615%	29.057%
szt	0.49144%	29.5484%
té	0.487179%	30.0356%
eg	0.480407%	30.516%
se	0.479765%	30.9958%

szt	0.470652%	29.119%
ho	0.465457%	29.5844%
se	0.464158%	30.0486%
ik	0.461993%	30.5106%
nd	0.455931%	30.9665%

A beszédadatbázis rögzítése

A beszédadatbázis felvételeit különböző helyszíneken: zajos, kevésbé zajos irodai helyiségekben, laborokban, otthonokban vettük fel. A felvételeknél szinkronban két különböző rendszerrel dolgoztunk. Az egyik az ún. *referenciarendszer*, ahol jó minőségű, közelbeszélő kondenzátor mikrofont (Monacor EMC 100), jó technikai paraméterekkel rendelkező hangkártyát (Hercules Muse Pocket USB 5.1), valamint egy adott Gericom Webshox típusú laptopot használtunk. A másik rendszernél az ún. *variált rendszer*, különböző, jobb, kevésbé jó mikrofonokat, hangkártyákat, PC-ket használtunk, a lehető legnagyobb variáltsággal. A felvételi elrendezést az 1. ábra mutatja.



1. *ábra* Az MRBA adatbázis felvételi elrendezése

Felvételeknél a bemondók, a PC-k, a hangkártyák, a mikrofonok és a környezet minél nagyobb variáltságára törekedtünk.

A beszélők demográfiai adatai

Régiók és dialektusok: Egy-egy dialektust reprezentáló tiszta anyagot nehéz felvenni, két okból is. Egyrészt Magyarországon a lakosság keveredése miatt nincsenek olyan egybefüggő területek, ahol jellemzően egy dialektust beszélnek. Másrészt a hivatalos oktatásban, a médiában egyeduralkodó a magyar köznyelv használata. A fiatalabb generációk hivatalos környezetben, mikrofon előtt jellemzően ezt a köznyelvet használják, még akkor is, ha otthon, kötetlenebb viszonyok között dialektusban beszélnek. A felvételeket Magyarország 4 különböző tájegységében lévő városban rögzítettük: Budapesten, Szegeden, Győrben és Miskolcon.

A felvételek során a beszélők születési helyét és jelenlegi lakhelyét jegyeztük fel.

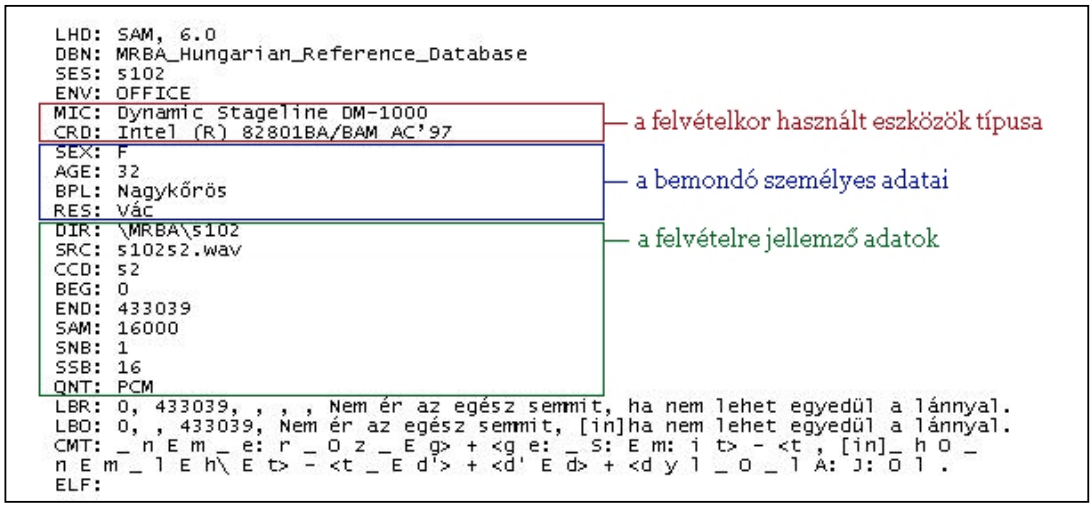
Életkor, nemek: A felvételek kor és nem szerinti eloszlása a következő táblázat szerint alakult.

Táblázat 3. A beszélők életkor és nem szerinti megoszlása

Korcsoport	Férfi beszélők	Női beszélők
16 év alatt	0,9 %	3,3 %
16 – 30 év	46,1 %	27,7 %
31 – 45 év	5,7 %	6 %
46 – 60 év	3,9 %	5,1 %
60 év felett	0,9 %	0,4 %
Összesen	*57,5 %*	*42,5 %*

Előfeldolgozás és annotálás

Az előfeldolgozás a felvételek lehallgatásából, ellenőrzéséből és esetleg feldarabolásából, esetünkben a két csatorna szinkronizálásából áll.
Az annotálás annyit jelent, hogy minden hangfájl mellé egy címkefájlt készítünk, amely különféle információkat tartalmaz a hangfájl paramétereivel és tartalmával kapcsolatban: az elhangzott szöveg ortografikus lejegyzését, hibás kiejtést, nem érthető szavakat, szótöredékeket, a beszélő nem beszédből származó hangjait, környezeti zajokat, stb. (Wells, J. 2001). Az alábbi ábrában egy ilyen címkefájlt prezentálunk.



2.*ábra* Az MRBA adatbázisban használt címkefájl.

Az átírás során az alábbi szabályokat vettük figyelembe:

- minden mondatkezdő nagybetű, tulajdonnévkezdő nagybetű, illetve mondatrész, mondatvégi írásjelek, gondolatjelek megmaradtak.
- a bemondó által elkövetett nyelvbotlásokat, értelmetlen szótöredékeket kiejtés szerint írtuk le.
- szó kihagyása esetén az elhagyott szót az átiratból töröltük.
- a betűszavakat (Rt., Kft.) betűzéseket, vagy idegen szavakat (EGIS) az átiratban a helyes kiejtésnek megfelelően jegyezzük le (Erté, Káefté, Égisz), amennyiben idegen szavaknál az ejtés a helyesnek tekinthetőtől eltér, csillaggal jelöltük.
- a számokat betűkkel, kiejtés szerint írtuk le.
- a háttérzajokat az elhangzás helyén jelöltük. Ha szó közben történt, akkor a zajos szó elején, ha szavak határánál, vagy a szavak közötti csendben, akkor a két szó közé a szavaktól szóközzel elválasztva jelöltük.

A fájlban lejegyzésre került a *CMT* sorban a bemondott szöveg fonotipikus átirata.

Fonotipikus automatikus betű – fonéma átírás, szóhatár megadás

A bemondott szöveg betűit átírtuk úgy, hogy figyelembe vettük a magyar beszéd fonetikai szabályait a szövegkörnyezet függvényében. Ilyenek például az alkalmazkodási folyamatok (hasonulás, asszimiláció, stb.).

A folyamatos beszédre tipikusan jellemző, hogy a szavak között nincs szünet. A fizikai paraméterek folyamatosan változnak a szóhatárokon. Ezért a további feldolgozás megkönnyítése érdekében bejelöltük a szavak határait is.

Kézi szegmentálás

Az adatbázis annotálása után következő feladat annak fonetikai szintű szegmentálása és címkézése, valamint a szóhatárok és a frázishatárok bejelölése. Kézi, illetve fél automatikus javított szegmentálással a beszéd időfüggvényében be kell jelölni a fizikailag megfigyelhető fonémák határait és beírni a megfelelő helyre a fonéma címkéket. A feladat „audio-vizuális fonetikai átírás" a 1997-ban elkészült BABEL nemzetközi project leírása alapján (Vicsi K., Vig A.). Az átírást a szöveg hallgatása, és az időfüggvény és/vagy a színkép elemzése alapján hajtottuk végre.
A feldolgozást három szinten végeztük: fonémák szintjén, szavak szintjén és a mondatok szintjén. **Fonémaszintű szegmentálás** esetén a következő szabályrendszer alapján végeztük el:

- a szegmenshatárokat a null-átmenetekhez illesztettük (zöngés hangok esetében a pozitív meredekségű null-átmenethez). A bejelölés pontossága lehetőleg min. 1 ms legyen.. Zöngétlen hangoknál ez azt jelenti, hogy 1 ms pontossággal jelöljük be a zöngétlen hangok kezdetét. A zöngés hangok esetében mindig a null-átmenetekhez igazodtunk szintén 1 ms pontossággal. Zöngés hangok között a határ bejelölésének bizonytalansága alap-periódusnyi hosszú.
- zöngétlen hang után a magánhangzó kezdetét a zönge indulásánál jelöljük.
- zárhangok, affrikáták jelölésénél ezen hangok kezdetét a megelőző hang utolsó lecsengő periódusa előtt jelöltük, vagyis a megfelelő hang utolsó periódusát a zárhangok és az affrikáták részének tekintettük.
- a magánhangzó-magánhangzó vagy magánhangzó-rezonáns mássalhangzó kapcsolatokban a határt az átmeneti rész 50%-ánál jelöltük be. A határ pontos bejelölése esetén egy-két periódusnyi ingadozást engedtünk meg. Ez itt nem hiba, hanem a feladat természetéből adódik, hiszen folyamatos mozgás által létrehozott hangproduktumok kettéválasztása ekkor bizonytalan. Rendszerint két különböző szakértő a határt két különböző helyre teszi.

Címkézésre SAMPA karaktereket használtunk. A karaktereket a határbejelölések között tüntettük fel. Azt jegyeztük le, ami ténylegesen elhangzott.
Allophonok esetén három különböző jelölést alkalmaztunk:

' Minden magánhangzónál létezik egy rekedt ejtés, ami igen gyakori. Jelölése a magánhangzó SAMPA karaktere előtt egy **'**. (Pl. **'o** jelenti az o hangot rekedt ejtésben.)
h A ch betű kapcsolatot h hangnak ejtjük.***
h A h hangnak csak egy allophonját ejtjük, a zöngés h hangot.

Az akusztikailag többkomponensű fonémákat – zárhangok, affrikáták – három komponensre bontottuk:

> – a zárképzés szakaszát a mássalhangzó SAMPA karaktere után jelöltük (pl**. t>**),
+/- – zöngés/zöngétlen zárszakasz jelölése,

< – zárfelpattanási zörej szakasza, affrikáták esetén s spiráns zörejjel együtt jelölétük a következőképpen: a mássalhangzó SAMPA karaktere elött egy < (pl. **<t**) .

Az aktuális kimondásnál kimaradt hangokat megjelöltük, a követő hang előtt zárójelbe tettük. Természetesen az annotálásban szereplő jelöléseket megtartjuk, azokat a kézi szegmentálás során pontosan bejelöltük. A beszéd közben tartott szünetet ~ jellel jelöltük. Szöveg betoldását vagy kihagyását a bemondó által *-al jelöltük a szegmentálás során.

A **szószintű szegmentálás** esetén a szavak határát jelöltük. A **mondatok határait** a frázisjelző írásjelek (vessző, kettőspont, pontosvessző, gondolatjel, macskakörröm, és, pont kérdőjel, felkiáltójel) adták.

A Magyar Referncia Beszédadatbázis adatkészleteinek átfogó műszaki tulajdonságai:

Magyar nyelvű, olvasott szövegű, személyi számítógépes környezetben felvett adatbázis, 16 bites, 16 kHz-es mintavételezéssel.

- *332 beszélő közvetlenül a számítógépbe rögzített hanganyaga;*
- *Beszélőnként 12 mondat és 12 szó, speciális fonetikai elvárásoknak megfelelően összeállítva;*
- *A felvételek többféle* mikrofonnal, hangkártyával, PC-kel *készültek, nagy variáltsággal;*
- *Felvételi környezet változó zajosságú irodahelyiség, laboratórium, otthoni környezet;*
- *Az adatbázis teljes anyaga annotált, az adatbázis harmada (100 beszélő) kézileg szegmentált és címkézett.*

Köszönetnyilvánítás

A kutatás az OTKA T 046487 ELE és az IKTA 00056 pályázatok keretén belül készül.

Irodalom

Vicsi, K. Beszédadatbázisok a gépi beszédfelismerés segítésére, Híradástechnika, Vol. 2001/1, Budapest, pp. 5-13, 2001.

Vicsi, K. Vig, A.: Az első magyar nyelvű beszédadatbázis, Beszédkutatás'98, Tanulmányok az elméleti és alkalmazott fonetika köréből. MTA Nyelvtudományi Intézet, Budapest, pp. 163-178, 1998.

Wells, J. at all.: Standard Computer-Compatible Transcription. Esprit Project 2589 (SAM), Doc. no. SAM-UCL-037. London: Phonetics and Linguistics Dept., UCL (1992).

Fourcin, A.J. and Dolmazon, J-M. „Speech knowledge, standards and assessment", Proceedings of XII International Congress of Phonetic Sciences, Aix-en-Provence, Vol. 5, 430-433 (1991).