Többnyelvű európai híranyag-adatbázis gyűjtése és feldolgozási módszereinek kutatása multimédiás műsorok automatikus feldolgozásához

TELEKI CSABA, VICSI KLÁRA

BME Távközlési és Médiainformatika Tanszék, Beszédakusztikai Kutatólaboratórium
{vicsi, teleki}@tmit.bme.hu

Lektorált

Kulcsszavak: digitális jelfeldolgozás, beszédfeldolgozás, beszédadatbázisok

Többnyelvű híranyag-adatbázisok (Broadcast News – BN) gyűjtése és ezek egységes elvű feldolgozási módszereinek kidolgozására nemzetközi munkacsoport jött létre a COST278 EU projekt keretein belül. A BME TMIT Beszédakusztikai Kutatólaboratóriuma a csoport tagjaként magyar híranyag-adatbázist hozott létre, amely 3 óra és 30 percnyi kép- és hanganyagot tartalmaz. Az adatbázis feldolgozásához a BN munkacsoport által kidolgozott módszereket és előírásokat használta fel, ilyen például az átíró és annotáló szoftver, amely a NIST (National Institute of Standards and Technology) ajánlásai alapján készült. Az átiratok egységes formátumra hozása érdekében, a NIST ajánlásai mellett, pontos címkézési módszereket, szabályokat hoztunk létre. Kutatócsoportunk másik célkitűzése az volt, hogy a beszéd akusztikai paramétereire támaszkodva különböző nyelvfüggetlen, kiértékelő eszközöket fejlesszen ki (beszéddetektálás, beszélő nemének meghatározása stb.). E tanulmányban laboratóriumunk magyar nyelvre vonatkozó feldolgozási módszereit mutatjuk be, valamint tesztelési eredményeinket hasonlítjuk össze a munkacsoport tagjai által elért eredményekkel.

1. Bevezetés

Köztudott, hogy napjainkban a média egyre nagyobb teret hódít és talán nem is tudatosul bennünk, fogyasztókban, hogy a rádiós, televíziós műsorok, híradások egy akusztikus számára is új kutatási területek szinte kiaknázhatatlannak tűnő tárházát jelenti. Egyre több kereskedelmi csatornát hallgathatunk a rádióban, nézhetünk a televízióban, és ezek mindegyike megpróbálja egyéni arculatát megteremteni, amely a képi ábrázolás mellett egyfajta „akusztikai arculatot” is definiál. A televíziós híradások világának akusztikai vizsgálatát céloztuk meg, támaszkodva a képi anyagra is, bár kisebb mértékben.

A COST278 BN munkacsoportban létrehozott európai többnyelvű (flamand, portugál, gall, cseh, szlovén, szlovák, görög, horvát és magyar) híranyag-adatbázis alapul fog szolgálni a multimédiás műsorok automatikus feldolgozásához, például a híranyag automatikus lejegyzéséhez, reklámsugárzás számlálásához, stb. A beszédtechnológiai kutatások egyik kiemelkedő területe a híranyagok automatikus lejegyzése. Ez a lejegyzés annál pontosabb, minél nagyobb és minél jobban feldolgozott adatbázissal történik a használt beszédfelismerő rendszer betanítása, tesztelése [1]. Éppen ezért igen nagy fontossága van annak, hogy milyen módszerrel történik az adatbázis feldolgozása. A munkacsoport célul tűzte ki, hogy az Amerikai Egyesült Államokban folyó kutatásokat is figyelembe véve (HUB4 amerikai híranyag korpusz [2]) egy európai feldolgozási és értékelési módszert dolgozzon ki [3].

A kifejlesztett algoritmusokat a többnyelvű BN adatbázisokon teszteltük, kiértékelve a kidolgozott algoritmusok hibáit és előnyeit, így érve el egyre jobb feldolgozási eljárásokat, amelyeket a munkacsoport minden tagja használ. Például az akusztikai és nyelvi feldolgozás során, a NIST ajánlásai alapján, elkészült egy újfajta annotáló eljárás [4], amelyet a munkacsoport minden tagja használt.

Az eredmények kiértékeléséhez két különböző szoftver került kifejlesztésre. A portugál partner által biztosított kiértékelő szoftvert [7] használtuk a beszéd-nem-beszéd detektáló algoritmusok eredményeinek kiértékelésekor, míg a beszélő csoportosító, a beszélő nemét detektáló szoftverek eredményeinek kiértékeléséhez a belga partner által közreadott szoftvert [8] használtuk.

E tanulmányban a magyar nyelvű híranyag-adatbázis gyűjtéséről, az akusztikai és nyelvi feldolgozásáról, valamint a szegmentálásról és kiértékelésről számolunk be, továbbá bemutatjuk az általunk kifejlesztett új és sikeres beszéd-nem beszéd detektáló eljárást.

2. Adatgyűjtés

Laboratóriumunkban korábban több adatbázist is gyűjtöttünk, melyeknek felhasználási célja különböző volt. Készült fonetikai kutatások céljából egy adatbázis (BABEL) [12], melynek szöveganyaga süketszobában került rögzítésre, így a felvételek során a jel-zaj viszony magas volt. A bemondott szöveganyag precízen összeválogatott mondatokból állt, mivel cél volt, hogy az adatbázis tartalmazza legalább kétszer a magyar nyelvben előforduló félszótagok 98 százalékát. Tartalmaz továbbá számokat és CVC (mássalhangzó-magánhangzó-mássalhangzó) kapcsolatokat is. A bemondók száma kicsi (60 beszélő) és nagy hangsúlyt fektettünk arra, hogy a beszélők szépen, artikuláltan beszéljenek.

Egy másik jellegű adatbázis a Magyar Telefonos Beszéd-Adatbázis (MTBA) [13]. Az adatbázis 500 magyar nyelvű beszélő hanganyagát tartalmazza, ezekből 297 vezetékes, 203 pedig mobil telefon-felvétel. Az adatbázis általános fonetikai, nyelvészeti kutatásokhoz szolgál alapul és statisztikai feldolgozási módszereken alapuló személyfüggetlen gépi beszédfelismerők, dialógusrendszerek létrehozását teszi lehetővé.

A Magyar Referencia Beszéd-Adatbázis esetén a cél egy olyan, olvasott folyamatos szöveget tartalmazó beszédadatbázis létrehozása volt, amely alkalmas PC-s beszédfelismerők betanítására, tesztelésére [13]. Az adatbázis szöveganyagát úgy terveztük meg, hogy az adatbázisba bekerülő mondatokban a felismerő rendszerekben tipikus felismerési egységek (beszédhangok, difón, trifón egységek) elegendően sokszor forduljanak elő. A mondatok mellett fonetikailag gazdag szavakat is kiválasztottunk, a nem kellő számban előforduló beszédhangok példányszámának növelése érdekében. Így a 332 adatközlő fejenként 12 különböző mondatot és 12 különböző, a mondatoktól független szót olvas fel. Az adatbázis felvételeit irodai helyiségekben, laborokban, otthonokban rögzítettük.

Látható, hogy a fentebb említett adatbázisok esetén bizonyos szempontok alapján megtervezett szöveget mondott be a beszélő egy (vagy több) számítógéphez csatlakoztatott mikrofonba, vagy egy telefon mikrofonjába. A COST278 munkacsoport által létrehozott adatbázist a fentebb említett adatbázisokhoz képest egy teljesen más szemlélet jellemez, hiszen a híranyag-adatbázis egy többrétű, multimédiás adatbázis, melynek feldolgozása során a felhasználónak alkalma nyílik egy akusztikai szempontból is sokkal gazdagabb anyagba betekinteni. A kutatócsoport minden tagintézménye egy legalább három órás adatbázist gyűjtött nemzeti (közszolgálati/kereskedelmi) televíziók hírműsoraiból. Jelenleg a teljes BN adatbázis 30 órányi anyagot tartalmaz, melyet 10 különböző televízióállomástól rögzítettünk 9 európai nyelven: flamand, portugál, gall, cseh, szlovén, szlovák, görög, horvát és magyar. A magyar nyelvű BN adatbázis körülbelül 3 óra 30 percnyi híranyagot tartalmaz, melyet közszolgálati és kereskedelmi adók műsoraiból rögzítettünk az *1. táblázatban* bemutatott arányban.

1. táblázat A magyar nyelvű BN adatbázis struktúrája

TV állomás	*Időtartam (perc:mp)*	*Híradások száma*	*Műsorvezetők száma*
MTV1	77:09	3	4
RTL Klub (k)	40:51	5	2
TV2 (k)	84:40	3	4

A táblázatban a (k) jelzéssel ellátott televízióállomások kereskedelmi adók. Szerepel továbbá a felvételek hossza televízió-állomásonként, a híradások száma összesen, illetve az adatbázisban rögzítésre került hírműsorok különböző műsorvezetőinek száma.

Látható, hogy a magyar nyelvű BN adatbázis összetett, hiszen az adatbázis három különböző televíziócsatorna különböző típusú hírműsorainak anyagát tartalmazza. Többnyire kereskedelmi adók hírműsorai kerültek be az adatbázisba (2 ó, 5 p és 31 mp, ami az adatbázis kétharmadát teszi ki), a fennmaradó egyharmad tartalmazza a közszolgálati televízió híradóit. Nyilván az, hogy a magyar nyelvű BN adatbázis igen heterogén, hatással lesz a kutatásaink eredményeire.

A felvételek egy személyi számítógéphez csatlakoztatott televíziós készülék segítségével készültek. A számítógép egy speciális jelfeldolgozó kártyával volt felszerelve, így lehetőség nyílt arra, hogy ne csak a hanganyagot, hanem a képi anyagot is rögzíteni tudjuk. A hanganyag digitalizálásakor 16 kHz-es mintavételi frekvenciát használtunk, tároláshoz pedig a hanghullám (wave) formátumot használtuk a következő paraméterekkel: 16 kHz mintavételi frekvencia 16 biten ábrázolva PCM kódolással, 256 kbit/s-os bitsebességgel. A képi anyag tárolásakor két szempontot vettünk figyelembe: egyrészt, hogy megfeleljen a COST278 BN kutatócsoport ajánlásainak, másrészt, hogy a képi anyag valós segítséget tudjon nyújtani az átírás során. Ezért két különböző tömörítési eljárást használtunk. Az egyik, a COST278 BN kutatócsoport ajánlásainak megfelelően a következő volt: Indeo® video 5.11 verziójú kodek 930 kbit/s-os bitsebességgel (a kép mérete: 180x144 pixel, 25 képkocka másodpercenként). A kép mérete miatt, az ily módon tárolt felvételek nem voltak igazán használhatóak az átírás, címkézés során, ezért belső használatra elkészítettünk egy 360x288 pixel méretű képi anyagot, melyet DivX 5.0.5 verziójú kodekkel tömörítettünk, 998 kbit/s bitsebességgel.

A magyar BN adatbázis, mely tartalmazza a hanganyagot, a képi anyagot és az átiratokat is, CD lemezeken és egy belső használatra létrehozott szerveren tároljuk, melyhez a kutatócsoport minden tagja hozzáfér. Minden adatfájl egyedi névvel rendelkezik, melyből kiderülnek a felvételre jellemző legfontosabb paraméterek, a következőképpen:

`<tv_csatorna_név>_<év>_<hhnn>_<óópp>`
`.<kiterjesztés>,`

ahol a `<tv_csatorna_név>` annak televíziós csatornának a neve, ahonnan a híradó rögzítésre került, az `<év>_<hhnn>_<óópp>` paraméterek a rögzítés dátumát és pontos kezdési idejét tartalmazza. A fájlok kiterjesztése pedig lehet *wav*, amennyiben hangfájlról van szó, *avi*, amennyiben a képi anyagról van szó és *trs, seg* vagy *stm*, amennyiben adatfájlokról van szó (átirat, címkézés).

3. Az adatbázis akusztikai és nyelvi feldolgozása

Az adatbázis akusztikai és nyelvi feldolgozása során nagyon fontos a hanganyag átírása, címkézése. A címkézés során a LDC (Linguistic Data Consortium) ide vo-

natkozó ajánlásait követtük. Mivel a LDC ajánlásai nem voltak eléggé konkrétak és sok hiba forrásául szolgálhattak, kiegészítettük ezeket a BN kutatócsoport által ajánlott szabályokkal [3]. Ezáltal nagyobb lett az esély arra, hogy a BN kutatócsoport résztvevői megfelelően pontos és hasonló módszerekkel készítsék el a címkefájlokat, elősegítve ezzel a közös munkát.

3.1. Címkézési szabályok

A címkézés során jelöltük a beszélőváltások során fellépő akusztikai változásokat, a beszélő által elmondott szöveg határait, a híradások szekcióit, a híradások szignáljainak kezdetét és végét, idegen nyelvű beszédet, háttérzajt és a beszélő által keltett zajokat.

A beszélőváltások során fellépő akusztikai változásokat az átviteli csatorna milyensége és annak minősége határozta meg. Két fajta átviteli csatornát különböztettünk meg (stúdióban elhangzott beszéd vagy telefonon keresztül elhangzott beszéd) és mindegyik csatornát minősítettük azok akusztikai minősége szerint (jelzaj viszony alacsony, közepes vagy magas). Jellemzően a stúdióban elhangzott beszélgetéseket, a stúdióból kommentált riportokat, illetve a műsorvezető beszéde során elhangzó hanganyagot a *„stúdió", „magas"* címkékkel láttuk el. A *„stúdió, közepes"* (jel-zaj viszony) címkével akkor jelöltük a beszédet, ha a riporter stúdión kívül beszél, jellemzően ezt a címkét az utcán, vagy nyílt terepen elhangzott beszéd kapta. A *„stúdió, alacsony"* címkével a különösen zajos környezetben készített felvételeket láttuk el. A telefonos beszéd esetén a tiszta beszédet a *„magas"* címkével, a zajos, de még érthető beszédet a *„közepes"* címkével, míg a nehezen érthető beszédet az *„alacsony"* címkével láttuk el. Ezt a kódolási eljárást a *2. táblázatban* foglaltuk össze.

Egyik fontos címkézési szabály az, hogy az egy bemondó által bemondott beszédet több, kisebb egységre bontottuk, ezért a magyar nyelvű BN adatbázisban minden belélegzés egy ilyen egység kezdete is egyben. Amennyiben a beszélőváltáskor keletkezett beszédszünet kisebb 0,5 másodpercnél, nem jelöltük. Amennyiben ez a fajta szünet 0,5 másodpercnél nagyobb, de 1,5 másodpercnél kisebb, akkor ezt jelölni kellett egy címkével a szünet közepén. Amennyiben 1,5 másodpercnél nagyobb szünet keletkezik, akkor a szünet elejét is és a végét is jelöltük.

		Csatorna	
		Stúdió *(Sávszélesség>4 kHz)*	***Telefon*** *(4 kHz)*
Minőség	*Alacsony*	Zajos	Érthetetlen
	Közepes	Stúdión kívül	Zajos
	Magas	Stúdióban	Tiszta

2. táblázat A csatorna fajtája és minőségének jelölése

Adatbázisunkban a szekciók határait is jelöltük. Szekció lehet egy riport (hírértékű esemény prezentációja), kitöltő szövegek (rövidhírek, címszavak stb.) és nem átírt események (reklámok és szignálok).

Minden szignál külön címkét kapott. Mivel előfordulhat, hogy a híradó elején, végén és közben mindig különböző szignálokat hallunk, az adás eleji és az adás végi szignált egy külön címkével jelöltük. Az idegen nyelvű szöveg kezdetét címke jelzi, de nem lett átírva.

A címkefájl tartalmazza továbbá a háttérzajok kezdetét és végét jelölő címkéket is. A háttérzajokat különböző kategóriákba osztottuk: zene, beszéd, susogás, egyéb. Ugyanakkor jelöltük a beszélő által keltett zajokat is, mint a belélegzés, kilélegzés, papírzörgés stb. Az átiratok minden esetben a Transcriber [4] nevű program segítségével készültek (http://www.etca.fr/CTA/gip/Projets/Transcriber) és XML formátumú, ISO-8859-2 karakterkódolású szövegfájlban kerültek elmentésre. Az *1. ábrán* látható a Transcriber program kezelői felülete.

A kezelői felület közepén látható a felvétel idő-amplitúdó függvénye, ez képezi a választóvonalat a kezelőfelület alsó és felső része között. A kezelőfelület felső részéhez fér hozzá a felhasználó, ide írhatja le az elhangzottakat, jelölheti be a szekciók elejét, a beszélő-

1.ábra
A Transcriber program kezelői felülete

váltásokat (természetesen ilyenkor az akusztikai paraméterek változását is jelölni kell, amennyiben ez megtörténik), a beszélő által keltett zajokat (belégzés *'[i]'*, kilégzés *'[e]'*, papírcsörgés *'[pap]'*, stb.), a háttérzajokat stb.

Amennyiben a felhasználó mindezeket helyesen jelöli, a kezelőfelület alsó részén megjelenik hierarchikus formában a bejelölt információ (fentről lefelé haladva): a háttérzajok sávja szürke, amennyiben létezik bejelölt háttérzaj (zene, susogás stb.), alatta jelenik meg a szekció sávja, amelyben a szekció fajtája van bejelölve (riport – „report", kitöltő szövegek – „filler", vagy nem átírt szöveg – „nontrans"). Amennyiben a szekció a „report" vagy „filler" címkét kapta, a felhasználó egy néhány címszavas leírást is írhat a szekció tartalmáról. A szekció sávja alatt található a beszélő sávja. Ebbe a sávba kerülnek az egy szekción belül előforduló bemondók nevei, vagy ennek hiányában valamilyen egyedi azonosító. Minden beszélő esetén egy adatlapot kell kitölteni a beszélőre jellemző adatokkal (a bemondó neve, annak neme, műsorvezető-e vagy sem, akusztikai környezet stb.)

Mivel nem mindig hangzik el a bemondó neve, nagy segítséget tud nyújtani ezen űrlap kitöltésekor a képi anyag. Amennyiben nincs beszéd, a „no speaker" címke kerül ebbe a sávba. A beszélő sávja alatt található a bemondott szöveg átiratának sávja, alatta pedig az időcímkék.

Említettük, hogy kimenetként ez a program egy XML kódolású, igen nehezen kezelhető, szövegfájlt produkál. Az alábbiakban egy részletet tekinthetünk meg belőle:

```
<?xml version="1.0" encoding="ISO-8859-2"?><!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="(unknown)" audio_filename="MTV1_2~1" version="17" version_date="040509">
<Topics>
<!— a híradóban elôforduló témakörök felsorolása-->
     <Topic id="to1" desc="Takarékossági csomag - egészségügy"/>
     […]
     <Topic id="to12" desc="elköszönés"/>
</Topics>
<Speakers>
<!— a híradóban elôforduló beszélôk felsorolása és azok paraméterei -->
     <Speaker id="spk1" name="Hajdú Andrea" check="yes" type="female" dialect="native"
accent="" scope="local"/>
     […]
     <Speaker id="spk6" name="reporterno1" check="no" type="female" dialect="native"
accent="" scope="global"/>
</Speakers>
<Episode>
<Section type="report" startTime="0" endTime="122.063" topic="to1">
<!— a szekció kezdete, idôcímkék -->
     <Turn startTime="0" endTime="16.431">
     <!— beszélô váltás kezdete, idôcímkék, akusztikai paraméterek-->
          <Sync time="0"/>
          <Event desc="jingle" type="noise" extent="instantaneous"/>
     </Turn>
     <Turn speaker="spk1" mode="planned" fidelity="high" channel="studio"
startTime="16.431" endTime="42.03">
          <Sync time="16.431"/>
           <Event desc="i" type="noise" extent="instantaneous"/>
            <!— zajesemény (belégzés) leírása -->
           Jó reggelt kívánok <!— az elhangzott szöveg -->
          <Event desc="e" type="noise" extent="instantaneous"/>
          !
          <Sync time="18.686"/>
     </Turn></Section>
     […]
</Episode></xml>
```

Látható, hogy a Transcriber program által kimenetként előállított fájl nehézkesen olvasható, nehézkesen dolgozható fel, ezért ezt a fájlformátumot egy könnyebben kezelhető fájlformátumra konvertáltuk. Erről bővebben a következő szakasz ad tájékoztatást.

3.2. Adatbázis értékelés (statisztikák)

A magyar BN adatbázisban 2425 mondatot különböztettünk meg, amelyek közül 2382 mondat került átírásra.

Összesen 22.500 szó szerepel az adatbázisban, melyek közül a különböző szavak száma 8147. Összevetve a COST278 BN kutatócsoport tagjainak adatbázisaival, azt tapasztaltuk, hogy a magyar adatbázis a cseh és a szlovák adatbázisokkal hasonlítható össze a fenti számok alapján. A cseh adatbázisban előforduló szavak száma 27.642, míg a különböző szavak száma 8834 (a cseh adatbázis 181 percnyi híranyagot tartalmaz). A szlovák adatbázisban 25.770 szó található, a különböző szavak száma 8887 (a szlovák adatbázis 191 percnyi híranyagot tartalmaz) [3].

Mint azt láthattuk, az átirat során keletkezett fájlformátum nem mondható ideálisnak automatikus gépi feldolgozáshoz. Ezért ezt egy olyan formátumra konvertáltuk, amelyben soronként a következő információkat rögzítettük:

```
[fájlnév] 1 [bemondó neve] [idôcimke1]
[idôcimke2] <o,[F állapot],
[beszélô neme]> [bemondott szöveg]
```

Például:

```
MTV1_2004_0220_1200 1 Rábai_Balázs 395.151 408.813
<o,F0,male> [i] Bizonytalanná vált a ^szegedi légi-
mentôk mûködése. [i] A szolgálatot fenntartó alapít-
vány kormányzati támogatása [e] több, mint harminc
százalékkal, huszonnyolc millió forinttal [pap] csök-
kent a tavalyihoz képest.
```

Látható, hogy a bemondott szövegben már bejelöltük a beszélő által keltett zajokat is. Az időcímkék határozzák meg a bemondás kezdetét és végét ms-ban. Ebben a formátumban a csatorna minőségét és milyenségét is átkódoltuk a könnyebb kezelhetőség és a pontosabb leírás érdekében (*F-állapotok–F-conditions,* további információk: http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/conventions.html).

A *3. táblázatban* bemutatjuk az F-állapotok szerinti statisztikát a magyar nyelvű BN adatbázis esetén. A táblázat világosan mutatja, hogy a telefonon keresztüli interjúkészítés igen kedvelt módszer a magyar médiában, hiszen közel az adatbázis közel 18%-a telefonon keresztül bemondott szöveget tartalmaz. A COST278 BN adatbázisainak statisztikái szerint a telefonos interjúkészítés hungaricum, hiszen partnereink adatbázisaiban elenyésző mértékben volt jelen az ilyen körülmények között rögzített beszéd (kevesebb, mint 4% felelt meg az F2 állapotnak).

F-állapot	*Időtartam (óó:pp:ss)*	*Százalék*
Szélessávú tiszta (olvasott) beszéd – F0	00:22:07	11.43 %
Szélessávú tiszta (spontán) beszéd – F1	00:17:24	9.1 %
Telefonos beszéd – F2	00:34:40	17.72 %
Beszéd háttérzene jelenlétében – F3	00:13:15	6.95 %
Rossz akusztikai viszonyok között elhangzott beszéd – F4	01:31:03	45.43 %
Nem anyanyelvű beszélő által mondott szöveg – F5	00:01:16	0.83 %
Egyéb beszéd – FX	00:16:27	8.54 %

3. táblázat
F-állapotok a magyar nyelvű BN adatbázisban

A televíziós híradás egy másik jellemzője az, hogy a riportokat a hírértékű esemény megtörténtének helyszínén készítik részben, vagy akár teljes egészében. Ez az adatbázisunk statisztikájában a rossz akusztikai viszonyok között elhangzott, az F4 állapotú felvételek formájában jelennek meg, amelyek a teljes adatbázis csaknem felét teszik ki. Ugyanakkor egy másik magyarázat erre a tényre az lehet, hogy a magyar BN adatbázis kétharmada kereskedelmi adók hírműsorait tartalmazza. Azért lehet ez is egy magyarázat, hiszen közkedvelt a kerekedelmi adók híradóiban a viszonylag hangos háttérzene alkalmazása.

Ez a 45%-os arány átlagosnak mondható, hiszen a többi BN adatbázis statisztikája is az F4 állapotra ezt a százalékos arányt prezentálja többé-kevésbé. Kivételként megemlíthető a két szélsőértéket produkáló BN adatbázis: a portugál nyelvű BN adatbázis, amelynek nagy része F4 állapotú beszédet tartalmaz (76.4%) és a szlovén BN adatbázis, amelynek igen kis részét teszi ki az F4 állapotú beszéd (8.1%) [3]. Messzemenő következtetéseket nyilván nem tudunk az előbb említett számok alapján levonni, de valószínűsíthető, hogy a portugál adatbázisban nagyobb arányban voltak jelen a kereskedelmi adóktól átvett híranyag a közszolgálati adóhoz képest, míg a szlovén adatbázis esetében ez az arány fordítva volt jelen.

A *4. táblázatban* látható, hogy a BN adatbázisokban milyen arányban jelentek meg férfi, illetve női beszélők.

4. táblázat
A beszélők eloszlása a BN adatbázisokban nemek szerint

	Női beszélők száma	*Női beszéd időtartama*	*Férfi beszélők száma*	*Férfi beszéd időtartama*
Belga BN	27	0:39:48	88	1:59:46
Cseh BN	128	1:05:51	285	1:45:06
Gall BN	34	1:39:47	100	1:20:30
Görög BN	42	0:42:01	111	1:32:04
Horvát BN	69	1:05:58	140	1:47:03
Magyar BN	39	1:15:28	102	1:35:04
Portugál BN	39	0:44:53	133	2:31:38
Szlovén BN	55	1:02:27	139	1:30:58
Szlovén 2 BN	28	0:42:01	69	1:24:14
Szlovák BN	29	1:08:39	95	1:51:22
összesen:	**490**	**10:06:53**	**1262**	**17:17:45**

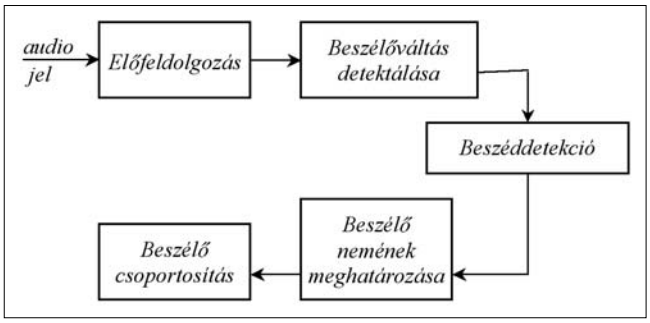
A táblázatban szereplő beszélők száma az összes olyan beszélőt takarja, akinek a hangja elhangzott a híradás során. Látható, hogy jóval nagyobb számban szerepelnek a híradásokban a férfi beszélők a női beszélőkhöz képest. Valószínűsíthető, hogy a televíziós társaságok a nagyobb hitelesség reményében inkább férfiakat bíznak meg a műsorvezetéssel, riportkészítéssel stb. Ugyanakkor látható az is, hogy annak ellenére, hogy általában jóval kevesebb a női beszélő a híradásokban, mégis az egy főre jutó beszélt percek száma a nőknél nagyobb, mint a férfiaknál. Egy női beszélő átlagosan 1 percet és 13 másodpercet beszél, míg egy férfi beszélőre jutó beszédidő 49 másodperc. Természetesen ez a szám adatbázisonként változik, például a gall BN adatbázis esetén az egy női beszélőre eső percek száma majdnem 3, míg a férfi beszélők csupán 1 percet és 24 másodpercet beszéltek, azaz feleannyit. A legkiegyenlítettebb arány talán a portugál adatbázisban fedezhető fel, ahol egy női beszélő 69 másodperc beszédidővel, míg egy férfi 68,57 másodperc beszédidővel rendelkezik.

Mindezek a különbségek, eltérések az adatbázisok között, sőt még a magyar adatbázison belüli különbségek is az adatbázis többrétűségét, újszerűségét emelik ki.

4. Szegmentálás, kiértékelés

A COST278 BN kutatócsoportjának célkitűzései között szerepelt az is, hogy a már rögzített és szabályosan átírt, címkézett adatbázist alapul véve olyan intelligens automata rendszereket fejlesszen ki, amelyek képesek néhány nyelvfüggetlen paraméter alapján feldolgozni a teljes BN adatbázist, majd ezeket a rendszereket egy egységes kiértékelő elv alapján osztályozni. A kutatócsoport a következő algoritmusok megvalósítását tűzte ki célul: beszélőváltás detektálása, beszéddetekció, beszélő nemének meghatározása, beszélők csoportosítása [3]. A kutatócsoport mindegyik tagja kifejlesztett egy vagy több olyan módszert, amely valamely, a *2. ábrában* is jelölt feladat elvégzésére alkalmas volt.

2. ábra
Az audio jel feldolgozásának egyszerűsített blokkvázlata



A Beszédakusztikai Kutatólaboratórium egyik fő célkitűzése a magyar nyelvű híranyag-adatbázis létrehozása és a fentebb említett algoritmusok közül a beszéddetektáló algoritmus implementálása volt.

A különböző algoritmusok különböző jellegű eredményeket produkálnak, ezért szükség volt egy közös kiértékelő szoftverre, mellyel az eredmények összehasonlíthatóságát biztosították. Ezt a szoftveres eszközt a BN kutatócsoport minden tagja használta és a portugál partner bocsátotta közre.

5. Beszéddetekció

A beszéddetektáló algoritmus *(speech-non-speech classification – SNC)* feladata az, hogy detektálja a legalább 1,5 másodperces beszédszünetet [3], tehát a rögzített anyag olyan részeit, ahol nincs beszéd, de előfordulhat háttérzaj vagy egyéb zaj, zene vagy egyéb hanghatás.

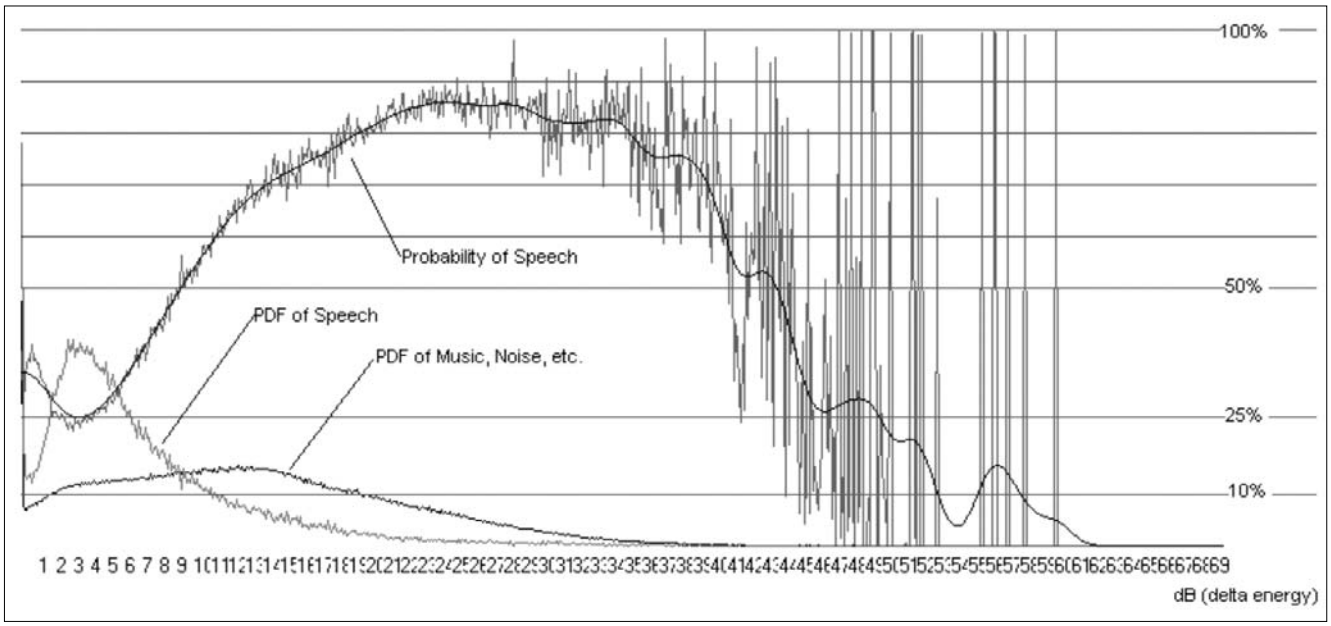
A laboratóriumunkban kifejlesztett algoritmus két különböző, ideiglenes döntésre alapozva hozza meg a végső döntést arról az akusztikai szegmensről, amit a bemeneten talál. Az első ideiglenes döntést egy statisztikai alapú (HMM) fonémafelismerő* kimenete alapján hozza. Kimenetnek a fonéma-bigram valószínűségeket tekintjük. Ezek a valószínűségek jellemzően másmás értékeket vesznek fel attól függően, hogy a bemeneten beszéd vagy egyéb akusztikai jelenség (zene, zaj stb.) található. A felismerő betanításához bármely, a BN kutatócsoportjában szereplő, nyelven elhangzott hanganyagot használhatunk. A betanítás során két különböző fonéma-bigram mátrix készül, egy a beszédre, egy pedig a „nem-beszédre”, azaz zene, zaj stb. A döntés a felismerés során születik meg a modell tranziensei és a beszéd vagy „nem-beszéd” fonéma-bigram mátrixok közötti távolság függvényében.

A második ideiglenes döntés a beszéd, illetve a „nem-beszéd” szegmens energiaváltozásának statisztikai analízisének eredménye alapján jön létre. A mért energiaváltozást valószínűségi változóként kezelve, egy valószínűség-sűrűség függvényt *(probability density function – PDF)* számolunk a beszédre és a „nembeszédre” is. A PDF által adott eredmények alapján a beszéd valószínűségét határozzuk meg *(3. ábra)*. Csak abban az esetben fog a rendszer beszédet detektálni, amennyiben mindkét algoritmus beszédet detektál. Minden egyéb esetben „nem-beszédet” fog detektálni.

A módszer kiértékelésekor azon szegmensek százalékos aránya dönt, amelyekre helyes döntést hozott a rendszer. Ez a százalékos arány a rendszer pontosságát fejezi ki *(„accuracy”)* [3].

Ezt a feladatot a COST278 BN kutatócsoportjának hét tagintézménye végezte el: a Ghent-i Egyetem –

** A BME Beszédakusztikai Laboratóriumban kifejlesztett folyamatos beszédfelismerő (MKBF 1.0) optimális működését az akusztikai, fonetikai [10] és nyelvi modellek változtatásával állítottuk be. A felvételek mindegyike – mind a betanításnál, mind a tesztelésnél – 16 kHz-en mintavételezett, 16 biten lineárisan kvantált jel, amely a megfelelő előfeldolgozás után kerül felismerésre. A fonémaszintű felismerőnk 16 kHz mintavételezésű, 17 Bark frekvenciatérbeli derivált, + 17 időbeni derivált, + 17 időbeni második derivált, + energia bemeneti jelvektor mellett, 4-5 állapotú kvázi-folytonos, 24 lépcsős, rejtett Markov-modellekkel (QCHMM), fonéma alappal dolgozik. Az akusztikai, fonetikai szint optimalizálásáról további információk [11]-ben találhatók.*



3. ábra
Valószínűség-sűrűség függvények beszéd, „nem-beszéd” (zene, zaj, egyéb) esetre és annak valószínűsége, hogy beszéd hangzott el

Belgium (ELIS), a Ljubljana-i Egyetem – Szlovénia (ULJ), a Maribor-i Egyetem – Szlovénia (UMB), a Liberec-i Műszaki Egyetem – Cseh Köztársaság (TUL), INESC ID – Portugália (INESC), a Vigo-i Egyetem – Spanyolország (UVIGO), és a Budapesti Műszaki és Gazdaságtudományi Egyetem (BUTE).

A kiértékelés eredményei *(4. ábra)* világosan mutatják az algoritmus létjogosultságát, hiszen a feladat megoldása során módszerünk, a többi módszer eredményeivel összevetve, kiemelkedő eredményeket prezentál. A probléma megoldása nem triviális, hiszen előfordulhat, hogy az adatbázisba reklám is belekerül, ami tartalmazhat beszédet is (megjegyzés: a reklámok a BN adatbázisokban nem kerültek átírásra). Az általunk bemutatott módszer sikeresen alkalmazható „nem-beszéd” (zene, zaj stb.) szegmensek detektálására, címkézésére. Az eredményeket a grafikonon prezentáljuk.

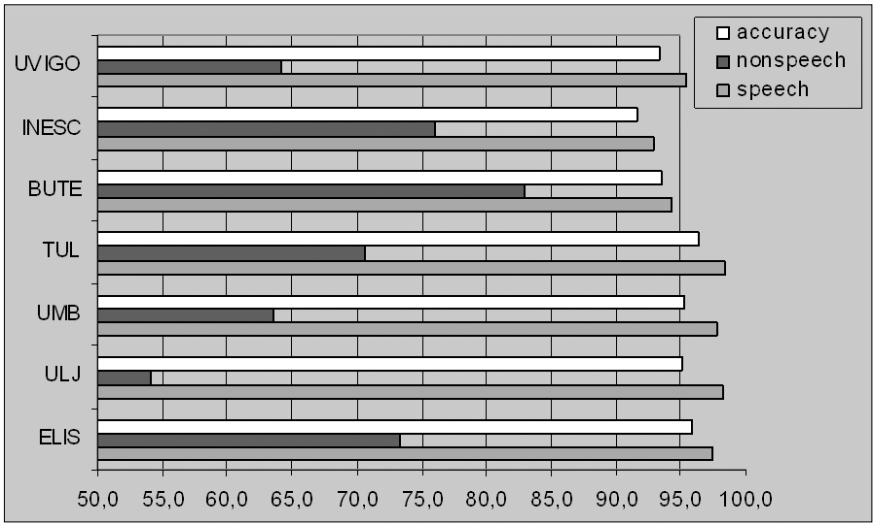
Az ábrán látható, hogy a fentebb bemutatott módszer a COST278 munkacsoport többi tagja által kifejlesztett módszerhez viszonyítva a beszédet majdnem 95% arányban osztályozza beszédnek, ami egy közepes eredmény ebben a kontextusban. Ugyanakkor látható az is, hogy a módszer igazi erőssége abban rejlik, hogy a nem-beszéd eseményt csaknem 85% arányban sorolja a nem-beszéd kategóriába, ami ebben a kontextusban egy kiváló eredmény, hiszen csak a portugál partner tudott olyan algoritmust kifejleszteni, amely 75% feletti arányban teszi ugyanezt. Tehát kimondható az, hogy erre a feladatra a legalkalmasabb módszer az általunk bemutatott módszer.

6 . Összefoglalás

E tanulmányban a szerzők bemutattak egy újszerű, multimédiás adatbázist, annak struktúráját, az adatbázison végzett statisztikai vizsgálatok eredményeit és egy a „nem-beszéd” detektálásához és annak címkézéséhez alkalmazható algoritmust, melyet a BME Beszédakusztikai Kutatólaboratóriumában fejlesztettünk ki. Az eredmények alapján kimondható az, hogy ez a módszer megbízhatóan és megfelelően nagy pontossággal ismeri fel a „nem-beszéd szegmenseket”.

Ugyanakkor Laboratóriumunk nem tekinti lezártnak a kutatást ezen a területen, hiszen még nagyon sok kiaknázatlan területe van még. Például, az adatbázis kiválóan alkalmazható automata reklámszámláló szoftverek betanításához, teszteléséhez.

4. ábra
A különböző beszéddetektálási módszerek eredményei



Irodalom

[1] Becchetti C., Ricotti L.P.,
'Speech Recognition, Theory and C++ implementation'
Fondazione Ugo Bordoni, Rome, (1999)
ISBN 0-471-97730-6.

[2] D. Graff, Z. Wu, R. MacIntyre, M. Liberman,
'The 1996 broadcast news speech and language-model corpus'.
In: Proceedings of the 1997 DARPA Speech Recognition Workshop, February 1997.
Chantilly, Virginia.

[3] A. Vandecatseye, J. Martens, J. Neto, H. Meinedo, C. Mateo, J. Dieguez, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A. Cizmar, H. Papageorgiou, C. Alexandris,
'The COST278 – pan-European Broadcast News Database',
In: Proceedings of LREC 04,
Lisboa, Portugal (2004)

[4] C. Barras, E. Geoffrois, Z. Wu, M. Libermann,
'Transcriber : Development and use of a tool for assisting speech corpora production',
Speech Communication,
Volume 33, Issues 1-2., pp.5–22. (2001)

[5] J. Zibert, F. Mihelic, J. Martens, J. Neto, H. Meinedo, J. Neto, L. Docio, C. Mateo, P. David, J. Nouza, M. Pleva, A. Cizmar, A. Zgank, Z. Kacic, Cs. Teleki, K. Vicsi,
'The COST278 Broadcast News segmentation and speaker clustering evaluation – overview, methodology, systems, results',
INTERSPEECH 2005,
Lisboa, Portugal (2005)

[6] Siegler, M. A., Jain, U., Raj, B., Stern, R. M.,
'Automatic segmentation, classification and clustering of broadcast news',
In: Proceedings of DARPA Speech Recognition Workshop, Chantilly VA, pp.97–99. (1999)

[7] K. Vicsi, Cs. Teleki, Sz. Velkei,
'Development and evaluation of a Hungarian Broadcast News database',
In: Proceedings of Forum Acousticum 2005,
Budapest, Magyarország (2005)

[8] Perez-Freire, L., Garcia-Mateo C.,
'A multimedia approach for audio segmentation in TV Broadcast News',
In: Proceedings ICASSP (2004)

[9] Vandecatseye A., Martens, J.P.,
'A fast, accurate and stream-based speaker segmentation and clustering algorithm'.
In: Proceedings Eurospeech (2003)

[10] Deng Y., Mahajan M., Acero A.,
'Estimating Speech Recognition Error Rate without Acoustic Test Data'
Elérhető: http://research.microsoft.com/srg/papers/2003-milindm-eurospeech.pdf

[11] Sz. Velkei, K. Vicsi,
'Beszédfelismerő modellépítési kísérletek akusztikai, fonetikai szinten, kórházi leletező beszédfelismerő kifejlesztése céljából',
MSZNY 2004, Szeged, Magyarország (2004)

[12] Roach, P., S. Arnfield, W., Barry, J., Baltova, M., Boldea, A., Fourcin, W., Gonet, R., Gubrynowicz, E., Hallum, L., Lamel, K., Marasek, A., Marchal, E., Meister, E., Vicsi, K.,
'BABEL:
An Eastern European Multi-language database'.
International Conference on Speech and Language Processing 1996, Philadelphia.

[13] Vicsi, K., Valyon, Z., Gordos, G., Csirik, J., Kocsor, A., Tóth, L.,
'MTBA – Magyar nyelvű telefonbeszéd adatbázis'.
Technical report. IKTA 3 project,
a.sz.: 11025888, (2000)
http://alpha.ttt.bme.hu/speech/hdbMTBA.php
„György Békésy" Acoustics Research Laboratory of the Budapest University of Technology and Economics (2002).

[14] Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László,
'Beszédadatbázis irodai számítógép-felhasználói környezetben',
II. Magyar Számítógépes Nyelvészet Konferencia, (2004)