A MAGYAR TELEFONBESZÉD-ADATBÁZIS (MTBA) KÉZI FELDOLGOZÁSÁNAK TAPASZTALATAI

Tóth László – Kocsor András

Bevezetés

A magyar telefonbeszéd-adatbázis (MTBA) olyan nagyméretű, telefonos beszédkorpusz, amely a magyar nyelvű fonetikai, beszédtechnológiai kutatások és fejlesztések támogatására készült. Az adatbázis 500 adatközlő hangfelvételeit tartalmazza, amelyek jelentős részét, beszélőnként 12 mondatot és 4 szót fonetikai szinten annotáltunk és szegmentáltunk. A feldolgozás során számtalan érdekes fonetikai és fonológiai jelenséggel találkoztunk, ami annak köszönhető, hogy a mondatok összeállításánál a hangkapcsolatokban való gazdagságra törekedtünk. Jelen tanulmányban a fonetikai szintű szegmentálás nehézségeiről, tapasztalatairól és érdekességeiről számolunk be.

A feldolgozott hanganyag tartalma és felépítése

Az MTBA magyar telefonbeszéd-adatbázis az Oktatási Minisztérium IKTA-3 pályázatának keretében készült a BME Távközlési és Telematikai Tanszékének és a Szegedi Tudományegyetem Számítástudományi Tanszékének együttműködésében 2000 szeptembere és 2002 augusztusa között. Az adatbázis 200 vezetékes és 300 mobilhívás felvételét tartalmazza. A telefonálók által bemondandó szövegeket úgy állítottuk össze, hogy az alkalmas legyen számítógépes beszédfelismerő-rendszerek betanítására. Ezért a szöveganyag tartalmazza például a legfontosabb magyar településneveket, a legjelentősebb intézmények neveit, dátumokat, családneveket, számneveket stb. Ebből fonetikai szintű feldolgozást adatközlőnként 12 mondaton és 4 szón végeztünk, így a továbbiakban csupán az adatbázisnak erről a feldolgozott részéről esik szó. Az adatbázis egyéb részeinek leírásáról vö. Vicsi (2002).

A fonetikailag feldolgozott mondatok és szavak azzal a céllal kerültek az adatbázisba, hogy azokon a beszédfelismerő-rendszerek fo-

netikai modelljeit lehessen betanítani. Ezért a mondatokat úgy kellett összeválogatni, hogy azokban a magyar beszédhangok, valamint bi- és trifónok (kettes és hármas hangkapcsolatok) kellően nagy számban forduljanak elő. A szöveganyag összeállítását végző budapesti partner e célból egy automatikus fonetikus átíróprogramot készített, amelynek segítségével ellenőrizte, hogy az adatbázisba bevett mondatok hangstatisztikája megfelel-e az említett elvárásnak. Összességében 1992 darab, újságcikkekből származó mondatot válogattunk össze az adatbázisba, és ezek mindegyikét összesen háromszor olvasták fel. A beszélőnként további 4 szóra azért volt szükség, hogy ily módon növelni tudjuk az olyan ritka hangkapcsolatok gyakoriságát, amelyek a mondatokban nem fordultak elő kellő számban.

A kézi feldolgozás feladata

Az adatközlők által felolvasott 12 mondat és 4 szó kézi feldolgozását egy speciálisan e célra készült program segítségével végeztük, amely képes a hangfájlok hullámformájának és spektrumának megjelenítésére, továbbá tetszőleges hangrészlet lejátszására is. E három információ alapján kellett a munkát végzőknek a hangmintát fonetikai szinten szegmentálni és címkézni. Címkézésre a nemzetközileg elfogadott SAMPA kódolást (vö. Sampa 2002) használtuk, ami az IPA számítógépes karakterekhez igazított változata. A magyar nyelvhez összeállított SAMPA kódtábla összesen 58 különböző fonetikai címkét tartalmaz. Ennél finomabb részletességű jelölés nem volt célunk, ugyanis a feldolgozás így is több mint egy évet vett igénybe.

A fonetikus átirat elkészítésének felgyorsítására a program mindig felkínál egy javaslatot. Ezt a korábban már említett átíróprogram készíti el a mondat ortografikus lejegyzéséből kiindulva. Az algoritmus képes kezelni az alapvető hasonulási szabályokat, viszont nyilvánvaló módon nem tudja megjósolni, ha a beszélő mást mondott, vagy máshogyan ejtett ki valamit. Ezért a gép által felajánlott átírást csak javaslatnak tekintettük, és szükség esetén módosítottuk.

A szegmentálási feladat ismertetésének befejezéseként szeretnénk hangsúlyozni, hogy az adatbázis legfontosabb tervezett alkalmazása a gépi beszédfelismerő-rendszerek tanítása, ezért a szegmentálás során végig az automatikus felismerők igényeit tartottuk szem előtt. A gépi

tanulás hatékonyságának növelése szempontjából a legfontosabb, hogy az azonosan címkézett szegmentumok minél jobban hasonlítsanak egymásra, a különböző címkéjűek pedig minél jobban eltérjenek. Ez a törekvésünk az oka annak, ha bizonyos esetekben valamelyest eltértünk a fonetika által javasolt címkézési vagy szegmentálási szabályoktól. Úgy véljük azonban, hogy ezek az eltérések nem számottevőek.

Általános tapasztalatok

A telefonhívások megszervezésénél gondosan ügyeltünk arra, hogy az ország minden területéről érkezzenek hívások, méghozzá életkor és nem szerint a magyar lakosság eloszlásának megfelelően. A felvételekhez az adatközlők megkapták a felolvasandó mondatok/szavak listáját és egy rövid tájékoztatást az adatbázisról. A felvételek feldolgozása során azt tapasztaltuk, hogy több beszélő is szebben artikulált, mint az természetes lett volna. Ezért az adatbázison végzett nyelvészeti vagy beszédtechnológiai vizsgálatok során tanácsos tekintetbe venni, hogy olvasott és nem spontán beszédről van szó[1]. Ennek legfeltűnőbb következménye az volt, hogy az automatikus átíró által feltételezett hasonulásokat gyakran kellett „visszaalakítanunk" az átiratban. Talán a rögzítésre kerülés tudatával magyarázható az is, hogy rendkívül kevés a tájnyelvi jellemzőket tartalmazó felvétel: csak néhány esetben találkoztunk például *ö*-ző vagy zárt *ë*-t használó beszélővel.

További általános tapasztalatunk volt, hogy nagyon kevés (a 8000-ből csupán 65) felvételnél találtunk olyan nagy mértékű háttérzajt, ami a szegmentálási munkát nehezítette. Valószínűleg ez sem felel meg a természetes telefonálási körülményeknek: a hívók többsége feltehetően keresett egy „csendes, nyugodt" helyet a hívás lebonyolítására. A feldolgozás során a háttérzajnál jóval több gondot okozott a sok esetben fellépő torzulás, illetve a telefonok recsegése. Az utóbbi kezelésére egy speciális szimbólumot vezettünk be, így a recsegés miatt használhatatlan beszédrészletek a fonetikai kódolás alapján kiszűrhetők,

[1] Az artikulációs gondosság két végletes példája lehet a 409. (túlartikuláló) és a 328. (alulartikuláló) adatközlő.

eltávolíthatók. Ugyanezt a jelet használtuk a különféle „beszélőkeltette zajok" – köhögés, krákogás, levegő kifújása – címkézésére is.

Mivel a felolvasandó szöveg összeállításánál a fonetikai tartalom volt a fő szempont, a mondatok többsége eredeti szövegkörnyezetéből kiragadva teljesen értelmetlenné vált, amihez még az is hozzáadódott, hogy többnyire viszonylag hosszú mondatokról volt szó (30-50 beszédhang). Ezért a beszélők gyakran tévesztettek a mondatok felolvasásakor, újra kezdtek egy-egy szót. A félbeszakadt szavakra szintén speciális jelölést használtunk, így a feldolgozás során ezeket is figyelmen kívül lehetett hagyni.

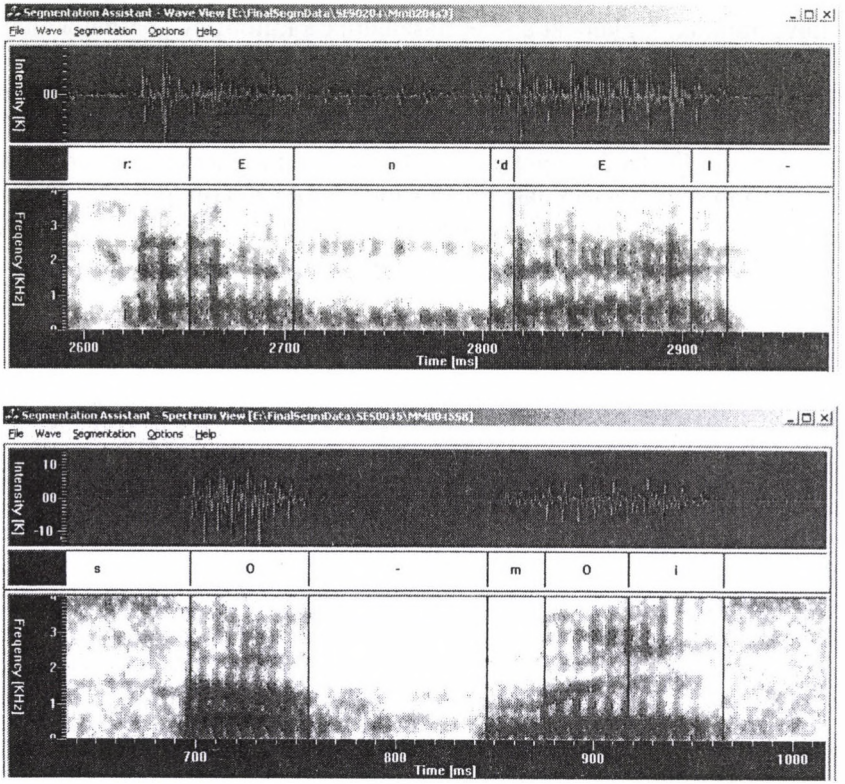
A szegmentálást nehezítő jelenségek

A szegmentálás során az egyes fonetikai szegmentumok kezdő- és végpontját egyetlen időpillanat hozzárendelésével kell megadni. Azonban az egyes fonetikai jegyek nem teljesen szinkronban kapcsolódnak ki és be, ami a hangok szintjén koartikuláció formájában jelenik meg. Ilyenkor a beszédhangok határát – jobb híján – általában az átmeneti rész közepére húztuk be. Ezzel a módszerrel a koartikulációt tudtuk ugyan kezelni, de néhány esetben feloldhatatlan nehézségekkel szembesültünk. Ezeket az eseteket, és az általunk alkalmazott áthidaló megoldásokat ismertetjük az alábbiakban.

A zárhangok és affrikáták zár/zörej felbontása

A zárhangokat és az affrikátákat a fonetika három képzési szakaszra bontja: a záralkotás, a zár és a felpattanás szakaszára (Roach 1991). Ezek a részek akusztikailag egymástól élesen elválnak és viszonylag önálló életet élnek, például a zár eltűnhet, ha a felpattanó zárhang nazális hang után áll – a nazálisból ugyanis közvetlenül fel lehet pattintani a zörejrészt (vö. 1/a ábra). Ugyanígy nincs értelme zárrészt bejelölni egy szövegkezdő zöngétlen zárhang esetén (a zár és a szó előtti csend nem különböztethető meg).

Más esetekben viszont csak a zár marad meg – a zörejrészt nem ejtjük ki. Tipikus példa a zárhang–nazális kapcsolat (1/b ábra). Ezért az eredeti SAMPA jelöléseket módosítottuk, ahogy az az ábrák jelölésmódján is megfigyelhető.



1. ábra
a) Nazális utáni zárhang zár nélkül; b) nazális előtti zárhang zörej nélkül: az [m] előtti [k] hang zöreje hiányzik

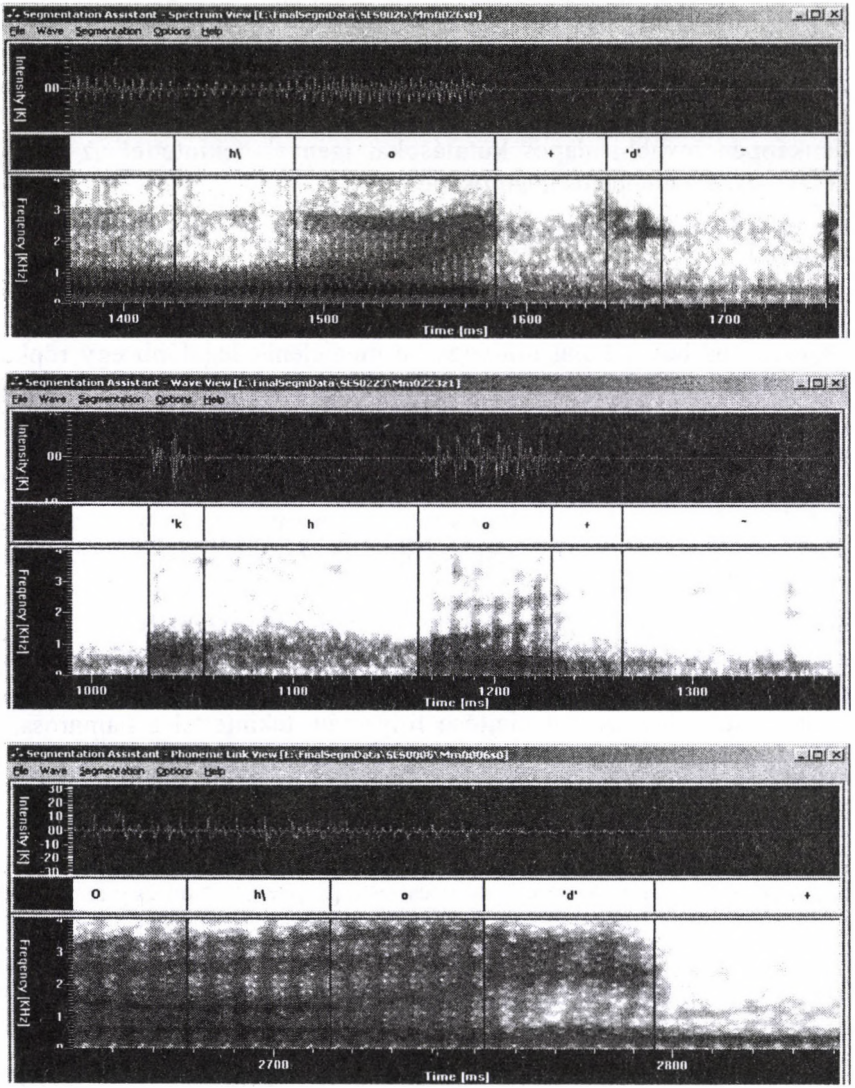
A zárhangokat és az affrikátákat felbontottuk két részre: a +, illetve – jeleket bevezettük a zöngés/zöngétlen zár jeleként, a hangokra vonatkozó jeleket pedig csak a zörejrészek címkézésére használtuk (egy kezdő aposztróffal kiegészítve, az eredeti jelöléstől való eltérést hangsúlyozandó). A záralkotás szakaszát nem illettük külön jellel, mivel általában nem önálló szegmentumként jelentkezik, hanem a megelőző hang képzését (formánsszerkezetét) módosítja. Ez azonban bizonyos hangkiesések esetén olyan szegmentálási eredményhez vezetett, ami

fonetikai szempontból semmiképpen nem kielégítő. Úgy gondoljuk, hogy a jelölésrendszert a hangkapcsolatokra és speciálisan a felpattanók kapcsolataira nézve további hasonló adatbázisok készítése esetén felül kell majd vizsgálni. A hangátmenetek akusztikai szerkezete mindenképpen további alapos kutatásokat igényel, tekintettel az átmenetek beszédpercepciós jelentőségére.

A hangkiesések problémája

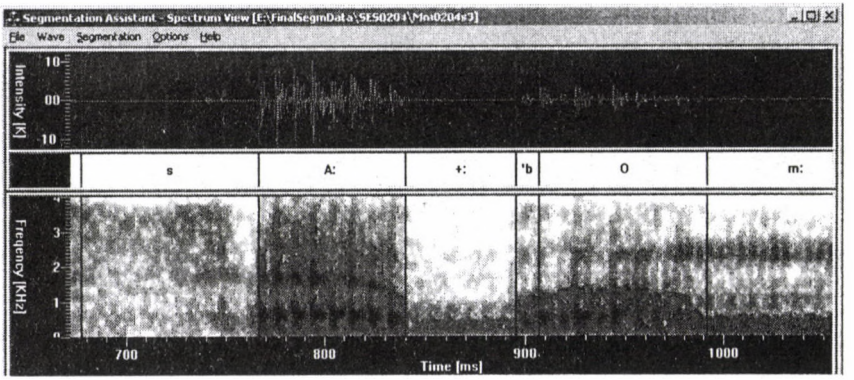
Az „átmenet-megfelezős" stratégia áthidalja ugyan a koartikuláció okozta határbehúzási nehézséget, de továbbra is feltételezi, hogy a két szomszédos beszédhang mindegyike megjelenik legalább egy röpke időintervallum erejéig. Az igazi nehézség akkor jelentkezik, amikor egy hang teljesen eltűnik, pusztán néhány jegye marad meg, amelyek a szomszédos hangok képzését módosítják. Ilyenkor a fonetikai szegmentumokra bontás egyszerűen lehetetlen, de legalábbis nem ad meggyőző eredményt. A vizsgált anyagban ilyesmivel ritkán találkoztunk, tekintve, hogy olvasott és nem spontán beszédről volt szó. Azonban néhány esetben kényszermegoldásokat kellett találnunk. Az egyik klasszikus példa a *hogy* szó. Pontos artikuláció esetén az [o], a zár [+] és a zörej ['d'] elválik, és jól bejelölhető a spektrumon (2/a ábra). Azonban szép kiejtés esetén is megfigyelhető a második formáns gyors emelkedése az [o] kiejtése folyamán, tekintettel a hamarosan következő palatális zárra.

Ez a formánsemelkedés olyan markáns, (és a *hogy* szó általában annyira megjósolható), hogy folyamatos beszédben a zörej realizálására nincs szükség kommunikációs szempontból (2/b ábra). Egy másik ejtéstípusnál még a zár sem tökéletes, és a [d']-re pusztán egy felszaladó formáns utal az [o]-ban. Ilyenkor a második formáns többnyire eléri a [d']-nek megfelelő locusokat (2/c ábra), ezért ilyenkor ahhoz az (az ábrán is látható) jelölésmódhoz folyamodtunk, hogy kitöröltük a zárat, de a ['d'] jelű szegmentumot meghagytuk. Sokszor azonban az egész szó olyan rövid, hogy azt éreztük volna a leghelyesebbnek, ha mind a [+], mind a ['d'], azaz a teljes [d'] kimarad. Habár lehallgatáskor ezekben az esetekben is egyértelműen *hogy*-ot hallunk, valójában nincs olyan időintervallum, amelyet [d']-nek lehetne jelölni.



2. ábra

A *hogy* háromféle ejtésben/jelölésben a) a [d'] zár- és zörejrésze is megjelenik, b) a zörej elmaradása, c) a zár elmaradása



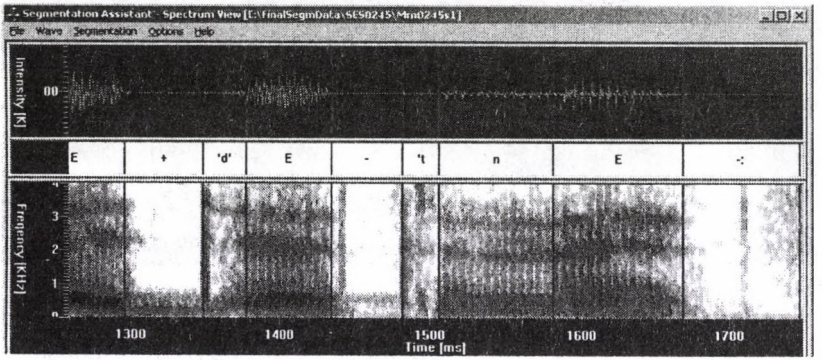
3. ábra.
A [g] és [b] találkozása (az *országba* szóban) – az [A:] formánsai a [g]-hez igazodnak, de a [g] zöreje hiányzik, így hosszú zárat kellett jelölnünk

A másik szituáció, ami hasonló nehézségeket okozott, a zárhang–zárhang kapcsolatok esete volt. Ezek kiejtésekor a legritkább esetben található meg mindkét hang zörejrésze. Ehelyett általában csak a második zörej marad meg, és a két zár egyetlen hosszú zárrá egyesül (vö. 3. ábra). Ennek ellenére lehallgatáskor tisztán érzékeljük mindkét zárhangot – ugyanis a megelőző magánhangzó formánsai az első felpattanó zörejének locusait veszik célba. Azonban hiába tudjuk mindezt, ha tartani akarjuk magunkat a szegmentum szintű jelöléshez, akkor az első zörejt egyszerűen nincs hova elhelyezni, így marad az ábrán is látható jelölésmód: hosszú második zárhang feltüntetése két rövid helyett. Ez az a szituáció, amikor a zárképzés önálló szegmentumként való kezelése segített volna, ugyanis akkor legalább ez a szegmentum megmaradhatott volna az első zárhangból.

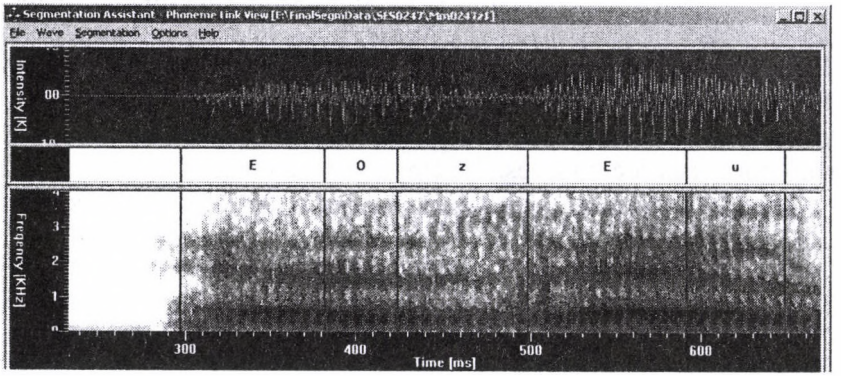
A spontán beszédre jellemző jelenségek

Azt gondoljuk, hogy spontán beszéd és a megértést nagyban segítő szövegkörnyezet esetén a nagyfokú, akár több beszédhangra is kiterjedő hangkiesés, elmosódás sokkal gyakoribb. Ez a jelenség azokat a

mondatépítő szavakat, kötőszavakat érinti leginkább, amelyek a leginkább megjósolhatóak a szövegkörnyezet alapján (vö. a *hogy* példáját).



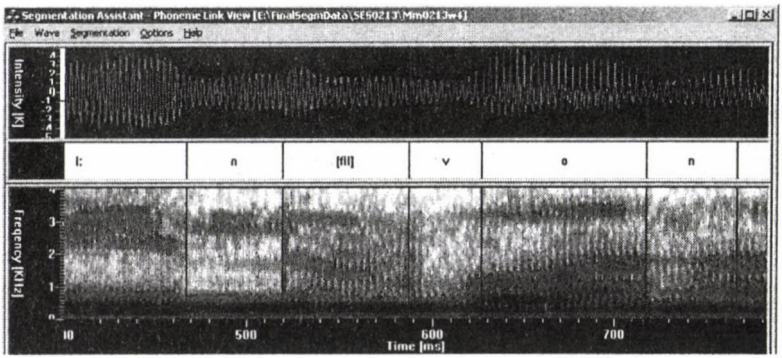
4. ábra.
Az *egyetlenegy* szóban a *len* összeolvadt egy leginkább [n]-ként címkézhető hanggá



5. ábra.
A mondatkezdő *ez az* összemosódása az „Ez az európai…" kezdetű példamondatban

E szavak lehallgatása esetén az a jellemző, hogy a mondatban tökéletesen érthetőek, kiragadva őket viszont nem, vagy alig azonosítható-

ak, és sem a spektrumon, sem a hullámformán nem lehet megtalálni a belőlük „elnyelt" hangokat. Két példát láthatunk a 4. és az 5. ábrán. Előbbi esetében az *egyetlenegy* szó *len* szótagja összeolvadt egyetlen szegmentummá, amelyet *n*-ként címkéztünk. Még erőteljesebb összemosódás történik az 5. ábrán, ahol a mondatkezdő *ez az* valami olyan egységgé olvad össze, ahol az eredeti hangok egyikét sem lehet pontosan bejelölni. Mivel ilyenkor is kellett valamilyen szegmentumokat bejelölnünk, nyilvánvalóan kényszermegoldásokhoz kellett folyamodnunk.



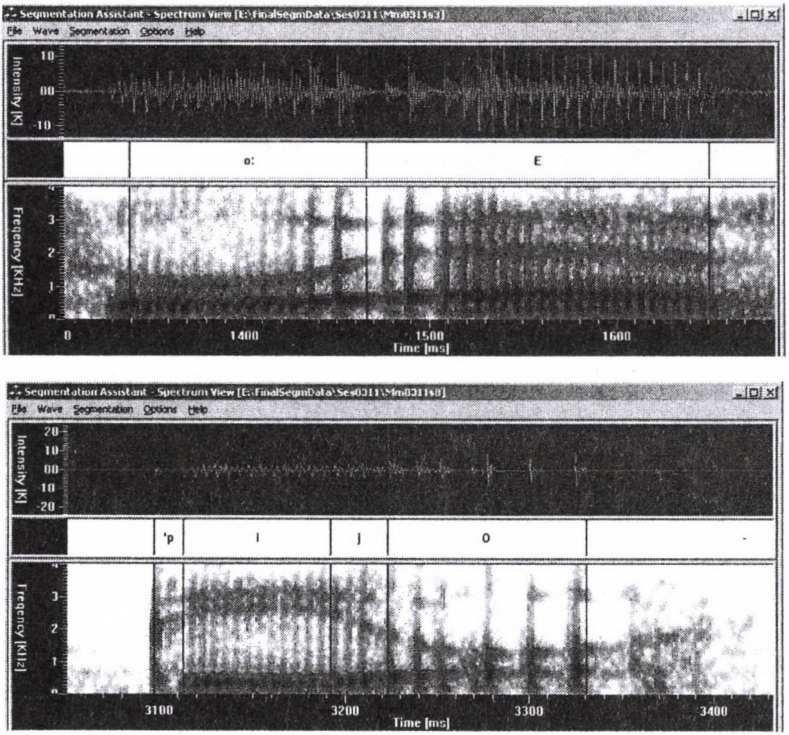
6. ábra.
Az [n] és [v] találkozása a *színvonalas* szóban. Az [n] zárfelpattanását külön szegmentumként jelöltük.

Mint azt már említettük, külön jel szolgált az úgynevezett „beszélőkeltette zajok" címkézésére. A fonetikai kódkészlet emellett tartalmaz egy speciális jelet a kitöltött szünet (hezitálás) megjelölésére is. Kitöltött szünetek csak ritkán fordultak elő az anyagban, azonban ezzel a jelöléssel – [fil] – címkéztünk fel olyan szegmentumokat, amelyek a beszéd részét képezték, de nem voltak pontosan meghatározhatók. Erre példát mutat be a 6. ábra, ahol egy [nv] hangkapcsolat szerepel a szóban. Ennek elvileg [Fv]-vé kellett volna alakulnia, ahol az [F] az [n] foghang változata. Az adatközlő azonban túl szépen akart beszélni, mindkét hangot kiejtette, ezért viszont kénytelen volt a két

hang közé beszúrni egy szegmentumot. Ez valójában az [n] zárfelpattanása, de semmiképpen sem akartuk az [n] részeként jelölni, mivel nagyon ritka jelenség, és így a statisztikai úton működő beszédfelismerőket megzavarhatná. Ezért [fil] címkével láttuk el.

További érdekességek

A pontos hanghatárok megtalálását rendkívül meg tudja nehezíteni az ún. laringalizált vagy „csikorgó" beszéd – „creaky voice", „vocal fry" (Durand–Siptár 1997). Ilyenkor a hangszalagok szabályos periodikus nyitódása–záródása helyett egyes periódusok kimaradoznak, és a spektrogram „szakadozottá" válik (7/a ábra).



7. ábra

a) Csikorgó beszéd; b) a szóhatáron fellépő hasonló jelenség

Ez az emberi beszédpercepciót egyáltalán nem nehezíti, azonban a számítógépes beszédfelismerő algoritmusok működését erősen zavarhatja. Úgy tapasztaltuk, hogy a fenti jelenség természetes módon jelentkezik a szóhatároknál: a szóhatáron a beszélő lecsökkenti a gégefőben a nyomást, de nem teljesen, így tökéletes csend helyett a hangszalagok továbbra is kinyílnak, de az elégtelen nyomás miatt nem szabályos periódusokban, hanem csak néha-néha. Szünet helyett a két szó között jellegzetes szaggatott jelszakasz keletkezik (7/b ábra). Ezt a jelenséget igen gyakran tapasztaltuk és nem csak mondathatároknál. Érdekes kérdésként vetődik fel, hogy vajon beszédérzékelésünk ki tudja-e használni a szóhatárok megtalálásánál.

További érdekességként találtunk egy olyan hangváltozatot, amelyet a SAMPA kódtábla nem tartalmaz, pedig erősen eltér a hang szokásos kiejtésétől, és viszonylag sokszor fordult elő. Ez a változat a zöngétlen [l] volt, amely jellemzően a *-tlan/-tlen* végződés esetében jelentkezett. Megjegyezzük, hogy a SAMPA táblázatban külön feltüntetett zöngétlen [j] viszont jóval kevesebbszer fordult elő, mint ahogy az automatikus átíró jósolta: az általa javasolt esetek többségében át kellett javítanunk a címkét zöngés [j]-re.

Végül megemlítjük, hogy az adatbázisban mind az automata által javasolt, mind a valódi kiejtésnek megfelelő fonetikus átirat megtalálható, így a kettő közötti eltérés statisztikai módszerekkel könnyen számszerűsíthető. Ezért reményeink szerint az adatbázist nem csak a beszédtechnológiával foglalkozó mérnökök, hanem a nyelvészek is nagy haszonnal tudják majd felhasználni.

Irodalom

Durand Jacques – Siptár Péter (1997): Bevezetés a fonológiába. Osiris kiadó. Budapest.

Gósy Mária (1995): Szükséges és szükségtelen hangátmenetek. In: Beszédkutatás '95. Szerk.: Gósy Mária. MTA Nyelvtudományi Intézet. Budapest, 20-31.

Roach, P. (1991): English Phonetics and Phonology: a Practical Course. Cambridge University Press. Cambridge.

Sampa (2002): http://www.phon.ucl.ac.uk/home/sampa/home.htm.

Vicsi Klára – Tóth László – Kocsor András – Gordos Géza – Csirik János (2002): MTBA – Magyar nyelvű telefonbeszéd-adatbázis. Híradástechnika 8: 35-39.

A szerzők köszönetet mondanak Sejtes Györgyinek nyelvészeti tanácsaiért.