MTBA – Magyar nyelvű telefonbeszéd adatbázis

Vicsi Klára,* Tóth László,+ Kocsor András,+ Gordos Géza* és Csirik János+

* Budapesti Műszaki Egyetem Távközlési és Telematikai Tanszék
+Szegedi Tudományegyetem Számítástudományi Tanszék

Összefoglalás

Magyar nyelvű, telefonon keresztül rögzített beszéd-adatbázist hoz létre a Budapesti Műszaki Egyetem Távközlési és Telematikai Tanszék (Beszédakusztikai Laboratóriuma) a Szegedi Tudományegyetem Számítástudományi Tanszékével együttműködve. A magyar telefonbeszéd-adatbázis (MTBA) a statisztikai feldolgozási módszereken alapuló, telefonon keresztül működő beszédfelismerő rendszerek betanítására és tesztelésére ad lehetőséget. Ilyen lehetséges alkalmazások az izolált szavas rendszerek, szókereső és azonosító rendszerek, dialógusrendszerek, valamint az ún. szótárfüggetlen felismerők, amelyeknél a felismerés a szónál kisebb felismerési egységek modellezésén alapul.

Az adatbázis szabályrendszerét európai szakértői bizottsági ajánlások alapján [1, 2] szerkesztettük meg. Az EU adatbázis-specifikációban nem szerepel a beszéd fonéma szintű szegmentációja és címkézése. Mivel a specifikáció összeállítása óta a beszédkutatás folyamatosan fejlődik, egy most létrehozandó adatbázisnál fontos az adatbázis egy részének fonémaszintű szegmentálása és címkézése, hiszen ez teszi lehetővé a szótárfüggetlen rendszerek betanítását, és így ilyen típusú felismerők létrehozását. A készülő új adatbázis fonémaszintű szegmentálást és címkézést is tartalmaz.

Az adatbázis hanganyaga 500 beszélő által telefonon bemondott szövegből (300 vezetékes, 200 mobil hívás) áll. Az összeállított szöveganyag a sokfeladatos elvárásoknak megfelelően igen sokrétű, változatos. Tartalmazza például a magyar településneveket, a Magyarországon működő legjelentősebb intézmények neveit, valuták neveit, dátumokat, család és keresztneveket, speciálisan a magyar nyelv sajátosságait tükröző mondatokat stb.

1 Bevezetés

A statisztikai beszédfeldolgozás elterjedésével általánosságban, de különösen a gépi beszédfelismerés területén, határozott és sürgető igény lépett fel a jól megszerkesztett, nagyméretű beszédadatbázisok létrehozására. A felismerők betanítása ugyanis nem egyéb, mint egy statisztikai alapú paraméterbecslés, a pontos becsléshez pedig nagyszámú minta alapján történő betanítás szükséges. E minták gyűjteményei – a szükséges jegyzetekkel, címkézésekkel és átírásokkal ellátva – képezik az adatbázist. Az adatbázisoknak tartalmazni kell azokat a megfigyeléseket, amelyek a paraméterbecsléshez szükségesek, tehát mindazokat a mintákat, amelyek egységesen lefedik a beszéd (és a környezeti zajok) változatosságát.

A beszédadatbázis nagyméretű hang-adathalmaz, amelyet igen sokféle csoportosítás szerint hozhatunk létre. Általában a felhasználási terület határozza meg az adatbázis nagyságát és belső szerkezetét. A telefonon keresztül történő felismerés megoldására olyan betanító és tesztelő anyagra van szükség, amely telefonon keresztül rögzített hanganyagot tartalmaz. Továbbá minden olyan tipikus esetváltozatnak, alapvariációnak elő kell fordulnia a tananyagban, ami a felismerés során előfordulhat, mert elegendően pontos felismerési biztonság csak így érhető el.

A személyfüggetlen, telefonon keresztül működő beszédfelismerők betanítására és tesztelésére alkalmas adatbázisok létrehozásának szabályrendszerét az Európai Közösség által létrehozott szakértői bizottság állította össze (MLAP LRE-63343 SPEECHDAT (M) [1, 2]). A nemzetközi szakértői bizottság által összeállított szabályrendszer biztosítja azt, hogy az előírások alapján létrehozott adatbázisok hasonlóak és egységes alapot képviselve hasonló beszédtechnológiai fejlesztési lehetőséget nyújtanak a feldolgozott nyelvekhez. Az MTBA adatbázist ezen szabályrendszerek betartásával hoztuk létre, kibővítve azt a fonémaszintű automatikus szegmentálással és címkézéssel.

Az adatbázis létrehozása három munkaszakaszban történt, ezeket ismertetjük az alábbiakban:

Az első munkaszakaszban a készítendő beszédadatbázis tartalmát, szabványait definiáltuk. Az adatbázis parancsszavakat, különállóan kiejtett szavakat, mondatokat és spontán beszédet is tartalmaz, valamint fonetikailag kiegyensúlyozott mondatokat és szavakat, amelyben a magyar fonémák és a leggyakoribb fonémakapcsolatok elegendő számban fordulnak elő, de a szöveg még kezelhető méretű.

A második szakaszban történt az adatbázis felvétele és annotálása. Ennek főbb részfeladatai a következők voltak:

- felvevő-berendezés (telefonszerver) létrehozása
- szövegbemondó lapok megszerkesztése
- az adatbázisgyűjtő dialógus elkészítése
- beszélők toborzása
- beszédfájlok készítése, a folyamat nyomon követése
- annotációk készítése a beszédfájlokhoz, a fonetikailag gazdag mondatok és szavak fonotipikus átírása
- automata szegmentáló szoftver elkészítése
- az adatbázis tartalmának dokumentálása

A harmadik szakaszban került sor a fonémaszintű audio-vizuális szegmentálásra és címkézésre, a teljes adatbázis ellenőrzésére és a CD-n való rögzítésre.

2 A telefonbeszéd adatbázis szövegének összeállítása

Az adatbázis számos különböző típusú beszédfelismerő betanítására és tesztelésére kell, hogy lehetőséget adjon. Ezek a felismerők az izolált szavas rendszerek, szókereső és azonosító rendszerek, dialógus rendszerek, valamint a folyamatos beszédfelismerő rendszerek. Az összeállított szöveganyag a sokfeladatos elvárásoknak megfelelően igen sokrétű. Tartalmaz tipikus szavakat, szókapcsolatokat, stb. A pontos összeállítást az 1. táblázat mutatja.

1. táblázat: Az MTBA adatbázis szövegeinek tartalmi felépítése.

Típus	**Darabszám**
parancsszavak	29 db
számjegysorozatok	10 db
Telefonszám	500 db
hitelkártyaszám	150 db
PIN kód	200 db
spontán dátum	500 db
relatív dátum	500 db
parancsszavas kifejezés	145 db
számjegy	166 db
betűzött spontán vezetéknév	166 db
betűzött városnév	500 db
betűzött szó	100 db
pénzmennyiség (forint/euró)	250 db
természetes szám	200 db
spontán vezetéknév	166 db
spontán városnevet, cégnév	166 db
vezeték + keresztnév	500 db
fonetikailag gazdag mondat	1992 db
Időpont	166 db
fonetikailag gazdag szavak	2000 db.

A fonetikailag gazdag mondatok szerepe az anyagban az, hogy a fonémaalapú felismerők részére elegendő mennyiségű betanító és tesztelő adatot szolgáltasson.

2.1 Egy beszélő szöveganyaga

A teljes szöveganyagot az egyes beszélők szöveganyagára kellett lebontani, tehát meg kellett szerkeszteni, hogy egy beszélő pontosan milyen szöveganyagot mondjon be. Ezeket a megszerkesztett szövegcsoportokat beszéd-szöveglapokon rögzítettük. A lapok két részből álltak:

- Az első rész olyan általános kérdéseket tartalmazott (pl. név, születési idő és hely, stb.), amelyek minden beszélőnél azonosak.
- A második rész különböző szempontok alapján összeválogatott szöveget tartalmazott (ld. a 2. táblázatot), ez beszélőnként különböző volt.

Az adatbázisban 500 beszélő szerepel, tehát 500 ilyen szöveglapot készítettünk el és küldtünk szét az országban. A szöveglapok alapján érkező bemondásokat a felvételi rendszer rögzítette, és típusok szerint csoportosította. Egy hívás ideje a teljes dialógusra átlagosan 12 perc volt, amelyből a hasznos felvett szöveg ideje hívásonként kb. 6 - 7 perc.

2. táblázat: A bemondólapok részletes szövegtartalma.

Szövegtest-azonosító	Szövegelem-azonosító	Szövegtest tartalma	Szövegcsoport
A	1-6	6 db applikációs kifejezés	applikációs szavak
B	1	1 db tíz önálló számjegyből álló sorozat	
C	1	1 db szöveglap sorszáma	4 hosszabb szám
C	2	1 db telefonszám (9-11 jegy)	
C	3	1 db hitelkártya-szám (14-16 jegy)	
C	4	1 db PIN kód (6 jegy)	
D	1	1 db spontán dátum (pl. születésnap)	3 dátumkifejezés
D	2	1 db kijelölt dátum (szavakkal)	
D	3	1 db dátumkifejezés	
E	1	1 db "word spotting" mintamondatok egy applikációs szóval	
I	1	1 db önálló számjegy	
L	1	1 db spontán betűzött keresztnév (saját)	3 betűzött szó (betűk sorozata)
L	2	1 db betűzött városnév	
L	3	1 db betűzött szó (a fonémák lefedése érdekében)	
M	1	1 db pénzmennyiség (Forint)	
M	2	1 db pénzmennyiség (Euró)	
N	1	1 db természetes szám	
O	1	1 db spontán keresztnév (saját)	6 listakezelését segítő szó
O	2	1 db spontán városnév (születési hely)	
O	3	1 db gyakori városnév	
O	5	1 db gyakori cégnév	
O	7	1 db vezeték- és keresztnév	
O	8	1 db vezetéknév	
Q	1	1 db döntően "igen" válaszú kérdés	2 kérdés, várhatóan igen/nem válasszal
Q	2	1 db döntően "nem" válaszú kérdés	
S	0-9	10 db fonetikailag gazdag mondat	
Z	0-1	2 db fonetikailag gazdag mondat	
T	1	1 db spontán időpont (óra és perc)	2 időkifejezés
T	2	1 db időkifejezés (szavakkal)	
W	1-4	4 db fonetikailag gazdag szó	

2.2. A fonetikailag gazdag mondatok összeállítása

Ma a legsikeresebb felismerők alapegysége a szónál kisebb nyelvi egység, rendszerint beszédhang, hangkapcsolat, vagy szótag. Szónál kisebb alapegység esetén a szótárkészlet nem kötött, és általában a felismerés is biztonságosabb. Ezen felismerők részére igen fontos a jól összeállított betanító és tesztelő anyag, ahol a megfelelő felismerési elemek elegendő számban szerepelnek. Ezért fektettünk nagy súlyt a fonetikailag gazdag mondatok előállítására, hogy ezek révén a magyar nyelvre is legyen elegendő, a fonéma-alapú felismerők tanítási kritériumainak megfelelő hanganyag.

Tehát a feladat egy olyan optimális méretű szöveganyag megszerkesztése volt, ahol a magyarban előforduló fonémák, valamint a leggyakoribb bi- és trifonok – kettős, hármashangzatok – megfelelő számban fordulnak elő [3].

A mondatok összeállításához használt szöveg újságcikkeken alapult, kb. 1.6 MB méretű volt és kb. 14000 mondatot tartalmazott. Először a szöveget megtisztítottuk az extra karakterektől, jelentés nélküli szavaktól, oldalszámoktól stb. Utána a szöveget fonémák sorozatára alakítottuk egy speciális algoritmus segítségével. A szöveg-fonéma átírást többféleképpen is el lehet végezni, ebben az esetben a fonotipikus fonetikai átírást volt célszerű alkalmazni, vagyis a karakterek átírását a nyelv fonetikai szabályainak alapján kellett elvégezni úgy, hogy a szövegkörnyezetet is figyelembe vesszük (pl. koartikuláció, hasonulás) [4].

A fonémaátírás után statisztikai vizsgálatokat végeztünk az anyagon. Megvizsgáltuk a fonéma, bifon és trifon eloszlást az eredeti 1,6 MB méretű szövegadatbázison. Ebből az eredeti szövegből válogattuk ki a bemondandó anyagot úgy, hogy a kapott anyagban minden fonémából, bifon és trifon hangkapcsolatból elegendő számú minta legyen.

Egy bemondó 12 különböző mondatot olvasott fel. A teljes szöveganyag 166 bemondásra készült el azért, hogy a szöveganyag 3-szor kerülhessen ismétlésre az 500 bemondás során. Így a teljes szöveganyag 12 x 166, azaz összesen 1992 db különböző mondatból áll. Az 1992 mondat 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 teljes szöveg és a válogatott szöveg bifon statisztikáinak alapján a leggyakoribb 98.8% bifon mindegyike szerepel mindkét anyagban, és csak itt kezdenek el hiányozni a válogatott szövegből a bifonok. Az előzőhöz hasonlóan a trifon statisztikák alapján a teljes szöveg trifonjainak leggyakoribb 72.2%-ának mindegyike szerepel a válogatott anyagban, a hiányzások csak itt kezdődnek el, de az ennél ritkább trifonok közül még így is jónéhány megtalálható. A bifon és trifon statisztikák nagyarányú egyezése alapján a kiválasztott szöveget jónak minősíthetjük.

A ritkább fonémák előfordulásának szaporítására készült egy szó-blokk is, amely fonetikailag gazdag szavakat tartalmaz.

3 A felvételek készítése és a beszédfájlok formátuma

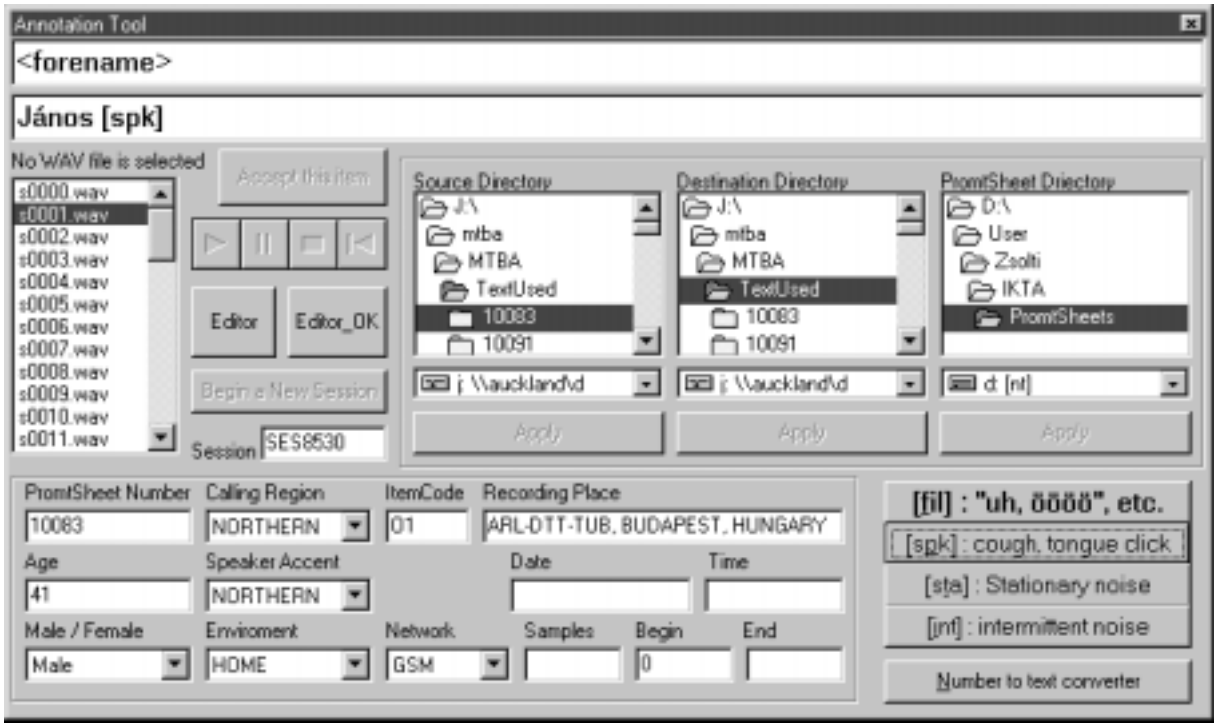
A szükséges technikai háttér létrehozása után ISDN vonalon keresztül vettük a hívásokat, egy zöld telefonszám biztosításával. Megszerkesztettük az adatbázisgyűjtő dialógust, amely a bemondásokat könnyebbé, természetesebbé tette. Minden egyes hívás számára önálló alkönyvtárat hoztunk létre, az egyes bemondásokat pedig külön WAV fájlokban rögzítettük. A rögzített hanganyagot 8 bites, 8 kHz-es *A-law* formátumban tároltuk. Minden egyes beszédfájlhoz egy ASCII formátumú *SAM* címkefájl is tartozik.

4 A hanganyag annotálása, címkézése, szegmentálása

A beszédadatbázisoknak a beszéd digitális tárolása mellett annak nyelvi információtartalmát is meg kell adniuk. Ezért a hullámforma tárolása mellett a hozzátartozó ortografikus karaktereket is rögzíteni kell. Különböző zajok, embertől származóak (ilyen a köhögés, nyelés, különböző szájmozgásból adódó zajok), vagy környezetiek (járművek, motorok zaja, székcsikorgás stb.) szintén bejelölésre kell, hogy kerüljenek a legtöbb adatbázisban, vagy a szöveganyagban, vagy magában az időfüggvényben.
A teljes hanganyagot annotáltuk, felcímkéztük. Továbbá a fonetikailag kiegyensúlyozott mondatokat és szavakat fonetikai szinten is szegmentáltuk [5], annak érdekében, hogy fonémaalapú felismerőt lehessen konstruálni a felhasználásukkal.

4.1 Az annotáció

Az annotálás annyit jelent, hogy minden hangfájl mellé egy címkefájlt készítünk, amely különféle információkat tartalmaz a hangfájl paramétereivel és tartalmával kapcsolatban. Jelen esetben az annotálás során megadtuk a hívás régióját, a beszélő életkorát és a többi szükséges információt. A felvétel dátumát és idejét, továbbá a hangminták számát a megfelelő WAV fájlokból nyertük ki. Mindezeket az adatokat a címkefájlokban rögzítettük. Bejelöltük továbbá a beszélő által elkövetett hibákat, illetve a felvételben hallható esetleges zajokat. Az annotáló program kezelő felülete az 1. ábrán látható.



1 ábra: Az annotálóprogram kezelői felülete.

4.2 Audio-vizuális fonetikai átírás

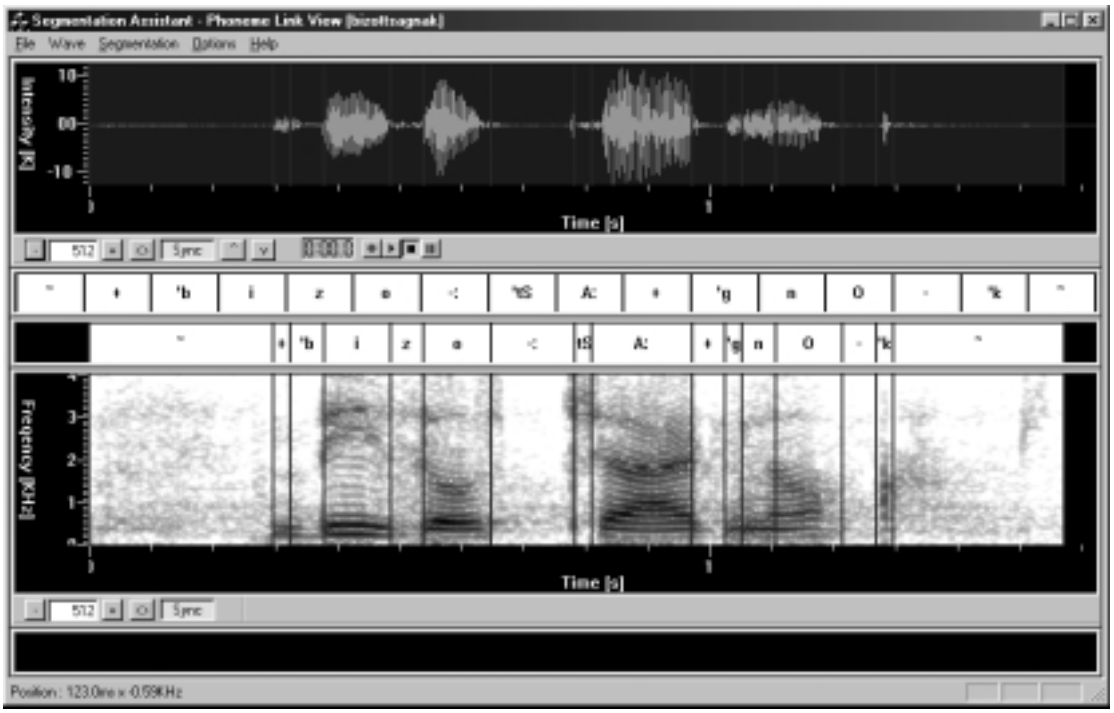
A kézi szegmentálásra előkészített anyagot, vagyis a fonetikailag gazdag szövegeket fonetikailag átírtuk, vagyis betűk helyett fonetikai szimbólumokkal jegyeztük le. A betűk átírását beszédhangokká a magyar nyelv fonetikai szabályai alapján végeztük, a szövegkörnyezet függvényében (pl. hasonulási szabályok figyelembe vételével). Ezt az átírást nevezik fonotipikus átírásnak.

4.2.1 A SAMPA jelölésrendszer

Általában a beszédhangok lejegyzésére fonetikai leírásoknál az IPA szimbólumrendszert szokás használni. Egy adott IPA szimbólum nyelvtől függetlenül egy meghatározott hangot szimbolizál. Sajnos ez a jelölésrendszer nem illeszkedik a számítógép billentyűzetéhez, ezért nemzetközi szinten bevezetésre került egy újfajta, úgynevezett SAMPA jelölésrendszer [6,7], amely alkalmazkodik a számítógéppel kezelhető karakterkészlethez. Így a számítógépes gépelés és továbbítás egyszerűen megoldható, ellentétben a hagyományos IPA jelölésrendszerrel.

4.3. Fonetikai szintű szegmentálás és címkézés

A fonetikai szintű szegmentálás és címkézés célja kézileg, vagy esetleg egy automatikus szegmentáló rutin támogatásával a beszéd időfüggvényében a fizikailag megfigyelhető fonémáknak és azok határainak a bejelölése. Ezt az ún. „audio-vizuális fonetikai átírást" az 1997-ban elkészült BABEL nemzetközi project ajánlása alapján [8] végeztük. Az átírásban a szöveg lehallgatása, továbbá az időfüggvény és/vagy a színkép elemzése nyújt segítséget. Ezen elemzések elvégzéséhez készítettünk egy speciális célprogramot, amelynek felületét a 2. ábra mutatja:



2. ábra: A szegmentálóprogram kezelői felülete

A program felső panelja a beszédjel hullámképét mutatja, az alsó panel pedig a színképelemzés révén előálló ún. spektrogramot. A lehallgatás mellett ez a két vizuális információ segíti a szegmentálás elvégzését. Középen látható a fonetikus szimbólumok sorozata, amelyet hozzá kell rendelni az adott hangfelvételhez. Ez a fonetikus átirat természetesen javítható is, ha a beszélő esetleg mást mondott, mint amit a szoftver feltételezett. A fonetikus jelek és a hangjel egyes részleteinek összerendelése a határvonalak segítségével történik, amelyek elhelyezése után a jelek automatikusan a megfelelő hangrészlet fölé ugranak, így segítve a tájékozódást.

A hanganyag szegmentálását nyolc, fonetikailag kellően felkészített szakember végezte. A munka során szakaszosan ellenőrizték egymást, a nagyobb pontosság érdekében. Az adatbázis teljes elkészülte után pedig még egy ellenőrzést végeztünk.

A tapasztalatok szerint a kézi szegmentálás optimális menete a következő: először lehallgatjuk a felvételt, figyelve az esetleges zajokra, hibákra, illetve a fonetikus átirat lehetséges hibáira. Ezután egy kisebb felbontás mellett (mint amilyen a 2. ábrán látható) érdemes végezni egy durva szegmentálást. Végezetül egy jóval erősebb nagyítás mellett véglegesítjük, pontosítjuk a fonémák határait, illetve kijavítjuk az esetleges ejtési eltéréseket.

A fonémák közötti szegmenshatárok helye nagyon sok esetben nem egyértelmű. Ilyenkor az alábbi, már a BABEL adatbázis során kifejlesztett[8] szabályrendszerhez igyekeztünk tartani magunkat:

- A szegmenshatárokat célszerű a null-átmenetekhez illeszteni. (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 jelöltük 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 legfeljebb alap-periódusnyi hosszú.
- A magánhangzó kezdetét a zönge indulásánál jelöljük (zöngétlen hang után).
- 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 egy-két periódust ingadozhat, ugyanis a hangok kettéválasztása ekkor bizonytalan (ilyenkor gyakran a szakértők is különböző helyre teszik a határt [8].

A címkézés szabályai:

- Címkézésre SAMPA karaktereket használtunk.
- A karaktereket a határbejelölések között tüntettük fel.
- Mindig azt jegyeztük le, ami ténylegesen elhangzott.
- Az aktuális kimondásnál kimaradt hangokat megjelöltük, a követő hang előtt zárójelbe tettük.
- A beszéd közben tartott szünetet ~ jellel jelöltük.
- A co-artikulációs zörejeket és a köhögést {spk} karakterekkel jelöltük.
- A kitöltött szünetet (az un. ö-zést {fil} karakterekkel jelöltük.
- Zajt akkor jelöltünk, ha az egyértelműen azonosítható volt, és a zaj nem a környezeti zajhoz tartozott. A tranziens zajt {int}, a stacioner zajt {sta} karakterekkel jelöltük.

5 Automata szegmentálás

A fonéma-szintű szegmentálást és címkézést tökéletesen csakis nagy figyelmet igénylő fáradságos és hosszadalmas kézi munkával lehet elvégezni. Megkönnyítheti és meggyorsíthatja viszont a munkát egy megfelelő, speciálisan erre a célra kialakított algoritmus, amely megkísérli automatikusan elhelyezni a fonetikai határokat. Habár ez a feladat jelenlegi tudásunk szerint tökéletesen nem megoldható, olyan program azért készíthető, amely a határok jó részét elég pontosan helyezi el. A szegmentáló személy feladata ilyenkor csak az automata szegmentáló javaslatainak ellenőrzése és korrekciója, ami által a szegmentálás elvégzése felgyorsítható.

Az automata szegmentálási algoritmusok közül a legjobbak azok, amelyek gépi tanuláson alapulnak. Speciálisan, a gépi beszédfelismerők is felhasználhatók a szegmentálási feladat elvégzésére. A beszédfelismerő algoritmusok ugyanis egy mondat felismerése közben egy keresését végeznek: a felismerendő hangjelre megpróbálják ráilleszteni az összes lehetséges fonetikai átiratot. Mivel a felismerés során a fonémák határai sem ismertek, ezért a felismerők végigpróbálják az összes lehetséges szegmentálást is. A felismerés eredményeképpen azt az átiratot illetve szegmenshatár-

sorozatot adják vissza, amelyet az adott jel esetén a legvalószínűbbnek találtak. Egy beszédfelismerési alkalmazás esetén persze nincs szükségünk a szegmenshatárokra, így az eredménynek ezt a részét figyelmen kívül hagyjuk. Viszont a felismerőnek ez az amúgy rejtve maradó „képessége" remekül kihasználható az automatikus szegmentálás céljaira. Ilyenkor ráadaásul könnyebb a feladat, mint a felismerés esetén, ugyanis a fonetikai átirat adott, így azt nem is kell keresni; csupán az adott átirat és a jel közötti legoptimálisabb fonémahatár-összerendelést kell megtalálni.

Miután az első 100 beszélő hanganyagának feldolgozása elkészült, mi is megkíséreltük egy automata szegmentáló rutin elkészítését, mégpedig a fent leírt módon, egy beszédfelismerő felhasználásával. A beszédfelismerő az MTA-SZTE Mesterséges Intelligencia Kutatócsoportnál fejlesztett „OASIS" rendszer volt, amely fonéma-alapú, és mesterséges neuronhálókat alkalmaz a fonémafelismerésre. Az OASIS rendszer fonémafelismerési technológiájáról bővebben [11]-ben tájékozódhat az olvasó; a felismerő mondat-szintű működési elvét pedig [12]-ben ismertettük. Az alábbiakban röviden összefoglaljuk a rendszer mögött meghúzódó matematikai elveket.

5.1 A beszédfelismerés és az automatikus szegmentálás kapcsolata

A gépi beszédfelismerő algoritmusok statisztikai, illetve valószínűségszámítási alapokon nyugszanak. Kiindulópontjuk az a döntéselméleti alaptétel, hogy a döntésből származó hibák számának minimalizálásához mindig az (a posteriori) legvalószínűbb lehetőséget kell választani [13]. Azaz, ha *X*-szel jelöljük a felismerendő objektumot (esetünkben a beszédjelet), *W*-vel pedig a lehetséges osztályokat (esetünkben a lehetséges mondatokat), akkor a felismerő által keresett $\hat{W}$ output:

$$\hat{W} = \arg\max_{W} P(W|X).$$

Ebből következik, hogy a gépi tanulás célja a $P(W|X)$ eloszlás minél pontosabb modellezése. Mivel azonban mind a lehetséges *W* mondatok száma, mind a lehetséges akusztikai megfigyelések *X* tere rendkívül nagy, ezért *W*-t és *X*-et valamilyen módon fel kell bontani. Magyarul, mivel képtelenség teljes mondatokat modellezni, ezért azokat valamilyen kisebb egységekre kell osztani. A legkézenfekvőbb ilyen egység a fonéma. Ezért tegyük fel, hogy a *W* fonetikus átirat fonetikai szimbólumokból áll, azaz $W = w_1 w_2 ... w_N$. A felismerés során viszont ezeknek az építőelemeknek nem csak a milyenségét nem tudjuk, hanem azt sem, hogy hányan vannak, és hogy akusztikus megfelelőjük hol helyezkedik el a jelben. Mivel – legalábbis elvileg – az egyes fonémák határa bárhol lehet, ezért minden eshetőséget meg kell vizsgálni. Ha *S* jelöli az *X* megfigyelés lehetséges szegmentációit, akkor a teljes eseményterekre vonatkozó alapvető valószínűségszámítási összefüggés szerint

$$P(W|X) = \sum_{S} P(W|X,S) \cdot P(S|X).$$

Ha feltételezzük, hogy a jelnek van egy „helyes" szegmentálása és a többi elég valószínűtlen, akkor a fenti képletben az összegzés jól közelíthető maximalizálással:

$$P(W|X) \approx \max_{S} P(W|X,S) \cdot P(S|X).$$

Ezt a képletet beágyazva a legelsőbe azt kapjuk, hogy a (fonéma alapú) beszédfelismerés voltaképpen két dimenzió szerinti keresést jelent: egyrészt meg kell találnunk az *X*-hez legjobban il-

leszkedő *W*-t, másrészt minden egyes *W* esetén meg kell találnunk a legjobb *S* szegmentálást. Ebből következik, hogy a (fenti elven működő) beszédfelismerő nyilvánvalóan felhasználható automatikus szegmentálásra is, hiszen a felismerés során implicit módon minden jelnek megkeresi az optimális szegmentálását. Az automatikus szegmentálás esetén annyi a könnyebbség, hogy a *W* fölötti keresést nem kell elvégezni, hiszen a fonetikai átirat adott. Tehát csak az adott átirat adott jelhez való legjobb illeszkedését kell megtalálni, ezért az ilyenfajta felhasználást „forced alignment"-nek nevezi a szakirodalom.

5.2 Beszédfelismerés és automatikus szegmentálás neuronhálókkal

A gyakorlatban használt beszédfelismerők túlnyomó része a Hidden Markov Model elnevezésű technológián alapszik [14], amely – a Bayes-tételt kihasználva – $P(W|X)$ helyett $P(X|W)$-et, azaz az ún. „class-conditional" eloszlást modellezi (ami nem befolyásolja lényegesen a fenti okfejtés érvényességét). Ez azt jelenti, hogy az egyes fonémákhoz tartozó lehetséges akusztikai jelek eloszlását tanulják, többnyire Gauss-eloszlások súlyozott összegeként.
Alternatív lehetőségként egyre több helyen alkalmazzák a mesterséges neuronhálókat, amelyek viszont az eredeti a posteriori, azaz a $P(W|X)$ eloszlás modellezésére használhatók fel [13]. Mivel a mi rendszerünk ez utóbbi elven alapszik, ezért a továbbiakban csak ezt ismertetjük.
Alakítsuk hát tovább az előbbiekben $P(W|X)$-re felírt közelítő képletet. Láthatjuk, hogy egy szorzatból áll, tehát két fő komponense van. Vizsgáljuk most az elsőt, $P(W|X,S)$-t. Mint már említettük, a lehetséges *W*-k és *X*-ek nagy száma miatt ezt az eloszlást tovább kell bontatni kisebb egységekre, esetünkben fonémákra. Ehhez függetlenségi feltevésekkel kell élnünk a $P(W|X,S)$ eloszlást illetően. Az első függetlenségi feltevésünk az lesz, hogy a szomszédos fonémák egymástól függetlenül fordulnak elő[1]. Ezzel

$$P(W|X,S)=\prod_{i=1}^{N} P(w_i|X,S).$$

A másik feltevésünk az lesz, hogy a fonémák milyensége nem függ a teljes hangjeltől, csak annak az adott szegmentálás mellett az adott fonémához tartozó részletétől. Jelölje ezt a részletet $X_{i(S)}$. Ezzel

$$P(W|X,S)=\prod_{i=1}^{N} P(w_i|X_{i(S)}).$$

A $P(w_i|X_{i(S)})$ eloszlás tehát azt adja meg, hogy egy adott akusztikus szegmentum milyen valószínűséggel tartozik az egyes fonémaosztályokba. A rendszernek ezt a komponensét ezért *fonéma-osztályozónak* fogjuk nevezni. Ez a valószínűség jól modellezhető neuronhálókkal, ennek egyetlen feltétele, hogy a szegmenseket (amelyek hossza változó) mindig ugyanannyi számú jellemzővel írjuk le.
A másik komponens a rendszerben $P(S|X)$. Ennek feladata az egyes lehetséges szegmentálások valószínűségének megadása. Ha ezt is neuronhálókkal akarjuk tanulni, akkor ismét fel kell bontani valamilyen módon. Ehhez ismét egy függetlenségi feltevésre van szükség: feltételezzük,

[1] Ez természetesen nem igaz; a fonémák egymástól való függését a beszédfelismerők nyelvi modulja szokta modellezni. Esetünkben azonban erre nem lesz szükség, hiszen a fonetikai átirat rögzített.

hogy az egyes s_i szegmentumok független módon befolyásolják a teljes szegmentálás valószínűségét. Továbbá, mivel az egyes lehetséges szegmentálások egymással „versenyeznek”, ezért a képletben nem csak az éppen vizsgált szegmentálás szegmentumait, hanem az összes többi lehetséges szegmentálás szegmentumainak értékét is szerepeltetni fogjuk. A közelítő képletünk tehát:

$$P(S|X) = \prod_{s_i \in S} P(s_i | X_{s_i}) \prod_{\bar{s}_i \in \bar{S}} P(\bar{s}_i | X_{\bar{s}_i}),$$

ahol $s \in S$ egy adott szegmentálás szegmentumait jelöli, $\bar{s} \in \bar{S}$ pedig az összes többi lehetséges szegmentálásban előforduló szegmentumokat. $P(s_i|.)$ azt a valószínűséget jelenti, hogy az adott szegmens egy fonéma, $P(\bar{s}_i|.)$ pedig azt, hogy nem-fonemikus szegmentumról, ún. „antifonémáról” van szó (azaz egy fonéma-részletről vagy több fonémát átfedő darabkáról). A szegmentumoknak ez a kétosztályos klasszifikációja ismét jól tanulható neuronhálókkal.
A gyakorlatban túl sok a lehetséges szegmentumok száma, ezért a fönti képletet tovább egyszerűsítettük azzal, hogy az $\bar{s} \in \bar{S}$ szegmentumok közül csak azt a kettőt vettük figyelembe, amelyek – határaikat tekintve – legközelebb esnek *s*-hez [12].

5.3 A tanításhoz használt jellemzők

Mint minden beszédfelismerési feladatnál, így a tanuláson alapuló automatikus szegmentációnál is kulcsfontosságú, hogy milyen akusztikus jellemzőket emelünk ki a jelből. Esetünkben az egyes szegmentumok leírására az alábbi jellemzőket használtuk (összesen 102 jellemző):

A, **Kritikus sávok energiája logaritmikus skálán**. Az emberi fül a hang frekvenciafelbontását logaritmikus skálán végzi, azaz a magasabb frekvenciákon rosszabb felbontóképességgel rendelkezik, mint az alacsonyakon. Továbbá az egymáshoz közel, egy ún. kritikus sávon belül eső frekvenciákat hasonló módon dolgozza fel. Ennek a logaritmikus skálának a leírására többen dolgoztak ki (lényegében hasonló) képleteket, mi az ún. bark-skálát használtuk. A telefonos beszédminták 8000Hz-es mintavételi frekvenciájából következően a jelekből megmaradó 4000Hz-es sávot 16 bark-szűrőre bontottuk fel; a szűrőkbe eső komponensek energiáját FFT-vel számolt energiaspektrumból kiindulva trianguláris súlyozással számoltuk. Mivel az egyes fonémák hossza változhat, ezért a szegmentumokat időben 5 sávra bontottuk: a szegmentumot magát három harmadra vágtuk, és ugyanilyen méretű egy-egy sávot figyelembe vettünk a szegmentum két szomszédjából is, a kontextus modellezésére. Ezeken a sávokon átlagoltuk a bark-szűrők kimenetét, így tehát 5*16=80 jellemzőt kaptunk.

B, **Modulációs spektrum és deriváltak.** Bizonyos kutatási eredmények arra mutatnak, hogy az emberi hallás a szomszédos kritikus sávokat összevonva kezeli. Klasszifikációs vizsgálataink is arra mutattak, hogy bizonyos hang-kategóriák megkülönböztetésére elégséges a spektrumnak egy durvább felbontását vizsgálni. Ehhez a spektrumot 3, bark-skálán egyenletes szélességű sávra bontottuk, és ebből a durva felbontásból kétféle jellemzőt vontunk ki: Egyrészt a sávok ún. modulációs spektrumát, pontosabban annak 4Hz-es komponensét, amely feltehetőleg nagy szerepet játszik a szótagok elhatárolásában. Továbbá minden sávon vizsgáltuk az energiákat illetve azok első két deriváltját a szegmentumok két határán – ezen jellemzőkkel a szegmensek pontos határának felismerését kívántuk elősegíteni. Ezek a jellemzők tehát további 3+2*3*3=21 jellemzőt adtak szegmentumonként.

C, **Hossz.** A harmadik felhasznált jellemzőtípus a szegmentumok hossza volt. Ennek rendkívül nagy szerepe van bizonyos rövid-hosszú hangpárok elkülönítésében, mivel azok spektruma a hossztól eltekintve nem különbözik lényegesen.

5.4 Automatikus szegmentálás a gyakorlatban

Az első száz beszélő hanganyagában összesen 62015 fonéma-előfordulás volt, ezeken tanítottuk be a rendszert. Az általános tapasztalat a rutinnal az volt, hogy általában jó helyre teszi a határokat, de nem azzal a nagy pontossággal, amit az adatbázis specifikációja során magunk elé tűztünk. Ezért csak a szegmentálási folyamat első lépését, a durva szegmentálást volt képes kiváltani; a határok finom beállítása a rutin elkészülte után is kézi feladat maradt.

5. A beszélők demográfiai adatai

5.1. 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 médiákban 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. Ennek ellenére Magyarországon belül négy dialektus-területet különítettünk el; ennek megfelelően a következő négy kategóriába soroltuk a beérkező hívásokat (ld. a 2. ábrát):

- Észak-Magyarország (NORTHERN)
- Magyarország középső része (SOUTHERN)
- Nyugat-Magyarország (WESTERN)
- Kelet-Magyarország (TISZA)

2. ábra. A Magyarországon kijelölt dialektus-területek

5.2. A beszélők neme és életkor szerinti megoszlása

Az adatbázisnak tükröznie kellett a várható felhasználók kor szerinti eloszlását. Statisztikai adatok alapján ezt az adatbázis 4. táblázatban közölt kor szerinti eloszlása jól közelíti.

4. táblázat: A beszélők neme és életkor szerinti megoszlása

Korcsoport	Férfi beszélők	Női beszélők
16 év alatt	2,20%	1,80%
16...30 év	22,80%	26,20%
31...45 év	11,60%	13,60%
46...60 év	9,40%	9,80%
60 év felett	2,00%	0,60%

6. Eredmények

A BABEL magyar tisztabeszéd adatbázis mellett tehát most létrejön egy magyar telefonbeszéd adatbázis amitől a magyarországi beszédtechnológiai kutatások és fejlesztések jelentős fellendülését várjuk. Az adatbázis alkalmas személy független beszédfelismerők betanítására és tesztelésére. Ezzel megnyílik az út különböző telefonon keresztül működő beszéd felismerési szolgáltatások kifejlesztésére. Ilyen szolgáltatások például szóval vezérelt alközponthívás. A dialógusrendszerek létrehozása közvetlen automatikus információszolgáltatást tesz lehetővé. A szokásos megoldásban a telefonáló egy dialógus keretében elmondja, hogy mit szeretne megtudni, és a rendszer beszéd szitetizátora közli a kívánt információt, pl. banki árfolyamokat, vonatmenetrendet, stb.
Nyelvészeti, beszéd tehnológiai kutatási célokra is kiválóan alkalmas lesz az összegyűjtött hanganyag, különösen a felszegmentált, címkézett szavak és mondatok.
500 felvétel természetesen nem sok, ha figyelembe vesszük a beszéd, a felvételi körülmények és az átvitel variáltságát. Igy ezzel a hanganyaggal egy meghatározott felismerési pontosság érhető el. Annak ellenére, hogy a beszéd gyűjtés és feldolgozás igen költséges feladat, további adatbázis készítésre van még szükség.

7. Köszönetnyilvánítás

A telefonbeszéd-adatbázist az Oktatási Minisztérium IKTA 3 kutatási és fejlesztési programja keretében hoztuk létre.

8. Irodalomjegyék

[1] Proceedings of the 1992 Workshop of the International Coordinating Committee on Speech Database and Speech I/Q Systems Assessment, Banff, Canada, October 1992.

[2] Pollak, P., Cernocky, J., Boudy, J., Choukri, K., Heuvel, H. van den, Vicsi, K., Virag, A.,Siemund, R., Majewski, W.,Sadowksi, J., Staroniewicz, P., Tropf, H., Kochanina, J., Ostrouchov, A., Rusko, M., Trnka, M.SpeechDat(E) - Eastern European Telephone Speech Databases. Proceedings LREC'2000 Satelite workshop XLDB - Very large Telephone Speech Databases, 29 May 2000, Athens, Greece, pp. 20-25.

[3] Vicsi K., Vig A.: Text independent neural network/rule based hybrid, Continuous Speech Recognition System Eurospeech'95, 4th European Conference on Speech Communication and Technology, Proceedings Volume 3, pp: 2201-2204.

[4] Vicsi, K.: Beszdéadatbázisok a gépi beszédfelismerés segítésére, 2001 Hiradástechnika 2001/I, Budapest, pp. 5-13

[5] LIAS: Language Independent automatic Segmentation Technique Using Sampa Labelling of Phonemes. First International Conference on Language Resources Education, Granada Spain, 1998. 1.317

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

[7] W.J. Barry and A.J. Fourcin „Levels of labelling", Computer Speech and Language, Vol. 6, pp. 1-14. (1992)

[8] Vicsi K., Vig A.: Az első magyarnyelvű beszédadatbázis, Beszédkutatás '98, MTA Nyelvtudományi Intézete, Budapest 1998, pp. 163-177

[9] 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).

[10] Chan, D., Fourcin, A. and others, „EUROM - A Spoken Language Resource for the EU", Proceedings of Eurospeech'95, Madrid, Vol. 1, pp. 867-870, (1995).

[11] Kocsor, A., Tóth, L., Kuba, A. Jr., Kovács, K., Jelasity, M., Gyimóthy, T., Csirik, J.: A Comparative Study of Several Feature Space Transformation and Learning Methods for Phoneme Classification, International Journal of Speech Technology, Vol. 3, Number 3/4, 2000, pp. 263-276

[12] Tóth, L., Kocsor, A., Kovács, K.: A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition, in: P. Sojka, I kopecek, K. Pala (eds.): TSD'2000, LNAI 1902, pp. 307-313, Springer Verlag, 2000

[13] Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification, Wiley and Sons, 2001

[14] Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing, Prentice Hall, 2001