AUTOMATIKUS BESZÉDFELISMERÉSHEZ HASZNÁLT BESZÉDHANGMODELLEK BETANÍTÁSI MÓDSZEREINEK ÖSSZEHASONLÍTÓ ELEMZÉSE

Fegyó Tibor – Mihajlik Péter – Tatai Péter

Bevezetés

A mai beszédfelismerő rendszerekben a felismerés legkisebb egysége általában a beszédhang. A beszédhangokhoz akusztikus modelleket rendelünk, amelyek paramétereit statisztikai módszerekkel határozzuk meg. A megfelelő betanító hanganyag alapvetően meghatározza a beszédfelismerő hatékonyságát. A mai technológiai és számítási kapacitás mellett nagy szótárak és bonyolult nyelvi modellek kezelésére is alkalmasak a gépi beszédfelismerő rendszerek. Mindezeknek azonban jelentős korlátot szab, ha az elemi mintaillesztési egységek nem elég hatékonyak.

Kísérleteink során megvizsgáltuk, hogy az akusztikus modellek betanításának különböző módszerei hogyan befolyásolják a felismerési eredményeket.

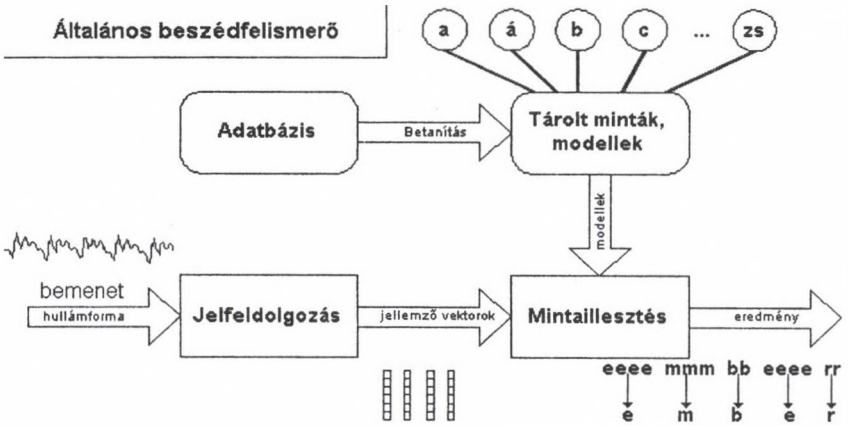
A beszédfelismerés alapjai

A beszédfelismerő rendszerek bonyolultságuktól függően különállóan bemondott szavakat, szókapcsolatokat, esetleg folyamatos beszédet képesek írott szöveggé átalakítani. A gépi beszédfelismerés elve alapvetően összehasonlításon alapul, tárolt modellekkel kell összevetni a beérkező jelsorozatot, és ki kell választani a lehetséges jelöltek közül a legjobbat. Először röviden bemutatjuk egy általános beszédfelismerő felépítését, majd részletezzük a mintaillesztő eljárás során használt akusztikus modellek betanítását.

1. A beszédfelismerők felépítése

Az 1. ábrán látható egy általános gépi beszédfelismerő felépítése. A bemenet egy hullámforma, a kimenet pedig az írott szöveg. A jel-

feldolgozás spektrális jellemzőket állít elő, amelyek közül leghatékonyabbnak a mel frekvencia alapú MFCC (Mel Frequency Cepstral Coefficients) vektorok bizonyultak (Young et al. 1999). Ezeket a vektorokat kell a tárolt minták vektorsorozatával összehasonlítani. A mintaillesztésre ma szinte kizárólag statisztikai alapú eljárásokat, főként rejtett Markov-modelleket (HMM) (Rabiner 1993) alkalmaznak. A mintaillesztéshez használt modelleket adatbázisok alapján kell betanítani.



1. ábra
Általános beszédfelismerő felépítése

Elvileg tetszőleges nyelvi egységhez rendelhetünk modelleket, ha van elegendő tanító adatunk. Így az ábrán látható beszédhang alapú megközelítés mellett lehetőség van hangkapcsolatok, szavak modellezésére is. A hosszabb egységek használatának a tanító adatbázis mérete mellett a rugalmasság igénye is korlátot szab. A mintaillesztő eljárás ugyanis nem tetszőleges hangsorozatot ismer fel, hanem a szótárak, és a nyelvi modellek is korlátozzák a felismerési halmaz méretét. Rövidebb mintaillesztési egységek esetén rugalmasabban lehet a szótárakat módosítani, és több a betanító elem is.

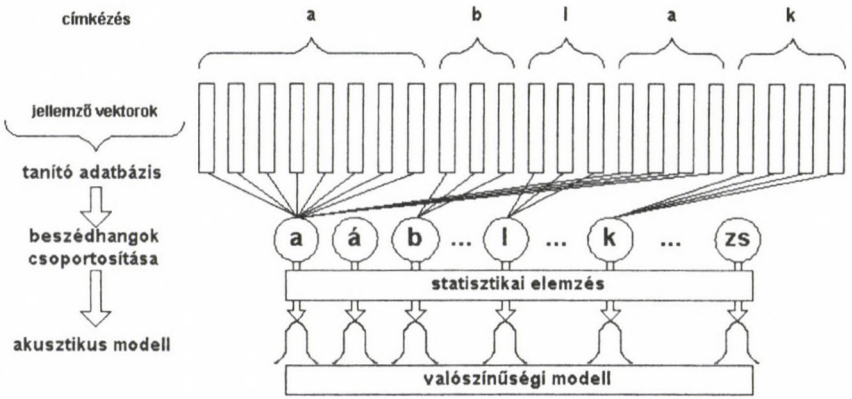
A beszédfelismerőkben a betanított modellek, (az úgynevezett

akusztikus modellek) jelentik a kritikus pontot. Rosszul betanított modellekkel nem lehetséges hatékony felismerőt készíteni.

A kutatások a kötött nyelvtanoktól a természetes nyelvi felismerés irányába haladnak. Ehhez a nyelvi hálózatnak egyre kiterjedtebbnek kell lennie, és ezzel nő a felismerő által elfogadott szavak, mondatok száma is. A növekedésnek az akusztikus modellek minősége szab határt. Minél pontosabbak az akusztikus modellek, annál nagyobb szótár vagy nyelvi hálózat alkalmazható és viszont, a szótárméret növelésével az akusztikus modelleknek is pontosabbaknak kell lenniük ahhoz, hogy a felismerési eredmények ne romoljanak.

2. Betanítás – az akusztikus modellek meghatározása

A beszédfelismerő rendszer betanítása alapvetően az akusztikus modellek meghatározását jelenti. A betanítás menete a 2. ábrán látszik. A konkrét matematikai lépésekre nem térünk ki, megtalálhatóak például (Rabiner 1993)-ban.



2. ábra
Az akusztikus modellek betanítása

Első lépésként szükség van egy adatbázisra, ami beszédhang-, szó- vagy mondatszinten címkézett hanganyagot jelöl. Ebből elő kell állítani az MFCC alapú jellemző vektorokat. Az egyes beszédhangok

összes előfordulását statisztikai módszerekkel elemezni kell, és ennek eredményeként áll elő az akusztikus valószínűségi modell.

Beszédhangonként egy-egy eloszlást tárolunk, ami a jellemző vektorok alapján kerül becslésre. Ez az eloszlás több normális eloszlás keverékeként áll elő. A mintaillesztés során a felismerendő jellemző vektorokat ezekkel a modellekkel hasonlítjuk össze.

Abban az esetben, ha az adatbázis nincs beszédhangszinten címkézve, akkor automatikus címkézésre van szükség. Az automatikus címkézés nagy adatbázisok esetén hatékonyan alkalmazható, részleteiről a következő fejezetben lesz szó.

Kísérleteink során a BABEL (Vicsi–Víg 1997), Huncities (Szarvas et al. 2000), MTBA (Vicsi et al. 2002) és Speechdat (Vicsi et al. 1999) adatbázisokat alkalmaztuk. Az első kettő 20-30 beszélő hangját tartalmazza, de jó minőségben, míg az utóbbi kettőben több száz beszélő telefoncsatornán felvett hangja található. Az adatbázisok részben beszédhangszinten címkézettek, részben csak a szó-, illetve mondatszintű annotálás történt meg. A betanítás során ezen adatbázisok különböző részeit alkalmaztuk. A felismerési kísérleteket általában a Speechdat adatbázis városneveket tartalmazó részén végeztük, amelyben közel ötszáz beszélő a szótárban szereplő 330 városnév egyikét mondta be.

A betanítás hatékonyságának vizsgálata

Megvizsgáltuk, hogy a különböző módon betanított akusztikus modellek hogyan befolyásolják a felismerési eredményeket.

1. A beszélők száma

A tanításhoz adatbázisra van szükség, ami sok beszélő hangját tartalmazza. Általános szabály, hogy minél nagyobb az adatbázis, annál jobbak lesznek a modellek. Az adatbázis elkészítése azonban meglehetősen idő- és munkaigényes feladat. A tanító adatbázisban szereplő személyek számát növeltük, és megvizsgáltuk, hogy 13, 25, 50, 100 beszélő hangjával tanítva hogyan változik a felismerési arány adott, kisszámú hangmodell mellett. A tanítást az MTBA adatbázis kézzel

címkézett részével, míg a felismerési kísérleteket a Speechdat városneveivel végeztük.

Az 1. táblázatban látható eredményekből kitűnik, hogy a beszélők számának emelése valóban hatékony, de 50-100 beszélő között már nincs érdemi különbség. A beszélők számának további növelése nem hoz jelentős javulást a modell egyéb paramétereinek (jelfeldolgozási paraméterek, eloszlás keverékszám stb.) módosítása nélkül.

1. táblázat: A beszélők számának hatása

Beszélők száma	**Felismerési hiba**
13	19,3%
25	15,0%
50	10,6%
100	11,1%

2. A fonetikus átírás hatása

A tanító adatbázis ideális esetben teljes egészében beszédhangszinten címkézve van. A teljes címkézés azonban nagyon erőforrásigényes feladat, ezért alternatív megoldásokat kell keresni. Különböző automatikus címkézési technológiákat dolgoztak ki, mi az ún. kényszerített illesztést (forced alignment) (Yuang 1999) alkalmaztuk.

A kényszerített illesztés egy betanított felismerő segítségével történik, amely betanításához kézzel címkézett adatokra is szükség van. A tanító adatbázis egy részét beszédhang szinten kézzel szegmentálták, így azzal lehetett a kezdeti felismerőt betanítani. Megjegyzendő, hogy vannak technológiák szegmentálatlan adatbázissal történő tanításra is (Yuang et al. 1999), azonban ezek kevéssé hatékonyak.

A címkézendő beszédnek annotáltnak kell lennie, azaz tudnunk kell, mi hangzott el a felvételen. Az írott formában adott szöveget át kell alakítani beszédhangok sorozatává, majd a felismerőnek mint egyetlen lehetséges felismerendő sorozatot kell ezt megadni. A mintaillesztési algoritmus azon túl, hogy a lehetséges jelöltek közül kiválasztja a legvalószínűbbet, implicit módon elvégzi a beszédhangszintű szegmentálást is, megadja, hogy melyik hang hol kezdődött és hol végződött. Mivel egy felvételhez egyetlen felismerendő sorozat tarto-

zik, itt a legvalószínűbb jelölt keresésének nincs értelme, a művelet eredménye hangszintű címkézés lesz.

A kényszerített keresés technológiája ismert, azonban kérdés, hogy pontosan mit is kell keresni, azaz hogyan történjen az írott szöveg átalakítása kiejtett formára. Ehhez meg kellett oldani szavak, szókapcsolatok, mondatok automatikus fonetikus átírását (Fegyó et al. 2001; Mihajlik et al. 2002). Szószinten felmerültek graféma szegmentálási problémák is, például: *kulcszörgés, láncszem*, ahol a *csz* szegmentálása magasabb szintű nyelvi információk nélkül nem egyértelmű. Továbbá, bizonyos grafémakapcsolatok esetén a kiejtés sem egyértelmű, mint például a *látja, átjáró* szavaknál, ahol a *tj* kapcsolatot [t j], illetve [ty]-nek kell ejteni, az *apátság* szónál, ahol a [ccs], illetve a [t s] egyaránt helyes kiejtés. Hasonló problémák lépnek fel az *ezüst, ezüstbánya, ébresztget,* illetve sok hasonló szónál is.

Szókapcsolatok szintjén a bizonytalan szünetek okozhatnak problémát. Bár az írott szövegben a vesszők, pontok egyértelműek, azonban a beszélők változó helyen tartanak szünetet, vesznek levegőt. Ahol a szavak között nincs szünet, ott figyelembe kell venni a szóhatáron fellépő hangmódosulásokat is, ahol pedig szünet van, ott nincs hasonulás.

Ha a kényszerített illesztéshez használt fonetikus átirat nem helyes, akkor pontatlan, inkonzisztens lesz a címkézés, és így az egyes hangok modelljeinek tanításakor, helytelenül, idegen hangokat is figyelembe veszünk.

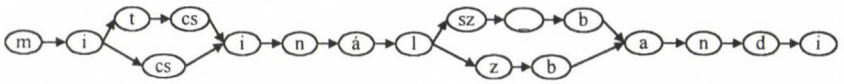
A címkézés hatását háromféle módszerrel vizsgáltuk:

a) automatikus címkézés kézi fonetikus átírás alapján, itt a kézi címkézésből elhagytuk a címkehatár-információt;
b) automatikus lineáris címkézés, amikor csak egy lehetséges kiejtési módot enged meg a rendszer;
c) automatikus opciós címkézés, amikor az összes lehetséges alternatívát megengedi a fonetikus átirat.

A lineáris címkézés nyilvánvalóan hibát visz a rendszerbe, tehát azt várjuk, hogy az kevésbé jól teljesít.

Az opciós modell a lineáris modellel szemben valódi felismerési feladatot is takar, mivel el kell dönteni, hogy a lehetséges kiejtési

variációk közül melyik hangzott el. Egy példa látható a 3. ábrán az alternatív kiejtési útvonalak ábrázolására.



3. ábra
Példa opciós kiejtési átiratra

A kísérletben tehát egy betanított felismerővel az MTBA adatbázist háromféle fonetikus átirat alapján automatikusan felcímkéztük, majd a címkézett adatokkal betanítottunk egy-egy felismerőt, és teszteltük a Speechdat városnevein.

2. táblázat: A fonetikus átírás módjának hatása

Fonetikus átírás módja	**Felismerési hiba**
a) kézi	15,5%
b) automatikus, lineáris	20%
c) automatikus, opciós	13,5%

A 2. táblázatból látható, hogy az opciós automatikus átirat teljesített a legjobban, mind a lineáris, mind a kézi átírásnál jobb felismerési eredményeket kaptunk. Az első táblázatban a címkézés kézzel, míg itt automatikusan történt, ezért magasabb itt a kézi fonetikus átíráshoz tartozó hibaarány.

3. A tanító hanganyag minőségének hatása

A tanító hangadatbázisok általában meghatározott körülmények között kerülnek felvételre. Amennyiben a betanított felismerőt más körülmények között szeretnénk használni, a felismerés hatékonysága romlik. Ennek oka, hogy az adott körülmények között jellemző zajok, torzítások máshol nem, vagy eltérően jelentkeznek.

Megvizsgáltuk, hogy telefoncsatornán felvett beszédet telefonos, illetve stúdióminőségű felvétellel tanított felismerővel milyen haté-

konysággal lehet felismerni. A 3. táblázatban látható, hogy a jobb minőségű adatbázis ellenére közel kétszer akkor a felismerési hiba, mint a zajosabb telefonos adatbázis esetén. A jó minőségű (BABEL) tanító adatbázis kisebb, mindössze 25 beszélőt tartalmaz, ezért a telefonos adatbázis (MTBA) esetén a szűkített tanítást is elvégeztük 25 beszélővel. Felismeréshez a telefonos Speechdat, illetve a jó minőségű Huncities adatbázis városneveit használtuk.

3. táblázat: A tanító hanganyag minőségének hatása

Tanító adatbázis	**Teszt adatbázis**	**Felismerési hiba**
Stúdió minőségű	Stúdió minőségű	13%
Stúdió minőségű	Telefon minőségű	24%
Telefon minőségű	Telefon minőségű	12%
Telefon minőségű (25 beszélő)	Telefon minőségű	15%

Jelentős kutatási területet képviselnek a kompenzációs, adaptációs módszerek, amelyek biztosítják az átjárhatóságot a különböző minőségű rendszerek között, azonban még csak speciális területeken születtek kísérleti megoldások. Megvizsgáltuk a csatornakarakterisztika kompenzálásának lehetőségét. A telefonon felvett adatok alapján meghatároztunk egy szűrő karakterisztikát, ami a csatornát jellemzi. Ezt a szűrőt alkalmaztuk a jó minőségű felvételeken, majd a szűrt jellel tanítottuk be a felismerőt. Ezzel a felvételek közötti lineáris torzítást kompenzáltuk. Számszerű mérések erre vonatkozóan még nem történtek, de szubjektív kísérletek alapján javult a felismerés hatékonysága.

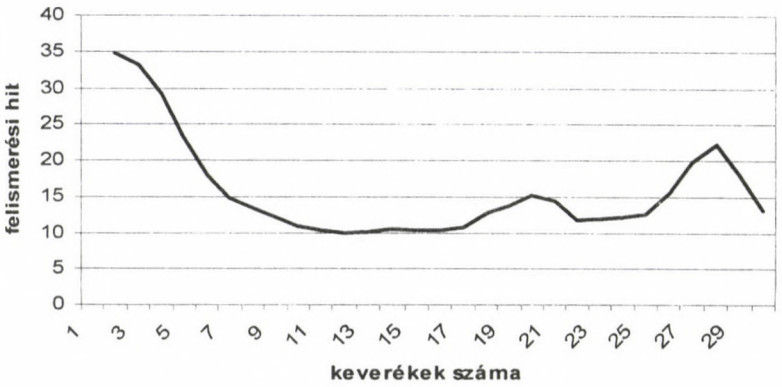
4. A modell bonyolultsági fokának hatása

Az akusztikus modellben az egyes hangok spektrális jellemzőinek tényleges eloszlását normális eloszlások keverékével reprezentáljuk. A keverékek száma szabadon megválasztható. Megvizsgáltuk, hogy milyen hatása van, ha alacsonyabb vagy magasabb keverékszámot választunk.

Ha mindenki közel azonosan ejtené az egyes hangokat, akkor elegendő lenne egyetlen eloszlás is, azonban vannak jelentős eltérések,

gondoljunk például a férfi és a női hangokra. A keverékek számának növelésével a kiejtési változatosságot hatékonyabban kezeli a modell.

Meddig érdemes növelni a keverékek számát? Erre több válasz is lehetséges. Amíg a felismerés pontossága nő, növelhető a keverékek száma. Bizonyos mennyiség fölött viszont már romlani fog a felismerés, mert nem lesz elég adat az egyes keverékek tanítására. További fontos szempont a tanítási, illetve feldolgozási idő, mivel a keverékek számának növelésével arányosan nő a feldolgozási idő is.



4. ábra
A keverékek számának hatása

A kísérletek alapján azt mondhatjuk, hogy 10 keverékig határozottan javulnak a felismerési eredmények, de itt telítésbe megy a görbe, majd 20 keverék felett bizonytalanná válik a betanítás (vö. 4. ábra). A bizonytalanság oka egyrészt a kevés tanító adat, másrészt a tanítás módja. A keverékek tanításánál ugyanis egy kezdeti modellt optimalizál a rendszer, a kezdeti modell azonban véletlenszerűen kerül kiválasztásra, ami lokális minimumhoz vezethet.

5. Beszédhang helyett hosszabb szegmensek bevezetésének hatása

Eddigi kísérleteinkben a beszédhangok betanítását vizsgáltuk. 11-

13%-os hibaaránynál alacsonyabbat nem sikerült elérni 3-400 szavas felismerési szótár esetén. Megvizsgáltuk annak a lehetőségét, hogy a beszédhangnál hosszabb egységeket tekintsünk felismerési alapegységnek. Olyan elemet kell választani, amihez van elegendő tanító minta, tehát szavak, szótagok nem alkalmasak erre a feladatra, rövidebb elemekre van szükség, mint például a félszótagok.

A félszótag a szótag fele, azaz a szótagot a magánhangzónál kettévágva kapjuk a félszótagot, például *bolt = bo + olt.* Kétféleképpen lehet elvégezni a félszótagra bontást, az írott és az elhangzott szövegen. Az előbbi nem alkalmas számunkra, mert ha az *asszony* szót *a asz szo ony* félszótagokból szeretnénk összeállítani hibás eredményt kapnánk. Az írott szöveget tehát először fonotipikus alakra kell alakítani, és azon lehet a szótagolást, majd a félszótagokká alakítást elvégezni a hagyományos szótagolási szabály szerint. A szóösszetételekkel itt nem foglalkozunk, a beszédfelismerés szempontjából nincs jelentősége.

A félszótagok száma néhány ezer, nagyobb korpuszokat megvizsgálva közel 2000, a tanító adathalmazban közel 1000 különböző félszótag adódott. A félszótagok gyakorisága azonban nem egyenletes, vannak sokkal gyakoribb és igen ritka elemek is. Minél kevesebb tanító minta áll rendelkezésre egy félszótaghoz annál pontatlanabb az akusztikus modell becslése. Megvizsgáltuk, a legalább 20, illetve 80 tanító mintával rendelkező félszótagok arányát. A 4. táblázatban látható, hogy a félszótagok csupán 30%-ához rendelhető legalább 80 tanító minta, de ezek a félszótagok gyakoriságuk alapján 94,6%-os lefedést biztosítanak.

4. táblázat: A gyakori félszótagok aránya

Gyakori félszótagok	**Aránya**	**Kumulatív gyakorisága**
legalább 20 tanító minta	49%	98,8%
legalább 80 tanító minta	30%	94,6%

Azon félszótagok, melyekhez nincs elegendő tanító adat, tehát csak 20, illetve 80-nál kevesebb, azokhoz nem rendelünk önálló félszótag-

modelleket, hanem beszédhangok sorozataként állítjuk elő őket, például *oksz=o + k + sz*. Betanítottunk háromféle felismerőt, kiindulásként minden félszótaghoz önálló modellt rendeltünk, majd a legalább 20, illetve 80 tanító mintával rendelkező félszótagokhoz rendeltünk önálló modellt.

5. táblázat: Félszótag-modellek alkalmazásának hatása

Alkalmazott modell		Felismerési hibaarány	
		független adatokon	tanító adatokon
beszédhangmodell		8,7%	10%
félszótagmodell	mind önálló	20,4%	1,8%
	>20 tanító adat	7,7%	1,8%
	>80 tanító adat	4,1%	3,2%

Az 5. táblázatból leolvashatóak a kísérlet eredményei. Nem szokás a tanító adatokon tesztelni, de tanulságos, hogy a ritka félszótagokhoz is önálló modelleket rendelve a tanító adatokon nagyon jó, míg független adatokon nagyon rossz eredményt értünk el. Ennek oka, hogy a néhány mintából vett statisztika az eredeti mintához nyilván hasonlítani fog, de az átlagtól nagyon eltérhet.

A tanító adathalmaz szókészlete nem volt azonos a tesztadatokéval, a szótár több elemet tartalmazott, ezért gyengébb a tanító adatokon a beszédhang modell alapú teszt, mint független adatokon.

Azt egyértelműen mutatják a kísérletek, hogy megfelelően tanított félszótagok esetén jelentősen, kevesebb, mint felére csökkenthető a felismerési hibaarány. A beszédhangmodellek esetén a korlátot a koartikulációból származó módosulás adja, amit a félszótag modellek hatékonyan kezelnek, mert ebben az esetben nem önálló hangok, hanem hangkapcsolatok kerülnek modellezésre.

Összefoglalás

Megvizsgáltuk, hogy a beszédfelismerők működését alapvetően meghatározó akusztikus modellek betanítása során milyen lehetőségeket lehet, és kell figyelembe venni, valamint, hogy ezek milyen hatás-

sal vannak a betanítás jóságára, azaz a későbbi felismerési eredményekre.

Az egyes kísérleteket egymástól függetlenül végeztük, azonban szükség lehet kombinált tesztekre is, mivel a beszélők számának vizsgálata azt mutatta, hogy 50-100 beszélő felett nem javul a felismerési arány, viszont a keverékek számának növeléséhez növelni kellene a tanító adatok mennyiségét.

Megállapítottuk, hogy a beszédhangnál hosszabb elem választása volt az egyetlen megoldás, mellyel 10% alá sikerült csökkenteni a felismerési hibát. Ennek oka a koartikuláció hatásának kezelése, amelyre a beszédhangmodellek önmagukban nem alkalmasak.

Irodalom

Fegyó, T. – Mihajlik, P. – Tatai, P. – Gordos, G. (2001): Pronunciation Modeling in Hungarian Number Recognition. In: Proceedings of the Eurospeech 2001. Aalborg, Denmark, 330-333.

Mihajlik, P. – Révész, T. – Tatai, P. (2002): Phonetic Transcription in Automatic Speech Recognition. Acta Linguistica Hungarica. Megjelenés alatt.

Rabiner, L. R. – Juang, B.H. (1993): Fundamentals of Speech Recognition. PTR Prentice Hall. Englewood Cliffs.

Szarvas, M. – Fegyó, T. – Mihajlik, P. – Tatai, P. (2001): Automatic recognition of Hungarian: Theory and practice. In.: International Journal of Speech Technology 3/3-4., 277-287.

Vicsi, K. – Víg, A. (1997): Babel — a multi-lingual database. Technical report. http://www.ttt.bme.hu/speech/database.htm. „György Békésy" Acoustics Research Laboratory of the Budapest University of Technology and Economics. Budapest.

Vicsi, K. (1999): Speechdat — Hungarian speech database for creation of voice driven teleservices. Technical report. http://luna.ttt.bme.hu/speech/speechdt.htm. „György Békésy" Acoustics Research Laboratory of the Budapest University of Technology and Economics. Budapest.

Vicsi, K. – Valyon, Z. – Gordos, G. – Csirik, J. – Kocsor, A. – Tóth, L. (2002): MTBA — Magyar nyelvű telefonbeszéd adatbázis. Technical report. http://luna.ttt.bme.hu/speech/MTBAhun.htm. „Békésy György" Akusztikai Kutató Laboratórium. Budapesti Műszaki és Gazdaságtudományi Egyetem. Budapest.

Young, S. et al. (1999): The HTK Book. Microsoft Corporation.