BEKE
ANDRÁS

BESZÉD • KUTATÁS • ALKALMAZÁS

GÉPI BESZÉLŐDETEKTÁLÁS MAGYAR NYELVŰ SPONTÁN TÁRSALGÁSOKBAN

ELTE EÖTVÖS KIADÓ

Beke András

GÉPI BESZÉLŐDETEKTÁLÁS MAGYAR NYELVŰ SPONTÁN TÁRSALGÁSOKBAN

Beszéd • Kutatás • Alkalmazás

Beke András

GÉPI BESZÉLŐDETEKTÁLÁS MAGYAR NYELVŰ SPONTÁN TÁRSALGÁSOKBAN

ELTE EÖTVÖS KIADÓ

Budapest, 2015

A kötet megjelenését az OTKA PUB-K 114596 ny. pályázat támogatta.

Lektorálták:
Adamikné Jászó Anna
Gósy Mária
Olaszy Gábor

© Beke András, 2015

ISBN 978-963-312-234-1
ISSN 2064-4442

ELTE EÖTVÖS KIADÓ www.eotvoskiado.hu

Felelős kiadó: Hunyady András ügyvezető igazgató
Felelős szerkesztő: Pál Dániel Levente
Nyomdai munkák: Multiszolg Bt.
Tördelés: Windor Bt.
Borítóterv: Csele Kmotrik Ildikó

Tartalom

Sorozatszerkesztői előszó 9

Előszó 11

1. Bevezetés 13

2. A beszélődetektáló általános felépítése 23
2.1. Akusztikai jellemzők a beszélődetektáláshoz 25
2.2. Beszélőszegmentálás 27
2.2.1. Metrikus alapú szegmentáló algoritmusok 28
2.2.1.1. Bayes-féle információs kritérium (BIC: Bayesian Information Criterion) 28
2.2.1.2. Általánosított valószínűségarány (GLR: Generalized Likelihood Ratio) 31
2.2.1.3. Gish-távolság (Gish-distance) 32
2.2.1.4. Kullback–Leibler-távolság (KL vagy KL2) 32
2.2.1.5. Más távolságmérési eljárások 33
2.2.2. Nem metrikán alapuló szegmentálók 33
2.2.2.1. Szünetalapú beszélőszegmentáló 33
2.2.2.2. Modellalapú szegmentáló 34
2.2.3. A beszélőszegmentáló algoritmusok összegzése 35
2.3. Beszélőklaszterezés 35
2.3.1. Hierarchikus klaszterezési technikák 36
2.3.1.1. Alulról felfelé (egyesítő, bottom-up) klaszterező eljárások 37
2.3.1.2. Fentről lefelé (lebontó, top-down) klaszterező technikák 40
2.4. Beszéddetektálás 40
2.4.1. A beszéddetektáló általános leírása 41
2.4.2. Jellemzőkinyerés a beszéddetektáló megvalósításához 41
2.4.3. A beszéddetektáló döntési modulja 42
2.4.4. A beszéddetektáló utófeldolgozása (simítás) 42
2.5. Beszélőspecifikus jellemzők a gépi beszélőfelismerésben 43
2.5.1. Kevert Gauss-beszélőmodell 48
2.5.1.1. Kevert Gauss-modell 48
2.5.1.2. Univerzális háttérmodell 50
2.5.1.3. A beszélőegyezés mérése 51
2.6. Az egyszerre beszélés detektálása 52

3. A kutatás célja, kutatási kérdések és hipotézisek 57
3.1. Kutatási kérdések 57
3.2. A kutatás célja 57
3.3. A kutatás hipotézisei 58

4. Kísérleti személyek, általános anyag és módszer 59
4.1. Anyag és kísérleti személyek 59
4.1.1. A beszélődetektáló kiértékeléséhez használt korpusz 60
4.1.2. A beszélőspecifikus jellemzők kialakításához használt korpusz 60
4.1.3. A beszéddetektálóhoz használt korpusz 61
4.1.4. Az egyszerrebeszélés-detektálóhoz használt korpusz 61
4.2. Kiértékelési módszer 61
4.2.1. Beszélődetektálási hibaarány (DER: Diarization Error Rate) 62
4.2.2. További kiértékelési technikák (DET: Detection Error Tradeoff) 64

5. Beszélődetektálás társalgásokban 67
5.1. A korpusz általános statisztikai jellemzői 67
5.2. A beszélődetektáló felépítése 69
5.2.1. Beszélőszegmentálás 69
5.2.1.1. Jellemzőkinyerés a beszélőszegmentáláshoz 69
5.2.1.2. Bayes-féle információs kritérium (BIC: Bayesian Information Criterion) 69
5.2.1.2.1. Növekedő ablakhosszmetódus a ΔBIC számításához 71
5.2.1.2.2. A BIC paraméterei 71
5.2.1.3. Téves riasztások csökkentése (False Alarm Compensation) 72
5.2.1.3.1. A KL2-alapú utófeldolgozás beállításai 72
5.2.2. Beszélőklaszterezés 73
5.2.2.1. Jellemzőkinyerés a beszélőklaszterezéshez 73
5.2.2.2. GMM-szupervektor 73
5.2.2.3. BIC-alapú klaszterezés 74
5.3. A beszéddetektálás felépítése 75
5.3.1. Jellemzőkinyerés 76
5.3.2. A beszéddetektáló döntési metódusa 76
5.3.3. A beszéddetektáló utófeldolgozása 77
5.3.4. Az általunk javasolt eljárás a küszöb meghatározására 78
5.3.5. A beszéddetektáló kiértékelése 79
5.4. Az egyszerre beszélések automatikus osztályozása spontán magyar társalgásokban 79
5.4.1. Jellemzőkinyerés 80
5.4.2. Lényegkiemelés 82
5.4.2.1. Korlátozott Boltzmann-gép 82

5.4.2.2. Az RBM előtanítási paraméterei 84
5.4.3. Osztályozás 85
5.4.3.1. Szupport vektor gép (SVM: Support Vector Machine) 85
5.4.3.1.1. Az SVM tanítási paraméterei 88

6. Eredmények 89
6.1. A beszélőszegmentálás eredménye az alapbeállítások mellett 89
6.2. A BIC beszélődetektáló beszélőspecifikus akusztikai jellemzővel 90
6.2.1. Beszélőspecifikus jellemzők 90
6.2.2. A beszélőspecifikus jellemzők implementálása a beszélődetektálóba 92
6.3. A BIC λ paraméterének optimális megválasztása 93
6.4. A beszéddetektálás implementálása 93
6.4.1. A beszéddetektáló eredményei spontán társalgásban 93
6.4.2. A beszéddetektáló implementációja a beszélődetektálóba 95
6.5. Az egyszerrebeszélés-detektáló eredménye 96
6.5.1. Az egyszerrebeszélés-detektáló eredménye spontán társalgásban 96
6.5.2. Az egyszerrebeszélés-detektáló implementációja a beszélődetektálóba ... 98

7. Következtetések 101
7.1. Beszéddetektáló 101
7.2. Beszélőspecifikus jellemzők a gépi beszélőfelismerésen keresztül 102
7.3. Az egyszerre beszélések automatikus osztályozása spontán magyar társalgásokban 103
7.4. Beszélődetektálás 104
7.4.1. A beszélődetektáló alaprendszere 105
7.4.2. Beszélőspecifikus akusztikai jellemzők implementálása 105
7.4.3. A BIC λ paraméterének beállítása 106
7.4.4. A beszéddetektálás implementálása 106
7.4.5. Az egyszerrebeszélés-detektáló implementálása 106
7.4.6. A kifejlesztett rendszer végső eredménye 106

8. Összegzés 107

9. Kitekintés 109

10. Irodalom 111

Automatic speaker diarization in Hungarian spontaneous conversations 131

Sorozatszerkesztői előszó

A Beszéd • Kutatások • Alkalmazások sorozat ötödik kötete jól mutatja, hogy a beszéd kutatása interdiszciplináris terület. A jelen munka szervesen építi össze a nyelvészeti (elsősorban fonetikai és pragmatikai) és a műszaki tudományok vonatkozó ismereteit, módszertanát.

Beke András magyar nyelvű spontán társalgásokban végez gépi beszélődetektálást – a kötet ennek a folyamatnak a lépéseit és aktuális eredményeit mutatja be. A kutatás és a téma jellegzetességeiből adódóan egy folyamatosan végzett munkáról és egy állandóan változó tudományterületről kapunk pillanatképet. Mind a technológiai eszközök, mind a témában végzett kutatások olyan dinamikusan változnak, hogy a kötet szükségképpen a fejlődésnek csak egy adott időszeletét rögzítheti. Arra azonban így is lehetőséget ad, hogy bepillantsunk ennek a beszéddel foglalkozó tudományterületnek a kérdésfelvetéseibe, éppen megoldásra váró problémáiba, kutatási módszereibe, az eredmények értékelésének lehetőségeibe.

A spontán beszéd legtipikusabb megjelenési formája a társalgás, amely ennek ellenére korábban kevéssé volt tárgya fonetikai kutatásoknak. Ennek hátterében elsősorban módszertani okok álltak – a spontaneitás és a jó minőségben rögzített hanganyag paradoxonából adódóan. A kétezres évek azonban áttörést hoztak ezen a területen, a spontánbeszéd-korpuszok létrehozásában egyre inkább figyelembe vették és a protokoll részévé tették ezt a beszédhelyzetet és -módot is.

A jelen kutatás módszertana meglévő algoritmusok implementálásából és finomhangolásából állt elő, a szerző kísérleti úton optimalizálta a komplex módszert, és ezzel sikerült javítania a beszélődetektálás hatásfokát. Jelentős a kutatás abból a szempontból is, hogy magyar nyelvű társalgások automatikus feldolgozása a célja. Az elemzett anyag mérete is számottevő: 100 hárombeszélős társalgás, azaz mintegy 55 órányi hanganyag képezte a kutatás alapját.

A társalgásokban a gépi beszélődetektálás jelentősége egyrészt az elemzési munka gyorsítása és segítése az automatikus módszerek bevonásával. Másrészt a társalgások gépi feldolgozása információt szolgáltat a társalgások felépítéséről, a beszélői szerepekről is. Mindez hosszabb távon – akár összehasonlítási alapként – hozzájárulhat az ember-gép kommunikáció sajátosságainak megismeréséhez, az ilyen jellegű új módszertanok létrehozásához.

A téma szakértői mellett ajánljuk a könyvet a nyelvészet és a mérnöki tudományok határterületei iránt érdeklődőknek, beszédtudománnyal foglalkozó egyetemistáknak és doktoranduszoknak.

Markó Alexandra

Előszó

Az emberi kommunikáció egyik leggyakrabban használt eszköze a nyelv. A nyelv hangzó változata, a beszéd a nyelvi kommunikáció legtermészetesebb és legtöbbet használt formája (Gósy 2005). A mindennapi életben a beszélt nyelvi kommunikáció a legtöbb esetben társas interakcióban jelenik meg, mint amilyen a társalgás. A beszédet akusztikai szempontból elemző kutatások elsőként szófelolvasásokon alapultak, majd szövegfelolvasásokon. Az utóbbi évtizedben azonban egyre nagyobb figyelem összpontosul a spontán beszéd vizsgálatára, azon belül a társalgás elemzésére. Számos tudományág (diskurzuselemzés, pszicholingvisztika, fonetika, beszédtechnológia stb.) foglalkozik a társalgás felépítésével, szabályaival, modellezésével. A konverzációelemzés eredményeiből tudjuk, hogy a társalgás nem rendezetlen struktúra, hanem szabályok mentén rendeződik, dinamikusan alakul a beszédpartnerek mentén (Iványi 2001). A konverzációelemzés által feltárt szabályosságokra támaszkodva a beszédtechnológiában is megindultak a vizsgálatok a társalgások gépi modellezésére. A beszédtechnológián belül az erre irányuló kutatási terület a gépi beszélődetektálás (speaker diarization) (beszélődetektálás alatt a jelen munkában mindig a gépi beszélődetektálást értjük, nem a humán percepción alapulót). A beszélődetektálás feladata, hogy a társalgásokban automatikusan jelölje, hogy mikor ki beszél. Ennek során a folyamatos társalgások automatikusan lejegyzett szövegeit újrastrukturáljuk (az elhangzott közléseket személyekhez rendeljük), így a szöveg sokkal könnyebben feldolgozható más, például tartalomkinyerő algoritmusok számára.

A jelen értekezés célja, hogy első ízben hozzon létre magyar spontán társalgásokra működő beszélődetektáló rendszert. A kutatás fő motivációja az volt, hogy spontán társalgásokra valósítsunk meg beszélődetektálót, mivel az eddigi beszélődetektálók híradós adásokra vagy telefonhívásokra készültek. A beszélődetektálás megvalósítása igen nehéz feladat mind a híradós felvételekre, mind a telefonos hívásokra. A legnagyobb kihívást azonban a spontán társalgások beszélőkre bontása jelenti. A dolgozat célkitűzése egyrészt az, hogy a beszélődetektáláshoz kapcsolódó tudományterületeket bemutassa, illetve hogy maga a beszélődetektálás főbb módszertani ismereteit leírja. A másik célja az, hogy a beszélődetektálóhoz szükséges algoritmusokat elkészítse (egyszerrebeszélés-detektálás, beszélőszegmentáló, beszélőklaszterező) és a már létező algoritmusokat implementálja a beszélődetektálóba (beszéddetektáló, beszélőfelismerő algoritmus).

Az általunk javasolt rendszer célja, hogy magyar nyelvű spontán társalgásokban automatikusan detektálja a beszélőváltásokat pusztán akusztikai információk alapján, vagyis megoldást adjon arra a kérdésre, hogy „mikor ki beszél?”. Az algoritmus kialakításához a BEA (BEszélt nyelvi Adatbázis; Gósy 2012) spontán társalgásait használtuk fel, amelyekben három résztvevő társalog. Az általunk javasolt beszélődetektáló rendszer lényegében nem felügyelt tanulási eljárásokon alapul.

Az értekezés 10 fejezetből áll. Az *első, bevezető fejezet*ben általános leírást adunk az automatikus beszélődetektálásról, helyéről a beszédtechnológiában, illetve a beszédtudományban. A *2. fejezet* módszertani áttekintést ad a beszélődetektálásban használt algoritmusokról. Itt kerül bemutatásra a beszéd/nembeszéd detektálásának folyamata, amelynek célja, hogy folyamatos akusztikai jelben jelölje, hogy hol van beszédrész, illetve nem beszédrész. Ez a fejezet ismerteti a beszélőfelismerés alapvető módszertani kérdéseit, valamint ebben a fejezetben kap helyet az egyszerre beszélések automatikus osztályozása is, amelynek igen nagy szerepe van a beszélődetektálás téves riasztásainak csökkentésében.

Saját kutatásunk céljainak, kérdéseinek és hipotézisének ismertetése a *3. fejezet*ben történik.

A *4. fejezet*ben a kísérleti személyek, az általános anyag és a módszer ismertetése történik. Itt mutatjuk be a kísérletekhez használt adatbázis felépítését, tartalmát, illetve itt kerül bemutatásra a beszélődetektálás kiértékeléséhez használt DER (Detection Error Rate) eljárásának és az osztályozásának kiértékeléséhez használt DET (Detection Error Tradeoff) algoritmus.

Az *5. fejezet*ben mutatjuk be az általunk felépített beszélődetektálót (részben már létező algoritmusokat, illetve a jelen munkában fejlesztett algoritmusokat). Ez a fejezet négy alfejezetet tartalmaz. Elsőként az általunk használt beszélődetektáló lépéseit írjuk le, majd a beszéd/nembeszéd detektálóét, és végül az egyszerrebeszélés-detektálóét.

A *6. fejezet*ben a kísérletek és az eredmények ismertetésére kerül sor. Elsőként a beszélődetektáló alapbeállításaival elért eredményeket mutatjuk be. Ezután vizsgáljuk, hogy az általunk javasolt beszélőspecifikus akusztikai jellemzőkkel milyen mértékű javulást lehet elérni a beszélődetektálásban. A harmadik vizsgálatban a beszéddetektáló implementálásának hatását vizsgáljuk a beszélődetektáló eredményeire. Az utolsó kísérletben az egyszerrebeszélés-detektáló rendszert mutatjuk be, illetve annak implementálásának eredményét a beszélődetektálóba.

A *7. fejezet* az általános következtetéseket tartalmazza, amelyet az általános összefoglalás követ *(8. fejezet)*. Ezután ismertetjük a beszélődetektálás felhasználási és további fejlesztési lehetőségeit *(9. fejezet)*. Ezt követi az Irodalom *(10. fejezet)*.

A beszélődetektálás nagyon fontos szerepet játszik a társalgások elemzésében, hiszen igen sok tartalom a beszélőváltások szerint strukturálható, amelyek nyelvészeti és metanyelvészeti információkat is tartalmazhatnak (domináns beszélő, szerepek a társalgásban, az interakció szintjei, érzelmek).

A kötet eredményei közelebb vihetik az olvasót az ember-ember kommunikáció megértéséhez, modellezéséhez, amely tovább mutat a mesterséges intelligencia, az ember-gép kommunikációja felé.

Ezúton szeretném kifejezni hálás köszönetemet témavezetőmnek, Gósy Máriának, valamint a jelen munka másik két lektorának, Adamikné Jászó Annának és Olaszy Gábornak, a hasznos megjegyzéseikért. Köszönettel tartozom Szaszák Györgynek szakmai és baráti támogatásáért. Külön köszönettel tartozom Markó Alexandrának, aki fáradhatatlan szerkesztői munkával lehetővé tette, hogy ez a könyv megjelenhessen.

A könyv az OTKA PUB-K 114596 ny. pályázat támogatásával jelenhetett meg.

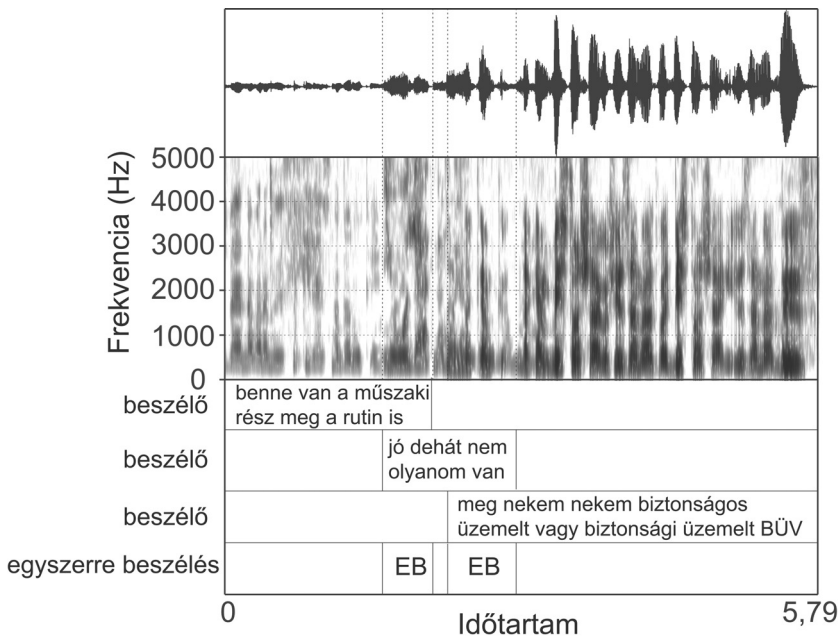
Beke András

1. Bevezetés

A kommunikáció alapvető feltétele a résztvevők megléte, azaz a feladó (forrás) és a címzett (vevő). A feladó az, aki különböző nyelvi és nemnyelvi jelek segítségével üzenetet küld a címzettnek (kódolja), aki ezt az üzenetet felfogja, értelmezi (dekódolja) és válaszol rá. A résztvevők szerepet cserélhetnek (az előző esetben a címzett válik feladóvá), illetve többen is részt vehetnek a kommunikációban. Az üzenetet kifejező összefüggő jeleket kódnak nevezzük. Használunk nyelvi és nemnyelvi kódokat. A kommunikáció csak akkor lesz sikeres, ha a résztvevők közös nyelvet beszélnek, azaz mindketten ismerik a kódot. A megfogalmazott üzenet a csatornán keresztül jut el a feladótól a címzettig, az továbbítja a közleményt. A csatorna lehet hallható (telefonbeszélgetés), látható (levél) vagy egyszerre többféle is (személyes beszélgetés). A tipikusnak mondható verbális kommunikációt mindig nonverbális elemek kísérik, amelyek természetesen csak akkor érvényesülnek, ha a kapcsolat nemcsak auditív, hanem vizuális formában is fennáll (vagyis nemcsak hallják, hanem látják is egymást a felek). Ilyen a testtartás, a prozódia, a mimika, a gesztikulálás stb. A beszédkommunikációban zajnak nevezzük azokat a tényezőket, amelyek megzavarják, torzítják az üzenetet, gátolják annak eljutását a címzetthez (például ha recseg a telefon).

A társalgás az 1960-as években került a középpontba, elsősorban a szociológiai érdekeltségű társas nyelvészet (Kiss 1995), a szociálpszichológia, a pszicholingvisztika, a modern filozófia és a logika együttműködéseként (Pléh 2012). A társalgásokkal elsősorban a diskurzuselemzés, illetve a konverzációelemzés foglalkozik (például Iványi 2001; Jakusné; Hámori 2006; Boronkai 2008, 2009).

A konverzációanalízis (conversation analysis) néven megjelent tudományág a hétköznapi társalgások verbális interakcióinak a szerkezetét vizsgálja, amely bizonyos szerkezeti szabályosságokat feltételez a társalgások felépítésében (Garfinkel 1967; Goffman 1983; Schegloff 1992; Sacks et al. 1974; Sacks 1992; Iványi 2001; Stokoe 2006). Fő elgondolásuk, hogy a beszélgetésnek interaktív, szekvenciális felépítése van, amelyben a beszélők váltják egymást. Ebben a keretben értelmezhetővé váltak olyan beszédelemek, amelyeket addig a rendszernyelvészet le nem írhatóknak jegyzett, mint például a megakadások, szünetek stb. Mindezen jelenségeket a konverzációelemzés a „beszélt nyelv szintaxisának" nevezi (Iványi 2001) *(1.1. ábra).*



1.1. ábra

A társalgás felépítésének szemléltetése

A konverzációelemzés adta keretben a társalgásnak belső struktúrát tulajdonítanak, amely nemcsak nyelvészeti szempontból fontos, hanem beszédtechnológiai szempontból is, hiszen ha a társalgás rendszerszerű, akkor feltételezhetően gépi úton modellezhető. A beszédtechnológia a mesterséges intelligencián belül a beszédalapú (verbális) gyakorlati alkalmazások kifejlesztésével és létrehozásával foglalkozik (Németh–Olaszy 2010: 209). Az ember-gép verbális kommunikációban számos részfeladatot modelleztek már magyar nyelven, mint a beszéd gépi megértését (beszédfelismerés), illetve a gépi beszéd-előállítást (beszédszintézis), a beszélő személy gépi azonosítását a hangja alapján (beszélőfelismerés). Ezek a részfolyamatok a társalgásban kapcsolódnak össze, ahol nem pusztán egyoldalú a folyamat, vagyis nemcsak beszédfelismerésről vagy beszéd-előállításról beszélhetünk, hanem ezek körkörös működéséről, ami a beszélők váltakozásából fakad, vagyis fontos lépés, hogy ezt a folyamatot gépileg tudjuk lekövetni, előjelezni (Jin et al. 2004).

Az elmúlt évtizedekben számos tudományos közösség figyelt fel a beszélődetektálás fontosságára, mint az amerikai Nemzeti Szabványügyi és Technológiai Intézet, NIST (National Institute of Standards and Technology, http://www.itl.nist.gov/iad/mig/tests/rt/). A beszélődetektálás fejlődését mindig valamilyen valós igény határozta meg. Az 1990-as évek végén és a 2000-es évek elején a korai munkákban a telefonos beszélgetések és a híradások voltak a kutatások középpontjában, amelyekben a beszélődetektálást a műsorok automatikus lejegyzéséhez használták fel. 2002-től nőtt az érdeklődés az élő, spontán társalgások iránt (meeting domain),

amelyek körül számos projekt jött létre, mint a European Union (EU) Multimodal Meeting Manager (M4) projekt (http://spandh.dcs.shef.ac.uk/projects/m4/index.html), a Swiss Interactive Multimodal Information Management (IM2) projekt (http://www.im2.ch/), az EU Augmented Multi-party Interaction (AMI) projekt (http://www.amiproject.org/), ezt követően folytatódott az EU Augmented Multi-party Interaction projekt a Distant Access (AMIDA) projekttel közösen (http://www.amiproject.org/), és végül az EU Computers in the Human Interaction Loop (CHIL) projekt (http://chil.server.de/). Ezen projektekben a multimodális technológiák kutatási és fejlesztési eredményeinek célja az volt, hogy elősegítsék az ember-ember kommunikációt azzal, hogy az automatikusan kivonatolt társalgás szövegét archiválni tudják, illetve elérhetővé tegyék a társalgó felek számára. A beszélődetektálás implementálható a multimodális rendszerekbe, amelyben fontos szerepet kap mind a tartalmi indexelés, tartalmi kivonatolás, mind a verbális és a nemverbális emberi kommunikációs eszközök archiválása (a testtartás, az érzelmek, a másokkal folytatott interakciók stb.). A multimodális technológia fejlesztéséhez olyan korpuszokat hoztak létre, amelyek egyszerre tartalmaznak audio-, videojelet és szöveges tartalmat. Mindezekből olyan információkat nyerhetnek ki, amelyek segítségével a társalgások tartalma automatikusan strukturálható, elemezhető (Ajmera–Wooters 2003; Barras et al. 2004; Wooters et al. 2004).

A társalgás alapegysége a beszédforduló (a terminus szinonimái: *beszédlépés,* illetve angol megfelelője, a *turn*). A beszédforduló során a társalgás egyik résztvevője beszél, amíg át nem adja, vagy amíg át nem veszik tőle a beszéd jogát: szóátadás (turn yielding), szóátvétel (turn-taking) (Sacks et al. 1974).

A beszélőváltás mechanizmusának leírásával a diskurzuselemzés, illetve a konverzációelemzés foglalkozik (például Brown–Yule 1989; Iványi 2001; Markó–Dér 2011). A beszédforduló lehet egyetlen mondat, egy frázis, vagy lehetnek különböző lexikai konstrukciók *(1.2. ábra).*

Jóllehet a beszédlépésváltás nem determinisztikus, azonban két komponense és azok szabályai befolyásolják és szabályozzák a beszélgetés struktúráját. Az egyik komponens az, hogy a társalgás résztvevői igyekeznek a szünet nélküli beszédátadásra, a másik komponens alapja, hogy a mindenkori következő potenciális beszélőváltás ideje, beszélője meghatározott.

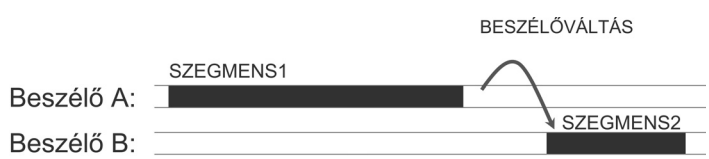
A beszédfordulók szerkezetét alapvetően meghatározza az a potenciális hely, ahol a társalgás résztvevői átvehetik a szót (átmeneti relevanciahely), vagyis alapvetően meghatározott, hogy a beszédpartnerek hogyan kövessék egymást. Ekkor az aktuális beszélő megnyilatkozása a hallgató számára lezártnak minősül, ezen a ponton a következő beszélőnek el lehet kezdenie a saját beszédlépését. A rendszer szabályai mind lokálisan lépnek érvénybe, és együttes működésüknek rekurzív jellege van: esetről esetre mindig csak két lépésegységet határoznak meg – azokat, amelyek az aktuális beszélőváltásban részt vesznek –, és átadásukat szabályozzák (Iványi 2001; Sacks et al. 1978).

A beszélőváltás legegyszerűbb formája az, amikor az aktuális beszélő a következő beszélőt külválasztással megjelöli a társalgás folytatására *(1.3.a ábra).*

A (nem is lehet) hát annak
T2 (ugyanúgy aszülés) sem mehet szerintem otthon meg meg! pont a mai világban amikor az ember már **aa**kutyájához meg a macskájához kihívja az állatorvost mikor az szül pedig őnáluk azér [azért] jóval (természetesebb)
A (igen)
T2 (ez a) folyamat mint (az embereknél)
A (meg nem egy olyan) nagy sterilitást (kíván)
T2 (igen)
A (mint) mondjuk az **or-** [orvos] a **mmm** emberi (szervezetnél)
T2 (ühm)
A a szülés és egyáltalán azon (körülmények amik)
T2 (igen)
A ott vannak akkor amikor az zajlik na most azért mondom h ez egy ehhez csak akkor lehet **megvalós-** [megvalósítani] lehetne megvalósítani ha tényleg olyan jól képzett jól gyakorlott bábák vannak na hát szülésznők!
T1 ühm
A de orvos háttérrel
T2 igen
A mindenképpen orvos háttér szükséges
T2 ühm
A tehát azt nem lehet megcsinálni hogy jaj szülök minttomén [mit tudom én] egy házat megcsinálnak szülő akárminek jó aztán akkor vagy ott van vagy nincs az orvos vagy majd jön majd hát**má-** [?] ez mondjuk kórházban is előfordult hogy jön majd a doktor úr jön majd csak nyugodtan (szüljön!)ajaj
T2 (igen) igen de ott legalább
A (na jó igen volt más) is
T2 közelebb közelebb van az orvos mint hogyha mondjuk (nem tom én [nem tudom] több kilométerről) kéne
A (igen szal [szóval] azér [azért])
T2 (nekem apukám)
A (igen de volt egy olyan is) ám mertén azt tudom ám hogy mentek ott a dolgok hozzá nem nyúlhatott a másik orvos az ügyeletes orvos nem nyúlhatott a másik orvos (betegéhez)

1.2. ábra

A társalgás szekvenciális szerkezetének reprezentálása

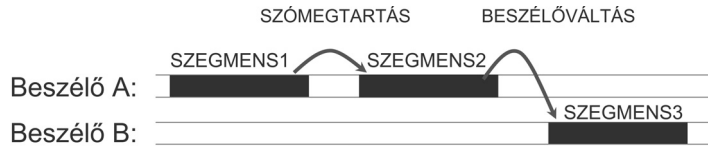


1.3.a ábra

A szóátadás sematikus ábrája

Ha nem történik meg a külválasztás, akkor a beszélgető partnerek egyike magához veszi a szót önkiválasztással.

Amennyiben sem a külválasztás, sem az önkiválasztás nem történik meg, abban az esetben az eredeti beszélő folytatja beszédét *(1.3.b ábra).*



1.3.b ábra

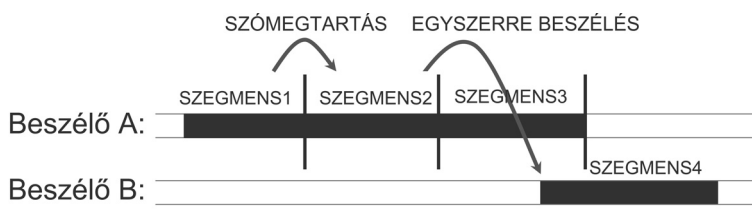
A szómegtartás sematikus ábrája

A társalgásokban a szóátadás, illetve a beszélőváltás sok más formában mehet végbe. Az egyik gyakori módozat, amikor *A* beszélő beszél, és *B* beszélő háttércsatorna-jelzéssel (például *ühüm, igen, oké, nem* stb.) átveszi a szó jogát *(1.3.c ábra)*.

1.3.c ábra

A szóátadás háttércsatorna-jelzéssel sematikus ábrája

A másik igen gyakori beszélőváltási mód, amikor *A* beszél, és mielőtt ő befejezné, *B* beszélő elkezdi beszédfordulóját, és ez egyszerre beszélést eredményez (*1.3.d ábra)*.



1.3.d ábra

A szóátadás egyszerre beszéléssel sematikus ábrája

Mind a háttércsatorna-jelzések, mind az egyszerre beszélések esetében nem szabályszerű, hogy ténylegesen meg is történik a beszélőváltás.

A lehetséges beszélőváltásra alkalmas helyeket általában a beszélő jelzi verbális, prozódiai (dallammenet, tempóváltozás, szünettartás) vagy nonverbális eszközökkel. Ugyanakkor a hallgató is jelezheti, hogy át kívánja venni a szót, amelyet a legtöbb esetben testtartással fejez ki. Az elmúlt évtizedekben számos jellemző szerepét vizsgálták a társalgások beszédlépéseinek előrejelzésében. Duncan (1972) azt feltételezte, hogy minden egyes interakcióban a beszélő és a hallgató bizonyos jeleket küldenek egymásnak, hogy milyen állapotban vannak a fordulóban. A beszélő különféle eszközökkel jelezheti a hallgatónak, hogy hol van lehetséges beszélőváltásra alkalmas hely: intonációval (csökkenő, emelkedő vagy monoton intonáció), gesztussal (kézmozdulat befejezésével vagy egy megfeszített kézpozíció ellazulásával),

konvencionális nyelvi jelekkel, szófordulatokkal – diskurzusjelölők *(tudod, de)* –, de kifejezheti paralingvális eszközökkel (hangerő vagy az alaphangmagasság csökkenése) vagy szintaxissal (egész szintaktikai egység).

Sacks és munkatársai (1974) a szintaxis szerepét hangsúlyozták a beszédjog átadásában. A teljes beszédlépés-szerkezeti egységet úgy lehet értelmezni, mint egy szintaktikai egységet, amely lehet egy mondat, mellékmondat, kifejezés vagy szó. Ezek az egységek mind szerepet játszhatnak a beszédlépés előrejelzésében: a hallgató el tudja dönteni, hogy a megnyilatkozás egy egészként zajlott-e le, vagy még kiegészítésre vár.

Selting (2000) szerint a műfaj és a tartalom is nagyon meghatározó a beszédlépések szerkezetében. A narratívák bevezető részében például a hallgató hosszan engedi a beszélőt megnyilatkozni.

A társalgás dekódolásában szintén fontos szerepet játszik az intonáció. Chafe (1994) szerint az intonációs egység egy alapvető egység, amelyet a lélegzetvétel szakít meg. Az intonációs egységet az alaphangmagasság változása, az időtartam, az intenzitás és a szünetek határozzák meg. Számos tanulmány foglalkozott az alaphangmagasság alakulásával a beszédlépések végén. Beattie (1982) Margaret Thatcherrel készített interjúkat elemzett, amelynek eredménye az volt, hogy több helyen is átvette a szót a riporter a beszélgetés során, még akkor is, ha Margaret Thatcher nem is akarta átadni a szót. Ezeknél a pontoknál az alaphangmagasság csökkenése volt megfigyelhető éppúgy, mint a szándékolt beszédlépés végénél. Tehát az alapfrekvencia mozgása eredményezte a riporter közbevágásait, amellyel bizonyította az alaphangmagasság fontos szerepét a beszélőváltásokban. Stephens és Beattie (1986) egy olyan kísérletet terveztek, ahol a résztvevőknek a társalgás átiratait kellett olvasni, illetve annak hanganyagát meghallgatni. Az átirat és a hanganyag társalgásokból kivágott beszédforduló belseji és végi megnyilatkozásokat tartalmaztak. Az eredmények azt mutatták, hogy a hanganyag alapján a résztvevők el tudták dönteni, hogy beszédlépésvégi megnyilatkozásról volt szó. Cutler és Pearson (1986) vizsgálatai szerint csak néhány dallammenet létezik, amely szóátadást jelezne, ezek karakterisztikái azonban egyértelműen nem meghatározhatók.

A szóátadás szándékát szintén jelezheti hosszabb néma vagy kitöltött szünet (Maclay–Osgood 1959). Beattie (1977) azt figyelte meg, hogy a társalgásban részt vevők gyakran szakítják meg a másikat, ha a beszédjelben hosszabb néma szünet van, illetve ahol kitöltött szünet realizálódik, bár ez függ attól is, hogy a hezitációt követi-e néma szünet, vagyis kombinált szünet jelentkezik a beszédben. Ugyanis ha a beszélő tovább kívánja folytatni a beszédét, akkor a legtöbb esetben csak kitöltött szünetet használ (Horváth 2009). Ugyanakkor a beszédtempó is alkalmas lehet a beszédlépés belseji, illetve végi megnyilatkozások elkülönítésére (Stephens–Beattie 1986).

Ford és Thompson (1996) eredményei azt mutatták, hogy a szünet segít befejezetté tenni az intonációs egységeket (0,3 másodperc vagy annál hosszabb szünet). Ugyanakkor a szünet nem minden esetben jelzi előre az intonációs egység végét. Local és Kelly (1986) két funkcióját feltételezték a szünetnek: az első, amikor a beszédjelben szünet keletkezik, amely

lezárásra utal; a másik, amikor a szünet a beszéd folytatását jelzi előre. Vizsgálataikban különös figyelmet fordítottak a kitöltött szünet előtti néma szünetre. Itt is két típust feltételeztek: az első típusban a hezitálást néma szünet követi, amely az utána lévő szóhoz kapcsolódik (ekkor a beszélő magánál tartja a szót); a második típusban a kitöltött szünetet kilégzésből adódó néma szünet követi (ekkor a hezitálás még centralizáltabb formában realizálódik), amelyet a legtöbb esetben szóátadás követ.

A társalgások beszédfordulóinak irányításában szintén nagy figyelmet kapnak a mozdulatok, a gesztusok. Számos kutatás kimutatta, hogy a mozdulatoknak igen fontos és integrált része van a spontán társalgás beszédfordulóinak szerveződésében (Lerner 2003). Kendon (1967, 2002) szerint a gesztus számos céllal jelenhet meg, ezek közül az egyik a diskurzus beszédlépéseinek előrejelzése. A beszélő és a hallgató mozdulatai jelként szolgálhatnak a beszédlépés határának kifejezésében: a kéz- vagy karmozdulat lezárása előre jelezheti a beszédlépés végét; ennek ellentéteként a mozdulat folytatása a szóátadást gátolhatja meg.

Mindezen jellemzők együttes megjelenése és vizsgálata sokkal eredményesebben mutatja a szóátadás folyamatát, mint az egyes jellemzők külön-külön. Duncan és Fiske (1985) számos tanulmányt publikáltak az egyes jellemzők interakciójáról, mint a testtartás, a gesztusok, a kitöltött szünetek, a szomszédsági párok struktúrája. Ford és Thompson (1996) a szintaktikai szerkezeteket, az intonációt és a pragmatikai lezártságot vizsgálták. Eredményeik azt mutatták, hogy a teljes szintaktikai egységet az intonáció (alaphangmagasság-emelkedés, -csökkenés az intonációs egység végén), a pragmatikai lezártság (olyan egység, amely komplett társalgási cselekménynek tekinthető) jellemzi, amely igen gyakran a szóátadás helyét mutatja, vagyis egy komplex lehetőséget a hallgatónak, hogy átvegye a szót. Wennerston és Siegel (2003) szintén a beszédlépéseket, mint komplex folyamatot, vizsgálták, főként fonológiai és szintaktikai interakciók együttes működéseként. Tanulmányaikban az intonáció, a szünet és a szintaxis szerepét elemezték. Megállapították, hogy mind a három bonyolult együttműködéseként jön létre a szóátadás, illetve hogy az intonáció sok esetben képes felülírni a szintaxis által kijelölt határokat. Elemzéseikből továbbá az is kiderült, hogy az az intonációs egység, amely erősen emelkedő mintázattal realizálódik, nagyobb valószínűséggel jelzi a beszédlépés végét, míg az az intonációs egység, amely alacsonyan emelkedő mintázattal valósul meg, ez a legtöbb esetben a beszéd folytatását jelzi. Megállapították továbbá azt is, hogy a korpuszban azon intonációs egységek, amelyek erősen emelkedő mintázattal realizálódtak, nem feltétlenül kérdő megnyilatkozások voltak. Kiemelték továbbá azt is, hogy amikor hosszabb szünet jelent meg (0,5 másodpercnél nagyobb), akkor a beszélő továbbra is folytatta beszédét. Ezt azzal magyarázták, hogy a hallgatónak 0,3 másodpercnél lett volna lehetősége átvenni a szót (Ford–Thompson 1996), de ezt elmulasztotta, így a beszélő folytatta megnyilatkozását. Ugyanakkor azonban ezt nagyban egyénfüggőnek találták.

Az utóbbi évtizedben egyre fontosabbnak tűnik a diskurzusjelölők szerepe a beszédfordulók előre jelzésében (Sacks et al. 1974; Schiffrin 1987; Wennerston–Siegel 2003; magyarra Dér 2010; Markó–Dér 2011; Shirm 2011). A diskurzusjelölőket (DJ) a magyar nyelvben több

különböző elnevezéssel szokás illetni: *konnektorok, pragmatikai kötőszók, társalgásszervező és -jelölő elemek, bevezető szók és kifejezések* stb. A DJ megnevezése az angol nyelvben sem egységes: *discourse markers, discourse deictics, discourse connectors, discourse particles, discourse operators, cue phrases* stb. (Fraser 1999: 932–937; Schourup 1999: 227–265). A diskurzusjelölők olyan nyelvi-pragmatikai egységek, amelyek a társalgásban ismertetőjegyei lehetnek a beszédfordulóknak, így nagyban hozzájárulhatnak a beszélőszegmentáláshoz, a diskurzus működésének megértéséhez (Markó–Dér 2011; Fraser 1999: 931; Louwerse–Mitchell 2003: 199). A szakirodalomban a diskurzusjelölőket a funkciójuk alapján szokás elkülöníteni a nem diskurzusjelölői szerepű szavaktól. Így alapvetően ezen elemeit a társalgásnak funkcionális csoportként tartják számon a szakirodalomban. Kategorizálását azonban nagyban megnehezíti, hogy eredetüket tekintve igen heterogén csoport, hiszen különböző szófajokból eredhetnek (határozószó, kötőszó, ige stb.), illetve különböző nyelvi szintű egységekből származhatnak (lexémák, különféle szintagmák stb.), és mindemellett nonverbális diskurzusjelölők is léteznek (Schiffrin 1987: 328; Markó 2005, 2006). Diskurzusjelölők nagyobb számban a beszélt nyelvben fordulnak elő, de egyes írott műfajokban is megtalálhatók (Dér 2006; Schiffrin 2001: 55). A kutatások többsége megegyezik abban, hogy a DJ-knek fontos szerepük van a beszédlépések szerveződésében, de önmagukban nem elégségesek. Schiffrin (1987) szintén amellett érvel, hogy számos különböző tényező vesz részt a társalgásszerkezet váltásában (1987: 117). Emellett a DJ-knek beszédlépésvéget jelölő szerepet feltételez (1987: 25), illetve beszédlépés-fenntartó szerepet, mint például a *you know* (Schiffrin 1987: 292). Sacks és munkatársai (1974) a DJ-k beszédlépéskezdő szerepét hangsúlyozták a *well, but, and, so* DJ-ket vizsgálva. Sok esetben az egyes diskurzusjelölő akár multifunkcionális is lehet: lehet beszédlépés-indító, -záró és -fenntartó szerepben is, mint ahogy az angolban az *uhm, yes* (Fisher 2000). Az angol nyelvű társalgásban a beszédlépések elejének 44%-ában szerepelt diskurzusjelölő (Heeman–Allen 1999), ami szintén megerősíti Dér (2012) kutatási eredményeit a magyar nyelvre vonatkozóan. Úgy tűnik tehát, hogy ez a funkció több nyelvnél is hasonló aránnyal fordul elő. A magyar nyelvben a verbális eszközök spontán társalgásokban való vizsgálatával, azon belül is a diskurzusjelölők szerepével csekély számú tanulmány foglalkozott (Dér 2010; Markó–Dér 2011). Az egyik legkiterjedtebb elemzést a témában Dér (2012) végezte el. Spontán társalgásokban elemezte a diskurzusjelölők gyakoriságát a beszédlépések kezdetén és végén. Az eredmények azt mutatták, hogy számos magyar diskurzusjelölő, a kötőszói eredetű jelölők, tipikusan az általa bevezetett beszédlépés elején fordulnak elő (például *hát, tehát, és, de*); kivételt azok képeznek, amelyek kötőszóként *sem* vagy *nem mindig* tagmondatkezdő helyzetűek (például *meg, pedig, bár*). Továbbá kimutatta, hogy a diskurzusjelölők előfordulása igen magas a beszédlépés kezdetén (43%), amelyek számos változatban jelenhetnek meg. Az előfordulások több mint felét (581 db, 52,7%) mindössze háromféle egyszavas diskurzusjelölő adta ki: a *hát,* a *de* és az *és*. Az elemzések során megállapította, hogy a beszédlépések 43,38%-ában szerepelt a szóátvételkor diskurzusjelölő elem. Mivel a DJ-k ilyen jelentős számban fordulnak elő a beszédlépések elején, ezért felmerült az igény ezen elemek automatikus osztályozására.

A háttércsatorna-jelzéseket az egyszerre beszéléseken belül a korai kutatások kezdetben csupán a társalgások egy igen érdekes jelenségeként vizsgálták, amelyeknek tipikusan szociális interakciós szerepet tulajdonítottak (Yngve 1970; Sacks et al. 1974; Duncan–Fiske 1985; Ward 1997). A háttércsatorna-jelzések többsége olyan jelenség, amely igen rövid, a hallgató a beszélő megnyilatkozása alatt mondja ki, illetve amely nem a szóátvételre irányul, sokkal inkább a beszélőt motiválja beszédének folytatására. Az általános definíció szerint ez a jelenség alapvetően arra szolgál, hogy a beszélőt informálja arról, hogy a hallgató az üzenetet megkapta, megértette, elfogadta vagy valamilyen hiba miatt a beszélőt kiegészítésre kéri (például *mmm, hm, aha, ja ühüm, aha igen, aha tudom*). A strukturális jellegzetességgel foglalkozó tanulmányok többsége a háttércsatorna-jelenségeket más nemverbális jelenségek, mint kézmozgás, gesztus, nevetés összekapcsoltságában vizsgálta a társalgásokban (vö. Birdwhistell 1962; Kendon 1967; Dittmann–Llewellyn 1968). Számos kutatás a háttércsatorna-jelzéseket mint nem beszédlépés státuszút elemezték a társalgásokban (Yngve 1970; Duncan 1972; Duncan–Niederehe 1974; Duncan–Fiske 1985). Az újabb kutatások szerint ezek a jelenségek nem beszédlépések, és nem hordoznak új információt, hanem segítik a társalgás folyamatosságát, dinamikus szerkezetét. Továbbá az is jellemző rájuk, hogy többségükben átfednek a beszélő megnyilatkozásának végével. Elmondható, hogy megjelenésük függ az aktuális beszélő következő beszédlépésétől. Azt is megfigyelték, hogy előfordulhat, hogy a háttércsatorna-jelzéssel megfordul a beszédlépés, és a hallgató veszi át a szót. Azt is kifejezheti továbbá, hogy a hallgatónak nem áll szándékában átvenni a szót, további folytatásra kényszerítve ezzel a beszélőt. Bizonyos megközelítésekben a háttércsatorna-jelzések beszédlépések szerveződésében betöltött szerepét vizsgálják. A konverzációelemzés irodalmában számos olyan háttércsatorna-jelzést találunk, amely részletes leírással rendelkezik. Az elemzések szerint mindegyiket meg lehet különböztetni elhelyezkedésük és szerepük szerint a szekvenciális környezetükben, illetve hogy milyen hatással vannak a későbbi beszédlépésre. Ezek a tokenek a következők: *yeah, uh huh, mm, hm* (Schegloff 1982; Jefferson 1984; Drummond–Hopper 1993); *oh* (Heritage 1984); *wow, good* (Goodwin 1986); *okay* (Beach 1995); *mm* (Gardner 2001). Schegloff (1982) az *uh, huh* háttércsatorna-jelzést vizsgálta angol nyelvű társalgásokban. A háttércsatorna-jelzéseket osztályozó rendszerek többsége Duncan és munkatársai munkásságán alapul (Duncan 1972; Duncan–Fiske 1985; Duncan–Niederehe 1974). Csoportosításukban a háttércsatorna-jelzéseket megkülönböztetik a többi hallgatói és beszélői viselkedéstől, mivel ezeknek nincs beszédlépésstátuszuk. Duncan és Fiske (1985) amellett érvel, hogy a háttércsatorna-jelzések nem alkotnak beszédlépést. A háttércsatorna-jelzés nem beszédlépésként való értelmezése igen problematikus, mivel maga a beszédlépés sem jól definiált. Ez vezetett Schegloff (1982) azon felvetéséhez, hogy a háttércsatorna-jelzések beszédlépésstátuszát eseti elbírálás alapján kell értékelni a lokális szekvenciális környezet alapján, utalva a szekvenciális és az interakciós célokra, amelyek megteremtik ezt a környezetet.

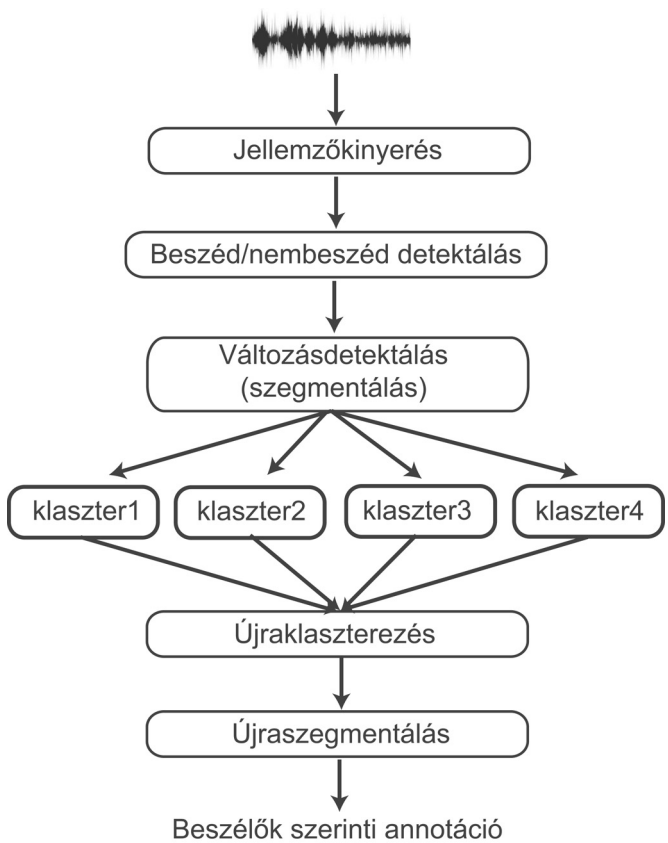
A beszédfordulókra irányuló elemzések többsége az angol nyelvre történt meg. Néhány vizsgálat létezik azonban más, főleg német (Auer 1996), spanyol (Placencia 1997), japán (Hayashi 1991; Tanaka 2001) nyelvre is. A magyar beszédfordulókra irányuló elemzésekre is

történtek már kísérletek, főként a prozódia és a szintaxis együttes működésével kapcsolatosan a beszédfordulókban (Németh 2007, 2008; Bata–Gráczi 2009).

Az automatikus beszélődetektálás megvalósítására jelentős mennyiségű kutatás történt idegen nyelvre (Tritschler–Gopinath 1999; Sivakumaran et al. 2001; Lu–Zhang 2002a; Cettolo–Vescovi 2003; Cheng–Wang 2003; Vescovi–Cettolo–Rizzi 2003). Magyar nyelvre azonban idáig nem született olyan munka, amely a beszélődetektálás megvalósítását tűzte volna ki céljául. A beszélődetektáló hasznos lehet mind a nyelvészek, mind a beszédtechnológusok számára. A nyelvészek használhatják a konverzációelemzéshez, hiszen automatikusan lehet a rendszerrel a társalgásokat beszélők szerint annotálni. A beszélődetektálás továbbá a beszédtechnológiában, azon belül a beszédfelismerésben a beszélőadaptált rendszerek megalkotásában játszhat fontos szerepet, illetve a törvényszéki beszélőazonosításban, ahol a folytonos társalgásban automatikusan lehet szegmentálni az egyes beszélőket, és azonosítani őket.

2. A beszélődetektáló általános felépítése

A beszélődetektálás során a folyamatos társalgás beszédfordulóit automatikusan detektáljuk, majd az így kialakított beszédrészeket hozzárendeljük a beszélgetésben részt vevő személyekhez (Jin et al. 2004). A beszélődetektálás feladata tehát kettős (Jin et al. 2004; Kotti et al. 2008). Az első feladat a beszélő szerinti szegmentálás (speaker segmentation), a második a beszélőosztályozás (speaker clustering). Az első feladat célja a beszédforduló automatikus detektálása, vagyis azon időpillanat megtalálása, amikor a beszélők váltják egymást. A második feladatban pedig ezeket a szegmentumokat kell osztályozni beszélők szerint, azaz az egyes beszélőkhöz rendelni. Egy általános beszéddetektáló rendszer felépítése a *2.1. ábrá*n látható.



2.1. ábra
A beszélőosztályozó leegyszerűsített blokkdiagramja

A két alapvető feladat mellett számos más algoritmus is fontos szerepet játszik a beszélődetektáló működésében, mint például a beszéddetektálás (más néven: beszéd/nembeszéd detektálás), egyszerre beszélések detektálása stb.

A beszélődetektálás megvalósítására számos megoldás készült különböző nyelveken. Jóllehet a beszélődetektálás beszédtechnológiai szempontból univerzálisnak tekinthető, a társalgás azonban sok tekintetben nyelvspecifikus, így fontos lehet, hogy magyar nyelvre is jól működő rendszert hozzunk létre.

A beszélődetektálást (speaker diarization) úgy lehet definiálni, mint az audiodetektálás (audio-diarization) egy alfeladatát, amelynek célja a hangfelvételen a különböző beszélők váltakozásának automatikus meghatározása (Reynolds–Torres-Carrasquillo 2004). A beszélődetektálás fő kérdése, hogy „ki mikor beszél?", amely sok esetben referál a beszélő szerinti szegmentálásra és a klaszterezésre.

Reynolds és Torres-Carrasquillo (2004) szerint a beszélődetektálásnak három fő alkalmazási területe van, amelyek felé az utóbbi években kiemelt figyelem fordult:

i) Híradások (broadcast news): rádió- és tévécsatornák hírei; jellemzője, hogy reklámszünetekkel és zenével megszakított, egycsatornás (illetve könnyen egycsatornássá alakítható).

ii) Felvett társalgások (multiparty meetings): spontán társalgások, megbeszélések vagy előadások, ahol egyszerre több beszélő lép interakcióba egymással egyazon szobában vagy telefonon keresztül. Ezek többsége többcsatornás felvétel, tehát több mikrofonnal vagy mikrofontömbbel van felvéve.

iii) Telefonbeszélgetések (telephone conversations): egycsatornás felvételek, ahol kettő vagy több személy között folyik a beszélgetés.

A beszélődetektálás részei, a beszélőszegmentálás és a beszélőklaszterezés a mintaosztályozás (pattern recognition) családjába tartoznak, ahol az a feladat, hogy az egyes (diszkrét) kategóriák legyenek megfeleltetve folyamatos beszédjelnek (időben illesztve legyenek a beszédjelhez), és ezáltal a köztük lévő határok definiálva legyenek. A mintaosztályozás célja általában, hogy egy x mintát a mintának megfelelő O osztályba sorolja a minta valamely jellemzője alapján.

Maga a beszélődetektálás, ahogyan a beszédfelismerés is, szintén a mintaosztályozás családjába tartozik. Mind a beszédfelismerésnek, mind a beszélődetektálásnak olyan jellemzőkkel kell dolgoznia, amelyek jól reprezentálják az akusztikai hanglenyomatokat, illetve olyan algoritmusokat kell használnia (ezek lehetnek szabály-, illetve statisztikai alapúak), amelyek alapján a jellemzővektorokat automatikusan csoportokba tudják sorolni.

Általánosságban elmondható, hogy az adatok osztályokba való csoportosítása igen széles körben kutatott statisztikai adatelemző eljárás, amelyet számos területen alkalmaznak, mint a gépi tanulás, az adatfeldolgozás, a mintafelismerés vagy az osztályozás stb.

Ahhoz, hogy meghatározzuk, hogy ki beszél a hangfelvételen (osztályozási technika alkalmazása), meg kell határoznunk először azokat a szegmenseket a hanganyagban, amelyeket klaszterezni fogunk, és amelyek különböző hosszúak, és különböző akusztikai karakterekkel rendelkezhetnek (beszéd, nembeszéd, zene, zaj). A csoportosítani kívánt egységek

kialakításához szegmentálási technikákat szokás alkalmazni, amelyek képesek a hanganyagot beszélők szerint felosztani (tehát a szegmentálás ebben az esetben nem szavakra vagy hangokra történik). A beszélőklaszterezés során meg kell határozni, illetve modellezni kell azokat a beszélői sajátságokat, amelyek az egyes beszélőkre jellemzők lehetnek (a feladat ebben rokon a beszélőfelismeréssel), és ki kell dolgozni azokat az eljárásokat, amelyek a beszédből származó adatokat hozzárendelik az egyes – akár előzetesen ismeretlen – beszélőkhöz. Ehhez a feladathoz megfelelő akusztikai modellek szükségesek, amelyeket számtalan algoritmussal elő lehet állítani (Reynolds–Torres-Carrasquillo 2004). A megfelelő algoritmus megválasztása azonban nem olyan egyértelmű. Gyakori, hogy a különféle osztályozó algoritmusok szignifikánsan különböző osztályozási eredményt adnak.

A beszélődetektáló általános felépítését Anguera munkája alapján mutatjuk be (2006). Ebben a fejezetben először azokat az akusztikai jellemzőket mutatjuk be, amelyek jól alkalmazhatók a beszélő személyek reprezentálására. A hagyományos jellemzőkinyerő technikák mellett egyre nagyobb hangsúlyt kapnak az alternatív akusztikai jellemzők, amelyek jobban mutatják a beszélő akusztikai sajátosságait, vagyis beszélőspecifikusak. Ezután bemutatjuk azokat az általános technikákat, amelyek a beszélőszegmentálásban, illetve a beszélőklaszterezésben használatosak. A legtöbb beszéddetektálóban a beszélőszegmentálás az első lépés, ezért először ezt, majd másodikként a beszélőklaszterezést mutatjuk be.

2.1. Akusztikai jellemzők a beszélődetektáláshoz

A beszélődetektáláshoz beszélőalapú jellemzőkinyerési technikákat szokás alkalmazni, ahogyan a beszélőazonosításhoz, illetve a beszélőfelismeréshez is. A jellemzőkinyerés célja, hogy azokat az információkat emelje ki a beszédből, amelyek a feladathoz hasznosak, és szűrjön ki minden lényegtelen információt. A jelen feladatban a beszéd azon tulajdonságait keressük, amelyek alapján az egyes beszélők hatékonyan megkülönböztethetők. A beszélőosztályozás során általában egy vagy több jellemzőt használnak a számtalan közül. A leggyakrabban használt akusztikai jellemzők a rövid idejű spektrális burkológörbe érzeti transzformációján alapuló eljárásokkal nyerhetők: MFCC (Mel Frequency Cepstral Coefficients, Mel-frekvenciás kepsztrális együttható; Sahidullah–Saha 2012), PLP (Perceptual Linear Prediction, perceptuális lineáris predikció; Hermansky 1990); ezek kimenete a legtöbb esetben 10–20 együtthatóból álló paramétervektor. Az MFCC- és a PLP-jellemzők alapvető kiindulási pontja, hogy az emberi hallás nem egyformán érzékeny az egyes frekvenciaközökre. Az ember nemlineáris hallásának modellezésére az MFCC esetében az adott keret spektrális energiaeloszlását lineáris Mel-skálán szokás transzformálni, míg a PLP esetében az emberi percepcióra épülő szűrőt alkalmaznak.

Ezen akusztikai jellemzőket más beszédtechnológiai rendszerekben, például beszédfelismerésben is használják. Jóllehet ezek a jellemzők jól alkalmazhatók, mégsem lehet őket kifejezetten beszélőspecifikus jellemzőknek tekinteni, mivel nem koncentráltan a beszélők elkülönítésére fejlesztették ki. Az MFCC és a PLP esetében is a legtöbb esetben magas számú koefficienseket szokás alkalmazni, mivel a magasabb együtthatók tartalmazzák/tartalmazhatják a beszélőkre vonatkozó ismertetőjegyeket (Anguera et al. 2006a).

A rövid idejű akusztikai jellemzők mellett az alaphangmagasságot is vizsgálták mint lehetséges beszélőspecifikus jellemzőt (Kajarekar et al. 2004; Friedland et al. 2009). Soenmez és munkatársai (1998) a beszélő alaphangmozgását lineáris függvényekkel közelítették, és az azokból származtatott statisztikai paraméterekkel jellemezték az egyes beszélőket. Soenmez és munkatársai (1998) munkájára alapozva Janin és munkatársai (2003) eső és emelkedő dallamkontúrokat határoztak meg a prozódiai jellemzők (alaphangmagasság és energia) alapján. Az így kapott tendenciákra bigram modelleket számoltak, amelyekkel reprezentálták az egyes beszélőket. A bigram modellezés az N-gram modellezéshez tartozik, amely jelen esetben a prozódiai tendenciák sorozatának valószínűségét becsli (a bigram esetében két prozódiai tendencia sorozatát). A szupraszegmentális jellemzőket a tanulmányok többsége valamilyen rövid időszakaszban mérte, de emellett vannak olyan kutatások is, amelyek ugyanezen akusztikai jellemzőket hosszabb időszakaszokra (a legtöbb esetben a statisztikákat egy teljes beszélgetésre) számolták, ahol a célszemély és az imposztor közötti távolságot minden egyes beszélgetésre mért jellemzővektor között valószerűségi arány teszt (log-likelihood-ratio test) módszerrel számolták (Peskin et al. 2003; Reynolds et al. 2003). Mindezek mellett a szótagszintű akusztikai modellezés is elterjedt. Ennek előnye az, hogy nagy mennyiségű mintát kapunk. A szótagokat ebben az esetben automatikusan, a beszédfelismerő kimeneteként kapják. Az akusztikai jellemzőket (alaphangmagasság, energia, időtartam) minden egyes szótagra kiszámítják, majd GMM-mel (Gaussian Mixture Model, kevert Gauss-modell) vagy SVM-mel (Support Vector Machine, szupport vektor gép) modellezik azokat (Shriberg et al. 2005; Ferrer et al. 2007).

A standardnak számító akusztikai jellemzők mellett egyre nagyobb hangsúlyt kapnak olyan alternatív akusztikai paraméterek, amelyek kifejezetten a beszélő karakterisztikáját igyekeznek reprezentálni, és amelyek kifejezetten a beszélő modellezésére alkalmazhatók (Anguera 2006). Yamaguchi és munkatársai (2005) beszélőszegmentáló rendszerükben például az energiát, az alaphangmagasságot, a frekvenciacsúcs középpontját, sávszélességét és még három új jellemzőt használtak: az energiaspektrum temporális stabilitása, a spektrális burkológörbe alakja és ezen jellemzők keresztkorrelációja az energiaspektrummal.

Nguyen (2003) egy új nemlineáris jellemző normalizációs eljárást (SWAMP: Sweeping Metric Parameterization) javasol a háttérzaj, illetve a nem a beszélőtől származó zajok csökkentésére. Kutatásában igazolta, hogy ha ezeket a normalizált jellemzőket kombinálja a nem normalizált jellemzőkkel (MFCC), akkor az eredmények javulnak.

Kotti és munkatársai (2006) az MPEG–7-alapú akusztikai jellemzők mellett érvelnek, mint például az AudioWaveformEnvelop és az AudioSpectrumCentroid. Ez a két akusztikai

jellemző az MPEG–7 Audio standard csomag részei (Salembier et al. 2002). Az AudioWaveformEnvelop jellemző néhány értékkel reprezentálja a szélső adatokat (minimum és maximum) a beszéd hullámformájából. Az AudioSpectrumCentroid jellemző pedig a spektrum log-frekvenciás energiaspektrum súlyközéppontját (CoG: center of gravity, súlyközéppont) határozza meg.

2.2. Beszélőszegmentálás

A beszélőszegmentálás sok esetben a beszélőváltás-detektáláshoz hasonlatos, és igen közel áll a beszéd/nembeszéd detektálásához. A jelet, beszélőszegmentálást/-váltást detektáló rendszer a folytonos akusztikai jelben megtalálja, hogy hol van beszélőváltás. Általánosabban az akusztikai változásdetektálás célja megtalálni azt az időpillanatot, ahol az akusztikai jelben változás történik a felvétel során, amely lehet beszéd/nembeszéd, zene/beszéd és egyéb más kategóriák. Jelen esetben az akusztikai változásdetektáló feladata a beszédlépések megtalálása (Anguera 2006).

Sok esetben tévesen a *beszélőszegmentálás* kifejezés egyszerre jelenti a beszélőváltások megtalálását, illetve ezen részek homogén csoportokba való klaszterezését. A beszélőszegmentálást és a -klaszterezést fontos megkülönböztetni, mivel két teljesen különböző feladatról van szó. A beszélőszegmentálás alapvető célja ugyanis, hogy megtalálja azon időpillanatokat, amikor beszédlépés történik, míg a beszélőklaszterezéskor ezek a beszédlépések kerülnek csoportosításra, vagyis a szegmensek az egyes beszélőkhöz rendelődnek (Anguera 2006).

A beszélőszegmentálás megvalósítására két fő megoldási technika létezik a szakirodalom alapján (Anguera 2006). Az első megoldásban a váltási pontok az akusztikai adatok alapján egyetlen lépésben kerülnek meghatározásra (vö. Kim et al. 2005). A második megoldásban ez több lépésben valósul meg úgy, hogy a kimenet iteratív módon pontosabbá válik (vö. Cheng–Wang 2004). Az első lépésben több váltási pontot feltételez a rendszer; többet, mint amennyi valójában létezik, ami magas téves elfogadási hibát (false alarm rate) eredményez. A második lépésben ezeket interatívan felülvizsgálja az algoritmus, és törli azokat, amelyek nem szükségesek.

Egy másik megközelítésben a beszélőszegmentálást három főbb kategóriába lehet sorolni (Chen–Gopalakrishnan 1998b; Kemp et al. 2000; Chen et al. 2002; Ajmera 2004; Perez-Freire–Garcia-Mateo 2004); metrikus alapú, szünetalapú, modellalapú algoritmusok.

2.2.1. Metrikus alapú szegmentáló algoritmusok

A metrikus alapú szegmentáló algoritmusok a legtöbbet használt eljárások. Az algoritmus alapja, hogy valamilyen távolságot mér az akusztikai szegmensek paraméterei között, és megállapítja, hogy vajon az előző beszélőhöz tartozik-e, vagyis hogy beszédlépésváltás történt-e. A két akusztikai szegmens általában egymást követi, vagyis nincs átlapolódás, illetve a beszélőváltás a két keret között jöhet létre. A legtöbb távolságszámításon alapuló eljárás, amelyet akusztikaiváltozás-detektálásra használnak, alkalmazható beszélőklaszterezésre is annak megállapítására, hogy a két beszélői csoport azonos beszélőhöz tartozik-e (Anguera 2006).

Legyen két audioszegmens *(i, j),* amelyeket az akusztikai jellemzővektorukkal reprezentálunk X_i és X_j, és amelyek hossza N_i és N_j. Ezek középértéke és varianciája μ_i, σ_i és μ_j, σ_j. Mindegyik szegmenst modellezzük Gauss-eloszlással: M_i (μ_i, σ_i) és M_j (μ_j, σ_j), amely lehet egygaussos vagy többgaussos. Másrészről a két szegmenst összevonva X, a középérték és a variancia μ, σ, amelyet Gauss-eloszlással közelítve M (μ, σ).

Általánosságban elmondható, hogy két különböző távolságalapú megoldással lehet a két szegmenst összehasonlítani. Az egyik típus a statisztikai alapú távolság (statistic-based distance), a másik a valószerűségalapú technika (likelihood-based technique). A statisztikai alapú eljárás a két szegmensből számított elégséges statisztikákat hasonlítja össze úgy, hogy közben nincs szükség modellekre. A statisztikák számítása normál esetben igen gyors és jó becslést ad, ha N_i és N_j elég hosszúak a statisztikák számításához, és az adatokból származó modellek meghatározhatók az egygaussos középértékkel és a varianciával (Anguera 2006).

A második csoport a valószerűségalapú, amely annak a valószínűségnek az értékelésén alapul, amely azt fejezi ki, hogy az adott modell mennyire reprezentálja az adott hipotézist. Ennek számítása jóval lassabb (hiszen a modelleket tanítani és értékelni kell), de sok esetben az eredmények jobbak, mint a statisztikai távolságalapúaké, illetve a nagyobb modellekkel komplexebb adathalmazra is alkalmasabbak (Anguera 2006). A következőkben néhány népszerű metrikus alapú algoritmust mutatunk be.

2.2.1.1. Bayes-féle információs kritérium (BIC: Bayesian Information Criterion)

A BIC az egyik legtöbbet használt algoritmus a szegmentálásban, illetve a klaszterezésben, mivel számítása igen egyszerű és hatékony (Anguera 2006). A BIC a feltételes valószínűségszámítás alapjain nyugszik. A BIC-ben a modellkiválasztás úgy történik, hogy a valószínűségi kritérium értéke annál magasabb, minél magasabb a modell komplexitása, tehát bünteti a modellkomplexitást (szabad paraméterek összege a modellben) (Schwarz 1971, 1978). Legyen X_i egy akusztikai szegmens, a BIC modell értéke M_i, ami azt jelenti, hogy a modell mennyire jól illeszkedik az adatokra, amely a következőképpen definiálható:

$$BIC(M_i) = \log L(X_i, M_i) - \lambda \frac{1}{2} \#(M_i) \log(N_i).$$

Mivel a *logL* (X_i, M_i) az adatok log-likelihood értéke (a valószerűségi érték logaritmusa) a szóban forgó modellből származik, λ egy szabad paraméter, amely a modellezett adatoktól függ; N_i a keretek száma a szóban forgó szegmensben, és a $\#(M_i)$ a szabad paraméterek száma a modellben lévő M_i becsléséhez (Ajmera 2004). Ilyen kifejezés a Bayes Factor (BF) közelítése (Kass–Raftery 1995; Chickering–Heckerman 1997), ahol az akusztikus modelleket ML (maximum likelihood) módszerrel közelítik, és ahol N_i nagynak tekinthető.

Ahhoz, hogy a BIC-et használni tudjuk arra, hogy vajon a váltás a két szegmens között van-e, értékelni kell azt a hipotézist, hogy X jobban közelíti az adatokat, mint az a hipotézis, hogy a $X_i + X_j$ jobban közelít – a GLR (általános valószerűség arány: Generalized Likelihood Ratio) algoritmushoz hasonlóan –, amelyet a következőképpen számolunk:

$$\Delta BIC(i, j) = -R(i, j) + \lambda P.$$

Az $R(i, j)$ a következőképpen írható fel abban az esetben, ha a modellt egy Gauss-eloszlással hozzuk létre:

$$R(i, j) = \frac{N}{2} \log\left|\sum x\right| - \frac{N_i}{2} \log\left|\sum x_i\right| - \frac{N_j}{2} \log\left|\sum x_j\right|,$$

ahol a P egy büntető kifejezés, amely a szabad paraméterek számának a függvénye a modellben. A teljes kovarianciamátrixra felírva:

$$P = \frac{1}{2}\left(p + \frac{1}{2} p(p+1)\right) \log(N).$$

A büntetőfaktor tulajdonképpen a valószerűséget növeli a nagyobb modell esetében, míg a kisebb modell esetében csökkenti.

Abban az esetben, ha az adatokat több Gauss-szal kívánjuk leírni (GMM), akkor azt a következőképpen tehetjük meg:

$$\Delta BIC(M_i) = \log L(X, M) - (\log L(X_i, M_i) + \log L(X_j, M_j)) - \lambda \Delta\#(i, j) \log(N),$$

ahol a $\Delta\# BIC(i, j)$ a különbség értéke a szabad paraméterekben a kombinált modell és a két különálló modell között (Anguera 2006).

Noha a $\Delta BIC(i, j)$ két $BIC(i)$ kritérium közötti különbség, amely azt határozza meg, hogy melyik modell illeszkedik jobban az adatokra, a beszélődetektálás szakirodalmában szokás magára a különbségre is BIC-kritériumként hivatkozni. A BIC-algoritmust elsőként Chen és Gopalakrishnan (1998a) alkalmazta a beszélődetektálásban, ahol egy teljes kovarianciájú Gausst használtak az adatok modellezéséhez (Chen et al. 2002). Bár nem létezik eredeti formula, a λ paraméter úgy van bevezetve, mint a büntetőfaktor hatása az összehasonlításban, amely rejtett küszöbértéket alkot a BIC-különbséghez. Mivel a küszöbérték megválasztása fontos az adatok illesztéséhez, ezért számos tanulmány foglalkozott azzal,

hogy milyen módszerrel lehet ezt a szabad paramétert optimálisan megválasztani. Néhány tanulmány a λ paraméter automatikus megválasztása mellett érvel (Tritschler–Gopinath 1999; Delacourt et al. 1999a; Lopez–Ellis 2000; Delacourt–Wellekens 2000; Mori–Nakagawa 2001; Vandecatseye et al. 2004).

Ajmera és munkatársai (2003) GMM-et használtak minden egyes modellhez *(M, M_i* és *M_j),* míg az *M* modell felépítéséhez a M_i és M_j modellek összegét használták, így el tudták kerülni a büntetőfüggvény alkalmazását, hogy ne kelljen a λ értéket használni. Az eredmény hasonló volt a GLR metrikai megoldáshoz.

A Schwarz (1978) által javasolt BIC-algoritmusban az akusztikai vektorok száma a modell tanításából származtatható, amelynek előfeltétele, hogy a BIC számításakor az adatok konvergálnak a végtelenhez. A valóságban ez ott okoz problémát, ahol nagy az eltérés a két hosszú szomszédos ablak között, vagy a csoportok között, amiket összehasonlít. Néhány kutató az eredeti formulát kis módosítással sikeresen alkalmazta, akár a büntetőfüggvényt (Perez-Freire–Garcia-Mateo 2004), akár az általános értékeket (Vandecatseye–Martens 2003), hogy csökkentsék azok hatásait.

Számos implementációban a BIC-et a szegmentálás metrikájaként javasolják. Kezdetben Chen és Gopalakrishnan (1998b) több váltási pontot feltételező kétutas algoritmust alkalmaztak, később számos tanulmány (Tritschler–Gopinath 1999; Sivakumaran et al. 2001; Lu–Zhang 2002a; Cettolo–Vescovi 2003; Cheng–Wang 2003; Vescovi et al. 2003) követte ezt, és vagy egyutas, vagy kétutas algoritmust alkalmaztak. Ezen tanulmányok többsége amellett érvel, hogy progresszíven növekvő ablakhosszt és különböző hosszúságú szegmenseket érdemes használni a váltási pontok detektálásához.

Tritschler és Gopinath (1999) az igen rövid idő alatt történő beszélőváltásokra készített számos gyorsabb algoritmust. Sivakumaran és munkatársai (2001), Cettolo és Vescovi (2003), illetve Vescovi és munkatársai (2003) gyorsabb megoldást javasoltak a modell középértékének és varianciájának kiszámítására. Roch és Cheng (2004) a MAP (Maximum A Posteriori) adaptációs algoritmust alkalmazta, Miró (2006) az ML (Maximum Likelihood) algoritmust használta a paraméterbecsléshez.

A BIC-algoritmus előnye más statisztikai alapú metrikákkal összehasonlítva, hogy a számítása abban az esetben jóval gyorsabb, ha nagy felbontású jelen futtatjuk. Ennek ellenére a BIC-algoritmust gyakran használják más algoritmusokkal együttesen (Anguera 2006). Például a BIC-et szokás a kétutas beszélőszegmentálás második lépcsőjeként alkalmazni (finomításként). A DISTBIC-algoritmusban, amely szintén egy kétutas beszélőszegmentáló algoritmus, fontos, hogy a GLR első szegmentálása után a BIC-et alkalmazzák mint utószegmentálót (Delacourt et al. 1999a, 1999b; Delacourt–Wellekens 2000). Szintén ezen irányban Zhou és Hansen (2000), Kim és munkatársai (2005), Tranter és Reynolds (2004) Hotelling's T^2 távolság használatát javasolják, míg Lu és Zhang (2002a; 2002b) KL2 (Kullback–Leibler) távolságot. Vandecatseye és munkatársai (2004) normalizált GLR-t (NGLR) használtak az előszegmentáláshoz, míg normalizált BIC-et az utószegmentáláshoz.

2.2.1.2. Általánosított valószínűségarány (GLR: Generalized Likelihood Ratio)

A GLR-t mint változásdetektáló algoritmust először Willsky és Jones (1976), illetve Appel és Brandt (1982) mutatták be. A GLR szintén valószerűségi alapú metrikai eljárás, amely két hipotézis közötti arányt fejez ki: a H_0 mindkét szegmens azonos személyétől származik, ezért $X = X_i \cup X_j \sim M(\mu, \sigma)$ reprezentálja jobban az adatokat. Másrészről H_1 azt feltételezi, hogy különböző beszélőktől származik a két szegmens, ezért $X_i \sim M_i(\mu_i, \sigma_i)$ és $X_j \sim M_j(\mu_j, \sigma_j)$ együtt jobban megfelelnek az adatoknak (Anguera 2006). A hasonlóság aránya tulajdonképpen a valószerűség arányaként számolható a két hipotézis között:

$$GLR(i, j) = \frac{H_0}{H_1} = \frac{L(X, M(\mu, \sigma))}{L(X_i, M_i(\mu_i, \sigma_i))L(X_j, M_j(\mu_j, \sigma_j))}.$$

Ebből meghatározva a két szegmens közötti távolságot, $D(i, j) = -\log(GLR(i, j))$, egy megfelelő küszöbértéket megválasztva el lehet dönteni, hogy a két szegmens azonos beszélőtől származik vagy sem. A GLR-algoritmus különbözik a hasonló elnevezésű standard valószerűségi aránytól (LLR), mivel a GLR-ben a valószínűségi eloszlásfüggvény nem ismert, és az adatokból direkt módon kell becsülni, míg a LLR-ben a modelleknek a priori ismertnek kell lenniük (Anguera 2006).

A beszélődetektálásban a GLR-t általában két azonos méretű szegmensre szokás alkalmazni. Ezt az időablakot gördítik végig az akusztikai jelben. A küszöbérték lehet előre meghatározott vagy dinamikusan adaptált.

Bonastre és munkatársai (2000) a GLR-t egyutas megoldásként alkalmazták a szegmensekre, hogy a beszélőváltásokat előre jelezze. A küszöbértéket úgy állították be, hogy a tévesztési arányt minimalizálják (gyakoribb téves riasztások árán). Beszélődetektálójukban minden egyes szegmenst önálló potenciális beszélőhöz tartozónak tekintettek.

Gangadharaiah és munkatársai (2004) két beszélőre alkalmas szegmentálót fejlesztettek, amely kétutas szegmentáló volt. Az első lépésben GLR-t, míg a másodikban Viterbi-algoritmust használtak a szegmenshatárok finomítására.

Egy hasonló kétbeszélős szegmentálóban Adami és munkatársai (2002) az első lépésben a beszéd első részét az első beszélőnek tulajdonították, míg a második beszélőt akkor feltételezték, ha váltási pontot jelzett a GLR. Algoritmusuk a második lépésben azokat a szegmenseket választotta ki, amelyek az elsőben egyik beszélőhöz sem tartoztak. Közben ezeket a szegmenseket hasonlította össze a két beszélői modell GLR-értékeivel, és ahhoz a beszélőhöz rendelte a szegmenst, amelyiknél magasabb a hasonlósági mérték.

A váltásdetektálásban és indexelésben a szegmentáló második lépéseként Liu és Kubala (1999) büntető GLR-eljárást alkalmazott. A váltási pont elfogadására/elutasítására egy előre tanított beszédhangalapú dekódert használtak. A büntetés, amelyet ez a GLR-ben használt, arányos a tanításkor rendelkezésre álló adatokkal a két szegmensben:

$$GLR'(i, j) = \frac{GLR(i, j)}{(N_i + N_j)^{\theta}},$$

ahol θ empirikusan meghatározott. Hasonló megfogalmazásban használták Metze és munkatársai (2004) a GLR-t szegmentációs lépésként a társalgást átiratozó rendszerükben.

2.2.1.3. Gish-távolság (Gish-distance)

A Gish-távolság egy valószerűségalapú metrika, amelyet a GLR variációjaként kapunk (Gish et al. 1991; Gish–Schmidt 1994). A GLR két részre oszlik ($\lambda_{kovariancia}$ és $\lambda_{középérték}$):

$$D_{Gish}(i,j) = -\frac{N}{2}\log\left(\frac{|S_i|^{\alpha}|S_j|^{(1-\alpha)}}{W}\right),$$

ahol S_i és S_j a kovarianciája a két szegmensnek, $\alpha = \frac{N_1}{N_1+N_2}$, és W a súlyozott átlaguk

$$W = \frac{N_1}{N_1+N_2}S_1 + \frac{N_1}{N_1+N_2}S_2.$$

Kemp és munkatársai (2000) a Gish-távolsági mértéket ötvözték más algoritmussal a beszélőszegmentáláshoz.

2.2.1.4. Kullback–Leibler-távolság (KL vagy KL2)

A KL és a KL2 (Siegler et al. 1997; Hung et al. 2000) igen hatékonyan és jó eredménnyel alkalmazható a beszélőszegmentálásban. Az információelméletben a Kullback–Leibler-divergencia vagy -távolság két valószínűségi eloszlás különbözőségét méri. Az egyik tipikusan az elméleti eloszlást, míg a másik ennek egy modelljét reprezentálja. A közöttük lévő távolság felfogható úgy, mint a modellezésből származó információveszteség vagy hiba. Adott két random eloszlás: *X, Y,* a köztük lévő KL-távolságot (vagy eltérést) a következőképpen tudjuk számolni:

$$KL(X,Y) = E_x\left(\log\frac{P_X}{P_Y}\right),$$

ahol E_X várható érték, tekintettel az X valószínűségi eloszlásfüggvényére. Ha a két eloszlást Gauss-eloszlással közelítjük, akkor a következőképpen fejezhetjük ki a KL-távolságot:

$$KL(X,Y) = \frac{1}{2}tr[(C_X - C_Y)(C_Y^{-1} - C_X^{-1})]\frac{1}{2}tr[(C_Y^{-1} - C_X^{-1})(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T].$$

A Kullback–Leibler-távolság ugyan nem negatív, de nem valódi metrika, mivel nem szimmetrikus, azaz megkülönböztetheti a modellt és a modellezett eloszlást. A KL-távolságot szimmetrikussá lehet tenni a következő lépéssel:

$$KL2(X,Y) = KL(X,Y) + KL(Y,X).$$

Delacourt és Wellekens (2000) a KL2-távolságot használták első két lépésként a beszélőváltási pont meghatározására. Zochova és Radova (2005) a KL2 egy továbbfejlesztett változatát használták. Elsőként a szüneteket és a lélegzetvételeket szűrték ki a beszédből rövid idejű spektrális energiát és ZCR-t (Zero-Crossing Rate, nullátmenetek aránya) alkalmazva.

Hung és munkatársai (2000) a beszélődetektálásban akusztikai jellemzőként MFCC-t nyertek ki a beszédből, illetve az egyes beszédszegmensek között KL2-vel és Mahalanobis- és Bhattacharyya-távolsággal határozták meg a beszélőváltás helyét.

2.2.1.5. Más távolságmérési eljárások

A fent említett eljárásokon kívül még számos más technika létezik két szegmens összehasonlítására. A DSD (Divegence Shape Distance, Kim et al. 2005; Lu–Zhang 2002a, 2002b); az XBIC (Cross-BIC, Anguera–Hernando 2004; Anguera 2005; Juang–Rabiner 1985); a Cu-Sum-távolság (Basseville–Nikiforov 1993); a Kolmogorov–Smirnov-teszt (Deshayes–Picard 1986); a Mahalanobis–Bhattacharyya-távolság (Campbell 1997); a VQ (Vector Quantization, vektorkvantáló) algoritmus (Mori–Nakagawa 2001); Hotelling's T^2-távolság (Zhou–Hansen 2000; Tranter–Reynolds 2004).

Jóllehet ezeket a technikákat előszeretettel alkalmazzák, az egyik nagy hátrányuk, hogy mindenképpen meg kell határozni a küszöbértéket az elfogadáshoz, illetve az elutasításhoz. Ennek a problémának a megoldására számos javaslat, munka került napvilágra. Ezek többsége az automatikus paraméterválasztást javasolja, vagyis azt, hogy a küszöbértéket dinamikusan, adaptívan kell beállítani. Lu több kutatásában (Lu et al. 2002; Lu–Zhang 2002a, 2002b) amellett érvelt, hogy az adaptív küszöbértéket tegyék függővé a *P*-től, amelyet a következőképpen lehet kifejezni:

$$Th_i = \alpha \frac{1}{P} \sum_{p=0}^{P} D(i-p-1, i-p),$$

ahol α erősítő tényező (általában az értéke közel van az 1-hez).

2.2.2. Nem metrikán alapuló szegmentálók

Ebben a fejezetben két technikát mutatunk be a beszélőszegmentálásra: szünetalapú eljárás, modellalapú eljárás.

2.2.2.1. Szünetalapú beszélőszegmentáló

A szünetalapú technikán nyugvó beszélőváltást előre jelző eljárások azt feltételezik, hogy a beszélőváltás előtt, vagyis a két beszélő mintái között szünet van. Ezek többsége a beszédfelismerő rendszereknek adják át a beszélőktől származó beszédszeleteket, így nagyon fontos,

hogy egy-egy beszédrész ne tartalmazzon félbevágott szót, vagyis a beszélőváltás átfedő beszéd nélkül jöjjön létre. Ebbe a technikai megoldásba tartoznak az energiaalapú (energy-based), illetve a dekóderalapú (decoder-based) eljárások (Anguera et al. 2006b).

Az energiaalapú eljárások általában valamilyen energiaszint-követést (energy detector) használnak, hogy megtalálják a lehetséges beszélőváltási helyeket. A kereső általában egy görbe minimumát/maximumát kapja meg értékelésre, hogy az adott szegmens potenciálisan szünet-e. A küszöbérték általában előre meghatározott (Kemp et al. 2000; Wactlar et al. 1996; Nishida–Kawahara 2003). Siu és munkatársai (2003) a MAD (mean absolute deviation statistic) algoritmust használták, amely az energia változását méri egy-egy szegmensen belül, így határozva meg, hogy az adott szegmens szünet-e.

Ezzel szemben a dekóderalapú szegmentálók általában egy teljes beszédfelismerő rendszer részei, és így keresik meg a szüneteket a beszédben (Kubala et al. 1997; Woodland et al. 1997; Lopez–Ellis 2000b; Liu–Kubala 1999; Wegmann et al. 1998). Ezen munkák többségében a szünetek minimális hossza előre definiált, hogy csökkentsék a téves elutasítások számát. Nyilvánvalóan belátható, hogy a szünetek jelenléte és a beszélőváltás csak csekély mértékben korrelálnak egymással, így általában ezen rendszerekben csak hipotetikusan megjelölt szünetváltási helyeket feltételeznek, és egyéb algoritmusokkal egyértelműsítik azokat.

2.2.2.2. Modellalapú szegmentáló

A modellalapú szegmentálók (például a leggyakrabban használt GMM) a tanuló mintákból származtatott akusztikai osztályokat használják (ezek lehetnek férfi-nő, zene-szünet stb. és ezek kombinációja). Az audioszegmenseket pedig a leggyakrabban ML-algoritmussal (Maximum Likelihood) rendelik hozzá a modellekhez (Gauvain et al. 1998; Kemp et al. 2000; Bakis et al. 1997; Sankar et al. 1998; Kubala et al. 1997). Ebben a rendszerben a modellek közötti határokat feltételezik váltópontnak. Ez igen közel áll a beszédfelismerésben használt dekódervezérelt rendszerhez, hiszen abban is modellt alkotunk minden egyes beszédhangra és a szünetre is. A különbség az, hogy itt igyekeznek szélesebb modelleket megkülönböztetni. Ez a szegmentálási technika igen közel áll a beszélőklaszterező megoldásokhoz, amelyekben a különböző beszélők identitása (vagyis akusztikai osztálya) a priori ismert. A modellalapú szegmentálás, illetve klaszterezés alapvető problémája, hogy előzetes ismeretekkel kell rendelkezni a modellekről, vagyis előzetes tanításra van szükség. A beszélőklaszterezés területén manapság már egyre gyakoribb, hogy olyan rendszereket hoznak létre, amelyben nem szükséges előzetes információ, például a társalgásban részt vevő beszélők száma. Ennek ellenére számos tanulmányban használják ezt a fajta szegmentálási technikát.

Ajmera és munkatársai (2002), illetve Ajmera és Wooters (2003) az iteratív dekódolást alulról felfelé végezték. Kezdetben magas számú beszélőváltást feltételeztek, majd ezt csökkentették addig, amíg az az optimális számot el nem érte. Meignier és munkatársai (2001) és Anguera és Hernando (2004) fentről lefelé irányuló eljárást használtak. Kezdetben egy váltási pontot feltételeztek, majd ennek a számát növelték, amíg az el nem érte az optimális számot.

Ezen rendszerek többsége a GMM-et használta a különböző osztályok modellezéséhez, és ML/Viterbi-algoritmust az optimális beszélőváltási pontok meghatározásához. Lu és munkatársai (2001) SVM-et (Support Vector Machines, szupport vektor gépek) alkalmaztak az akusztikai osztályok modellezéséhez.

2.2.3. A beszélőszegmentáló algoritmusok összegzése

Összességében elmondható, hogy számtalan megoldás létezik a beszélőszegmentálás megoldására. A kutatások többsége azonban mégis a BIC-algoritmust alkalmazza vagy csak önmagában, vagy kiegészítve más algoritmussal, például KL2-vel. Ennek oka, hogy a BIC igen gyorsan számolható, és nincs szükség előzetes tanítási folyamatra, vagyis felügyelet nélkül működő rendszer. Emellett a bonyolultabb megoldások nem, vagy csak alig javítottak a beszélődetektálás eredményein. Ezért a jelen értekezésben mi is emellett az algoritmus mellett döntöttünk.

2.3. Beszélőklaszterezés

A beszélőklaszterezés azon technikák és algoritmusok összességét jelenti, amelyekkel az egyes beszédszegmensek homogén csoportokba, egy-egy beszélőhöz rendelhetők. A beszédszegmens természetesen nem feltétlenül jöhet csupán egy hangfájlból, hanem több különbözőből is. Ugyanakkor az is elmondható, hogy a hangfájlnak nem kell akusztikailag homogénnek lennie (tartalmazhat zajt, zenét stb.). A beszélődetektáló tehát olyan rendszer, amelyben megvalósul a bemenő akusztikai jel beszélő szerinti szegmentálása, illetve ezután megtörténik a beszélőklaszterezés, amely ezeket a szegmenseket homogén csoportokba rendezi. Léteznek olyan hibrid rendszerek is, amelyek ezt a két lépést azonos időben végzik (Anguera 2006).

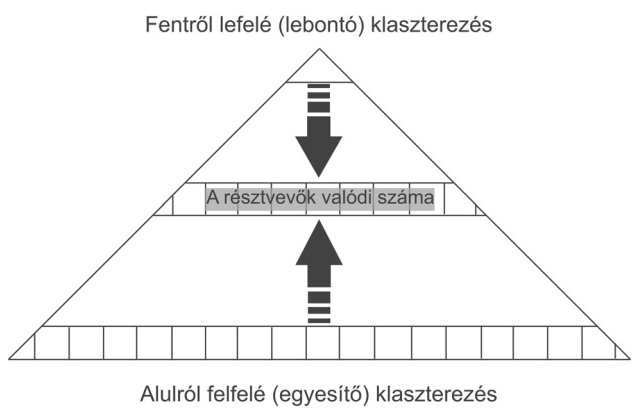
A beszélődetektálásnál is fontos megkülönböztetni az online és az offline rendszereket. Az offline rendszerben az adatok a rendszer működése előtt rendelkezésre állnak. A legtöbb kutatás ilyen rendszerekről számol be. Az online rendszerek esetében az adatok a felvétel során dolgozódnak fel. A legtöbb beszélődetektáló rendszer kezdetben egy beszélőt feltételez (aki elkezd beszélni a felvétel elején), és ezután iteratívan növeli a beszélők számát mindaddig, amíg az el nem éri az optimális beszélői indexet. Mivel az online rendszerek működése még nem elégséges, ezeket csak érintőlegesen mutatjuk be. Mori és Nakagawa (2001) olyan online rendszert épített, amelynek a klaszterező algoritmusa vektorkvantáló (Vector Quantizion), és amely a torzítás mérésén alapul (Nakagawa–Suzuki 1993). A rendszer kezdetben egy beszélőt szerepeltetett a kódkönyvben, majd fokozatosan adta hozzá a beszélőket, ameddig a kódkönyvben meg nem haladta a VQ-torzítás a küszöbértéket. Rougui és munkatársai

(2006) GMM-alapú rendszert javasoltak, amelyben módosított KL-távolságot használtak a modellek között. A beszélőváltási pontot a beszéd elérhetővé válásakor detektálta, és az adatokat hozzárendelte egy, az adatbázisban meglévő beszélőhöz, vagy új beszélőt készített. Ebben az esetben a hangsúly a beszélői szegmensek gyors, beszélők szerinti osztályozásán van, amelyben döntési fát használtak a beszélői modellek kialakításához.

A legtöbb beszélődetektáló azonban offline rendszer, amelyben az algoritmusokat úgy mutatják be, mint egy online is működtethető rendszert. Ezen rendszereket két csoportba lehet sorolni. Az egyikbe tartoznak a hierarchikus klaszterezési technikák (hierarchical clustering technique), amelyek az optimális beszélői számot iteratív módon keresik a lehetséges csoportok felosztásával vagy egyesítésével. A másik csoportot az olyan klaszterezési technikák alkotják, amelyek elsőként megbecsülik a klaszterek számát, és úgy alakítják ki a beszélők számát, nem számítva, hogy a csoport kisebb vagy nagyobb lesz-e (Anguera 2006).

2.3.1. Hierarchikus klaszterezési technikák

A legtöbb offline klaszterező algoritmus a hierarchikus technikát használja, ahol a beszédszegmensek vagy -klaszterek iteratívan vannak felosztva vagy egyesítve, amíg az optimális beszélőszámot el nem érik. Ezen technikák alapvetően két csoportra bonthatók *(2.2. ábra)* (Anguera 2006).



2.2. ábra
A klaszterező eljárások sematikus ábrázolása

Az alulról felfelé építkező felhalmozó eljárások kezdetben nagyszámú szegmenst/klasztert feltételeznek, majd az eljárás a legközelebbi klasztereket egyesíti, a hierarchiában egy szinttel feljebb újabb klasztert alakítva ki.

A fentről lefelé építkező lebontó módszerek egyetlen, minden adatpontot tartalmazó klaszterből indulnak ki, amit kisebb klaszterekre particionálnak, majd ezeket is tovább bontják.

Mindkét módszerben két lényeges feladatot kell megoldani:

1. A távolság mérése a klaszterek között, meghatározva ezzel az akusztikai hasonlóságot.
2. A megállási feltétel (stopping criterion) meghatározása, vagyis hogy mennyire, milyen magas szintig épüljön fel a hierarchikus klaszterezés fája, illetve az elkészült dendrogram melyik vágásával tudjuk a feladatot a legjobban megoldani.

2.3.1.1. Alulról felfelé (egyesítő, bottom-up) klaszterező eljárások

Ez a leggyakrabban használatos módszer a beszélőklaszterezésre, mert a beszélőszegmentálás technikája felhasználható a klaszterek kezdeti számának meghatározásához. Általában az aktuális klaszterek közötti távolságmátrixot számolja ki az algoritmus, és a legközelebbi párokat vonja össze iteratív módon mindaddig, amíg a rendszer el nem éri a megállási feltételt.

Az egyik korai munkában (Jin et al. 1997) a beszélőklaszterezést a beszélőfelismerőhöz működtették, amelyben a Gish-távolságot távolságmátrixként használták a közeli klaszterek összevonásához. A megállási feltételt a büntetősúlyt minimalizáló függvény adta: a büntetés súlya növekszik, ha túl sok klasztert hoz létre a rendszer (szabályozva a túlzott összevonást):

$$W_{Jin} = \left| \sum_{k=1}^{K} N_k \sum k \right| \sqrt{k},$$

ahol a K a klaszterek száma, a $\sum k$ a k klaszter kovarianciamátrixa, az N_k az akusztikai szegmens és $|\cdot|$ a determináns jele.

Közel egy időben Sigeler és munkatársai (1997) KL2-eltérést használtak a távolság mérésére. Megállási kritériumként az egyesítő előfeltételt alkalmazták. Az eredmények azt mutatták, hogy a KL2 jobban használható, mint a Mahalanobis-távolság a beszélőklaszterezésben. Zhou és Hansen (2000) szintén a KL2-metrikát használták a klaszterek közötti hasonlóság mérésére.

Általában elmondható, hogy a statisztikai alapú távolsági metrikák (nem igényelnek tanítást) korlátozottan működnek a beszélőklaszterezésben, mivel impliciten határozzák meg a távolságot a középérték- és a kovarianciamátrix között mindegyik szegmensre, ugyanakkor a beszélőklaszterezésben sokszor nem áll rendelkezésre egy beszélőtől elégséges mennyiségű adat a modellezéshez.

Rougui és munkatársai (2006) azt javasolták, hogy a két GMM-modell közötti távolságot a KL-metrika alapján mérjék. Adott két modell M_1, M_2 (K_2 kevert Gauss-modell mindkét modellre és a Gauss-súlyai $W_1(i)$, $i = 1 \ldots K_1$ és $W_2(j)$, $j = 1 \ldots K_2$), a távolság M_1 és M_2 között:

$$d(M_1, M_2) = \sum_{i=1}^{K_1} W_1(i) \min_{j=1}^{K_2} KL(N_1(i), N_2(j)),$$

ahol az $N(i)$ a modellből származó egyik Gauss-komponens.

Beigi és munkatársai (1998) az egyes kevert Gauss-komponensek között mérték a távolságot. A távolságmátrixban $d(i, j), \forall i, j$ a két modell minden lehetséges Gauss-párja között

meghatározásra kerültek a távolságok (a távolságot euklideszi, Mahalanobis- és KL-függvénnyel mérte), és a végleges távolságot úgy határozza meg, hogy minimalizálja a mátrix oszlopaiban és soraiban a súlyokat.

Ben és munkatársai (2004), valamint Moraru és munkatársai (2005) a klasztermodelleket MAP-adaptációval származtatják az előzetesen tanított GMM-ből. A távolságot a GMM-modellek között úgy számítják, hogy egy sajátos KL2-távolságot mérnek, ahol csak a középértékek vannak adaptálva (a variancia és a súlyok a modellből származnak). A távolság tehát a következőképpen számolható:

$$D(M_1, M_2) = \sqrt{\sum_{m=1}^{M} \sum_{d=1}^{D} W_m \frac{(\mu_1(m,d) - \mu_2(m,d))^2}{\sigma_{m,d}^2}},$$

ahol a $\mu_1(m,d)$ és a $\mu_2(m,d)$ a *d*-edik komponens átlaga, amely az *m* Gauss-komponens átlag vektora, a $\sigma_{m,d}^2$ az *m* Gauss *d*-edik komponense, és az *M, D* a kevert Gauss-komponensek száma és GMM-modell dimenziója.

A statisztikai alapú rendszereken kívül Gauvain és munkatársai (1998), valamint Barras és munkatársai (2004) a GLR-metrikát ismertették (hangoló paraméterekkel), ahol büntették a nagyszámú szegmenseket és a klasztereket a modellben. A klaszterterezés optimumát iteratív Viterbi-dekódoló és egyesítő iteráció detektálta, amelyben a rendszer megállási feltételét ugyanezzel a metrikával valósították meg.

Solomonov és munkatársai (1998) szintén GLR-t használtak, majd távolsági mátrixonként hasonlították össze a KL2-vel és iteratívan egyesítő klaszterezéssel mindaddig, amíg maximalizálva nincs az ún. becsült klaszter tisztasága (purity).

A legtöbbet alkalmazott távolsági és megállási kritérium a beszélőklaszterezés esetében is a BIC-algoritmus (Chen et al. 1998; Chen–Gopalakrishnan 1998a). A páronkénti távolságmátrix minden egyes iteráció során meghatározott, és az a pár, amelyik a legnagyobb BIC-értékkel rendelkezik, összevonásra kerül. Ez a folyamat akkor áll le, amikor az összes pár esetében $\Delta BIC < 0$. A későbbiekben ezt az eljárást fejlesztették tovább (Chen et al. 2002; Tritschler–Gopinath 1999; Tranter–Reynolds 2004; Gettolo–Vescoli 2003; Meinedo–Neto 2003).

Sankar és munkatársai (1995) és Heck–Sankar (1997) a szimmetrikus relatív entrópiatávolságot használták a beszélőklaszterezéshez, amelyet az ASR-rendszerben alkalmaztak a beszélőadaptáció megvalósításához. A távolság hasonló az Anguera (2005) és a Malegaonkar és munkatársai (2006) által használthoz, amelyeket a beszélőszegmentációhoz alkalmaztak. A következő egyenlettel kifejezve:

$$D(M_1, M_2) = \frac{1}{2}[D_{\lambda_1, \lambda_2} + D_{\lambda_2, \lambda_1}],$$

ahol D_{λ_i, λ_j} a következőt jelenti:

$$D_{\lambda_i, \lambda_j} = \log p(X_i | M_i) + \log p(X_i | M_j).$$

A megállási kritérium egy empirikusan megállapított küszöbérték volt. Később ugyanezen szerzők (Sankar et al. 1998) egy komponenst tartalmazó GMM-alapú klaszterezést valósítottak meg.

Ezen módszerek mellett megjelentek olyan technikák, amelyek a beszélőazonosítás és a beszélőfelismerés területéről érkeztek (Barras et al. 2004; Sinha et al. 2005; Zhu et al. 2005; Zhu et al. 2006). A rendszerben standard összevonáson alapuló (agglomeratív) klaszterezést használtak BIC-val. A büntetőparamétert, a λ-t úgy állították be, hogy több klaszter legyen mint az optimális. A beszélődetektáláskor először osztályozták a klasztereket nem és sávszélesség szerint (műsorhírekben). Minden egyes klaszterből univerzális háttérmodell segítségével (általános háttérmodell, UBM: Universal Background Model) és MAP-adaptálással alakították ki a beszélői modelleket (Wu et al. 2003a, 2003b, 2003c). A legtöbb esetben lokális jellemzővetemítő normalizációs (local feature warping normalization) algoritmust használtak azért, hogy a jellemzőkből eltávolítsák a nem stacionárius részeket. A beszélői modelleket metrikusan hasonlították össze kereszt-likelihood algoritmussal (Reynolds et al. 1998), amelyet a következőképpen lehet megfogalmazni:

$$D(X_1, X_2) = \frac{1}{N_1} \log \frac{p(X_1 | M_2 - UBM)}{p(X_1 | UBM)} + \frac{1}{N_2} \log \frac{p(X_2 | M_1 - UBM)}{p(X_2 | UBM)},$$

ahol $M_i - UBM$ azt fejezi ki, hogy a modell MAP-adaptált az UBM-ből. Empirikusan megállapított küszöbértéket használtak mint megállási kritériumot.

Néhány tanulmányban a beszélőszegmentálást a beszélőklaszterezéssel integrálták egy szegmentáló/klaszterező rendszerbe. A kezdeti szegmentálást használták fel a beszélői modellek tanításához, amelyet iteratívan dekódoltak, és újratanították az akusztikai adatokon. A küszöbértékmentes (threshold-free) BIC-metrikát használták ahhoz, hogy egyesítsék a közeli klasztereket minden egyes iterációval, és ezt a módszert használták a leállási kritériumhoz is (Ajmera et al. 2002; Ajmera–Wooters 2003; Wooters et al. 2004). Wilcox és munkatársai (1994) büntető GLR-algoritmust hoztak létre, ezen belül hagyományos agglomeratív (összevonó) klaszterezést. A büntetőfaktor azokat a klasztereket vonja össze, amelyek időben közel állnak egymáshoz. A klaszterek modellezéséhez általános HMM-et építettek az összes adatot felhasználva, és csak a súlyokat adaptálták minden egyes klaszterhez (Sankar et al. 1998). A végleges állapotot iteratív Viterbi-dekódolással érték el.

Moh és munkatársai (2003) egy új megoldást vezettek be a beszélőklaszterezésbe. A módszer lényege, hogy a beszélők klaszterezéséhez beszélő-háromszögelést (triangulation) használtak. Ennek lényege az, hogy az egyes beszédszegmenseket egy koordináta-rendszerben helyezik el, és ebben a koordináta-rendszerben keresik a klasztereket. Adott a klaszterek csoportja C_k, $k = 1 \ldots K$ és a nem átfedő beszédet tartalmazó szegmensek csoportja X_s, $j = s \ldots S$, amely a különböző alcsoportok/csoportok tagjait tartalmazza. Az első lépés során létrehozza az algoritmus a koordinátavektorokat minden klaszter egyes szegmenseihez (teljes kovarianciájú GMM-mel modellezve), amit úgy visz véghez, hogy kiszámolja minden

egyes klaszter valószerűségét az egyes szegmensekhez. A hasonlóságot két klaszter között úgy definiálja, mint keresztkorrelációt a két vektor között:

$$C(k, j) = \sum_s p(C_k | X_s)\, p(C_j | X_s),$$

összevonva azokat a klasztereket, amelyek nagy hasonlóságot mutatnak.

2.3.1.2. Fentről lefelé (lebontó, top-down) klaszterező technikák

A szakirodalomban csak néhány olyan beszélődetektáló rendszer van, amelyben a beszélőklaszterezés egyetlen klaszterből indul ki, és iteratívan bontja azt kisebb csoportokra addig, amíg a megállási kritérium nem találkozik a kívánt klaszterszámmal.

Johnson és Woodland (1998) lebontó klaszterezési technikát alkalmaztak a beszélők csoportosításához az ASR-rendszerben (Johnson 1999; Tranter–Reynolds 2004). Az algoritmus addig fut iteratívan, ameddig négy alcsoportot nem képez, majd ezután összevonja azokat, amelyek nagyon hasonlók egymáshoz.

2.4. Beszéddetektálás

A beszélődetektáláshoz fontos és alapvető feladat a beszéddetektálás (VAD: Voice Activity Detection), amely a beszédfelismerésben, de a hagyományos telefóniában is fontos szerepet tölt be. A beszéddetektálás során azon részeket határozzuk meg az akusztikai jel alapján a folyamatos beszédben, ahol beszéd detektálható, kiszűrve ezzel a szüneteket, a köhögéseket, illetve egyéb zajos részeket.

A beszéddetektálás igen egyszerűnek tűnhet, de a technológiai megvalósítása korántsem triviális, jóllehet számos beszédfelismerő tartalmazza, a mai napig nem megoldott kihívás a beszédtechnológia számára.

Az elmúlt évtizedekben számos kísérletet végeztek a beszéddetektálás tökéletesítésére (Sohn et al. 1999; Cho–Kondo 2001; Gazor–Zhang 2003; Armani et al. 2003). A beszéddetektáló algoritmusban különféle akusztikai jellemzőket lehet felhasználni, mint például a jel energiáját (Woo et al. 2000), az alaphangmagasságot (Chengalvarayan 1999), a spektrumanalízist (Marzinzik–Kollmeier 2002), a nullátmenetek számát (zero-crossing rate) (ITU-T 1996), a periodicitás mértékét (Tucker 1992) vagy a magasabb rendű statisztikai jellemzőket, mint az LPC-analízist (Nemer et al. 2001) vagy ezek különféle kombinációit (ITU-T 1996).

A tipikus beszéddetektáló három részből áll: *i)* zajcsökkentés, *ii)* jellemzőkinyerés, *iii)* döntés. A zajszűrés általános megoldása a Wiener-szűrő, mely a szűrő kimenete és a kívánt jel átlagos négyzetes távolságát minimalizálja. A döntési feladat alapvetően két

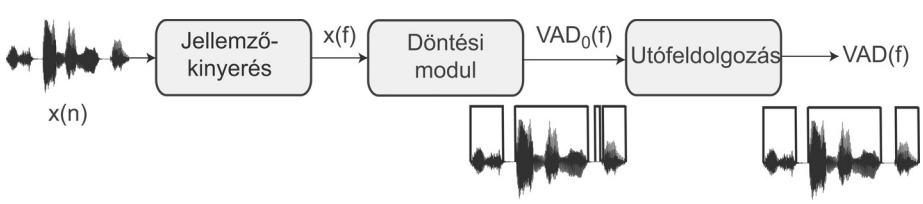
módszertani megoldásra vezethető vissza: a szabályalapúra és a statisztikai alapúra. A jellemzőkinyerés során olyan akusztikai paramétereket keresünk az akusztikai jelben, amelyek alapján elkülöníthetők a beszéd és a nembeszéd szegmensek. A Sohn és munkatársai (1999) által kifejlesztett beszéddetektáló rendszerben egy statisztikai alapon működő algoritmust alkalmaztak, amely a jel energiáját használja akusztikai jellemzőként, az adatok eloszlásának modellezéséhez pedig Gauss-eloszlást alkalmaztak. A Ying és munkatársai (2011) által javasolt beszéddetektáló felügyelet nélküli tanuláson alapuló eljárással (szekvenciális kevert Gauss-modell) automatikusan osztályozta a beszéd és a nembeszéd részeket. Akusztikai jellemzőként pedig Mel-frekvenciás spektrális szűrőt használt. Mind a beszéd, mind a nembeszéd szegmens modellezése két GMM-mel történt. Az ITU-T által kidolgozott G.729B beszéddetektáló szabvány egyszerre négy akusztikai paramétert is alkalmaz: a teljes és az alacsony frekvencia energiáját, a lineáris spektrális elemzést (LPC-vel) és a nullátmenet számát (zerocrossing rate).

2.4.1. A beszéddetektáló általános leírása

A VAD kihívása, hogy detektáljuk a jelen levő beszédjelet a zajos jelben. A beszéddetektálóban alkalmazott döntés a bemenő jellemzővektorokra történik, amely legyen x. Feltételezve, hogy a beszédjel és a zaj additív, a beszéddetektáló modul alapvetően a két következő hipotézis közül választ:

$$H_0 : x = n$$
$$H_1 : x = n + s.$$

A VAD tipikus megvalósítását a következő blokkdiagram mutatja *(2.3. ábra),* amely három fő modult tartalmaz: *i)* jellemzőkinyerés, *ii)* döntési modul, *iii)* utófeldolgozás.



A beszéddetektáló tipikus megvalósításának blokkdiagramja

2.4.2. Jellemzőkinyerés a beszéddetektáló megvalósításához

A beszéddetektáló létrehozásakor olyan akusztikai jellemző(ke)t szokás figyelembe venni, amely diszkriminatív tulajdonsággal bír a beszéd és a nembeszéd szegmensek automatikus osztályozásához. Számos megvalósítás alkalmazza: *i)* a fullband (energia a teljes spektrumra)

és a subband (részsáv: a spektrum felbontása kisebb frekvenciatartományokra) energiát (Woo et al. 2000), *ii)* a spektrális eltérést a beszéd és a háttérzaj között (Marzinzik–Kollmeier 2002), *iii)* az alaphangmagasságot (Tucker 1992), *iv)* a nullátmenetek számát (zero crossing rate) (Rabiner et al. 1975) és *v)* a magasabb rendű statisztikákat (Nemer et al. 2001; Ramírez et al. 2006; Górriz et al. 2006; Ramírez et al. 2007).

A legtöbb beszéddetektáló az éppen érkező hangjelre (frame: keret) határozza meg a döntést, és nem veszi figyelembe a kontextust, vagyis az előtte és a mögötte álló jelet. Ugyanakkor vannak olyan eredmények, amelyek szerint a hosszú idejű akusztikai jellemzők jól használhatók magas zajjal terhelt akusztikai jelre (Ramírez et al. 2004; Ramírez et al. 2005).

2.4.3. A beszéddetektáló döntési modulja

A beszéddetektáló döntési modulja tartalmazza azt a szabályt vagy eljárást, amely meghatározza a jellemzővektorra (x-re), hogy beszéd vagy nembeszéd. Sohn és munkatársai (1999) olyan VAD-algoritmust hoztak létre, amely statisztikai eljárást, valószínűségiarány-tesztet (likelihood-ratio test, LRT) alkalmazott egyetlen jellemzőre. Az eljárás két hipotézisteszt használ, ahol az optimális döntési szabályt a valószínűségi hiba minimalizálásával éri el Bayes-osztályozóval.

Adott egy megfigyelt vektor az osztályozáshoz; az alapvető feladat két hipotézistesztre redukálható, amely szerint két hipotézis közül (H_0 vagy H_1) a legnagyobb feltételes valószínűséggel rendelkezőt választjuk $P(H_i|x)$:

$$P(H_1|x) \underset{<H_0}{\overset{>H_1}{}} P(H_0|x).$$

Felhasználva a Bayes-szabályt az LRT kiszámolásához:

$$\frac{P(H_1|x)}{P(H_0|x)} \underset{<H_0}{\overset{>H_1}{}} \frac{P(H_0)}{P(H_1)}.$$

2.4.4. A beszéddetektáló utófeldolgozása (simítás)

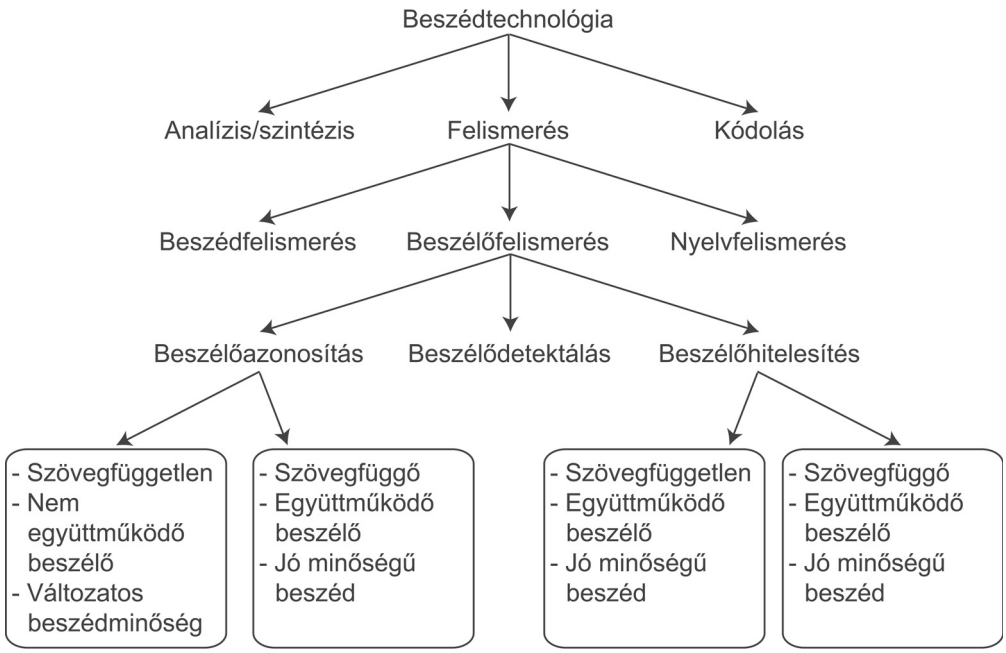
A legtöbb VAD a döntését az akusztikai jel keretekre bontása után minden egyes keretre hozza meg. Ezt a döntési sorozatot vizsgáljuk felül a beszéddetektáló utolsó lépéseként. Ebben a modulban használhatunk előre meghatározott küszöböt arra vonatkozóan, hogy minimálisan milyen hosszú lehet egy-egy szünet, például 100 ms.

2.5. Beszélőspecifikus jellemzők a gépi beszélőfelismerésben

Az utóbbi évtizedekben egyre nagyobb figyelmet kap az automatikus beszélőfelismerés megvalósítása a kriminalisztikai fonetikában és a beszédtechnológiában. Ezt jól reprezentálja, hogy a beszédtechnológiai konferenciákon elhangzó előadások közel egyharmadát a beszélőfelismerés témaköre adja. A beszélőfelismerő rendszerek számos más beszédtechnológiai alkalmazásba integrálhatók, ilyen például a beszédfelismerés, de a napjainkban a legdinamikusabban fejlődő beszélődetektálónak is szerves része.

A mindennapi életben képesek vagyunk akár néhány másodperces hangmintából azonosítani az általunk ismert személyeket. Ez azért lehetséges, mert a beszédhang olyan akusztikai jegyeket tartalmaz, amelyek jól reprezentálják az adott egyént (Bőhm 2007). Kutatások kimutatták, hogy a hangfelismerésért, akárcsak az arcfelismerésért, egy külön agyi terület felelős. Képalkotó eljárások ugyanis bizonyították, hogy más-más agyterület aktiválódott az ismert és nem ismert személy beszéde hallatán (Belin et al. 2004). Az ismert személyek felismerése mellett képesek vagyunk a nem ismert személyekről is profilt készíteni, vagyis általános információkat adni például a nemre (Lass et al. 1976), az életkorra (Ptacek–Sander 1966; Gocsál 1998), testalkatra (Dommelen–Moxness 1995; Gósy 2001) vagy hangulatra (Scherer et al. 2001) stb. vonatkozóan.

A gépi beszélőfelismerés három területre osztható (Campbell 1997) (vö. *2.4. ábra*). Megkülönböztetünk beszélőazonosítást (speaker identification), beszélőhitelesítést (speaker verification) és beszélődetektálást (speaker diarization). A beszélőhitelesítés célja, hogy egy személyről eldöntse, hogy ő az, akinek állítja magát. Ez a cél megegyezik a többi biometrikus személyazonosítás (például ujjlenyomat-, íriszvizsgálat) céljával. Ebben a feladatban bináris döntést kell hoznia a gépnek: elfogadás/elutasítás. Ekkor a beszélőnek érdeke, hogy a gép felismerje a hangját, ezért a beszédminőség igen jó. Ezzel ellentétben, a beszélőazonosítás célja, hogy a beszélők egy lehetséges köréből kiválasszuk az aktuálisan beszélőt (Campbell 1997). Ez a feladat osztályozási problémára vezethető vissza. Lehetséges azonban az is, hogy a lehetséges beszélők halmaza nyílt, vagyis a beszélő nincs benne a halmazban, ekkor a rendszer ismeretlen személyként azonosítsa a beszélőt. A beszélődetektáláskor két- vagy több-beszélős társalgásokban kell azonosítani azt, hogy ki mikor beszél. A beszélőazonosítás és beszélőhitelesítés lehet szövegfüggő vagy szövegfüggetlen. A kutatók általában a szövegfüggetlen osztályozásra törekednek, mivel ekkor tetszőleges tartalmú beszédminta alapján történhet a beszélő azonosítása vagy hitelesítése (Campbell 1997).

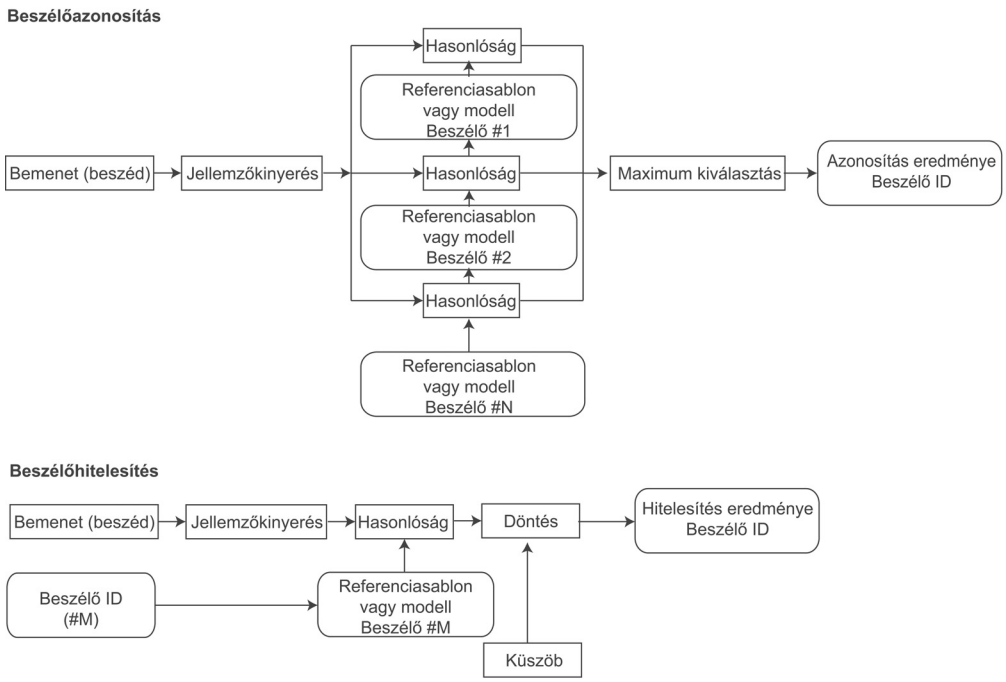


2.4. ábra

A gépi beszédfeldolgozás területei

A beszélőhitelesítés napjainkban egyre inkább megoldottnak tűnik, mivel közel 98–99%-os eredménnyel működik (Bimbot et al. 2011). A beszélőazonosítás eredményei ehhez képest jóval változatosabbak. Az eredmények nagyban függnek a felvétel minőségétől, azaz hogy milyen a jel/zaj viszony, és a beszédminta hosszától. A gyakorlatban legtöbbször igen rövid akusztikailag feldolgozható minta áll rendelkezésre az azonosításhoz (Nikléczy 2001; Nikléczy–Gósy 2008). Kutatások szerint a legrövidebb beszédminta hossza, amely még alkalmas az azonosításhoz, 16 másodperc (Nikléczy–Gósy 2008).

A beszélőfelismerés öt lépésből áll: a beszédjel tisztítása, jellemzőkinyerés, beszélőmodellek létrehozása, mintaillesztés, döntés *(2.5. ábra)*.



2.5. ábra

A beszélőazonosítás (fent) és a beszélőhitelesítés (lent) folyamatábrája

A bemeneti beszédjelből eltávolítjuk azokat a részeket, amelyek nem járulnak hozzá a beszélő személy felismeréséhez, vagy nehezítik azt. Ilyen tipikus eljárás a zajszűrés, beszédjeltisztítás, amely során a beszédből eltávolítjuk a zaj minél nagyobb részét, javítva ezzel a jel/zaj viszonyt. A másik eljárás a beszéddetektálás (voice activity detection), amely során csak azokat a részeket tároljuk el, ahol a beszélő valóban beszél, kiszűrve ezzel a szüneteket, hosszabb lélegzetvételeket, zajos részeket. A beszédjel megtisztítása után számítjuk ki belőle az akusztikai jellemzőket. Az akusztikai jellemzők igen sokfélék lehetnek. A jellemzőkinyerés célja az, hogy megtaláljuk azon akusztikai jellemzőket, amelyek mentén az egyes beszélők elkülöníthetők, azaz beszélőszemély-specifikusak. Az akusztikai jellemzőknek ugyanakkor egyszerűen mérhetőknek, minden beszélőnél jól mérhetőknek, érzelmi állapottól függetlennek kell lenniük.

Jóllehet az akusztikai jellemzők közül az MFCC (Mel-Frequency Cepstral Coefficients) együtthatók az egyik legtöbbet használt jellemzők, továbbra is kérdés maradt, hogy a spektrumban mely frekvenciasáv tartalmazza a beszélőspecifikus jegyeket. Az MFCC-t használó beszélőfelismerő rendszerek eredményei azt mutatták, hogy a spektrumban a 2,5–3,5 kHz közötti sáv az, amely beszélőspecifikus jegyeket hordoz (Furui 1981). Parthasarathi és munkatársai (2009) a beszélődetektáló rendszerükben szintén azt tesztelték,

hogy melyik az a kritikus sáv, amely a beszélőre utalhat. Az eredmények azt mutatták, hogy a vizsgált három tartomány közül (1,5–2,5 kHz; 2,5–3,5 kHz; 3,5–4,5 kHz) a 2,5 kHz és a 3,5 kHz közötti tartományban számolt Mel-frekvenciás kepsztrális együtthatókkal érték el a legjobb eredményt.

A jellemzőkinyerés után előállnak az úgynevezett jellemzővektorok, amelyek alapján elvégezhető az osztályozás. Az osztályozáshoz a beszédfelismerésben használt algoritmusokat szokás alkalmazni (kevert Gauss-modell: Gaussian Mixture Model, GMM; rejtett Markov-modell: Hidden Markov model, HMM; Mesterséges Neuronhálózatok: Artificial Neural Network, ANN; Szupport Vektor Gép: Support Vector Machine, SVM; Döntési fák: Decision tree és ezek kombinációi). A beszédfelismeréshez képest azonban a beszélőfelismerésben a modellek közötti hasonlóság mérését végezzük, ami a referencia-adatbázisban található személyek modelljei és az aktuálisan azonosításra kerülő személy modellje közötti hasonlóság mérését jelenti. Emellett hangsúlyos szerepet kap a konfidenciaszint, vagyis az, hogy mennyire biztos a döntés: elfogadás vagy elutasítás. Az osztályozás mindig két lépésben történik. Az első lépésben a tanulóalgoritmus elsajátítja az egyes személyekre jellemző paramétereket, majd a második lépésben a tanításnál fel nem használt személyekre alkalmazzuk a tanítás során elsajátítottakat.

A korábbi kutatások száma igen jelentős, a jelen munkában csak néhány bemutatására kerül sor. A beszélőfelismerés kiindulásakor a beszédből kinyert jellemzőkből számolt jellemzővektorok között számítottak távolságmértéket különböző távolságfüggvényekkel (vö. *2.1. táblázat)*.

2.1. táblázat
A beszélőfelismerő rendszerek kronológiai áttekintése
(Campbell 1997 alapján);
N: beszélők száma; i: identifikáció; v: verifikáció; s: szekundum, a beszéd hossza

Szerző	**Korpusz**	**Jellemzők**	**Osztályozás**	**Szöveg**	**N**	**Hiba**
Atal 1974	labor	kepsztrum	Mintaillesztés (távolságfüggvény)	függő	10	i: 2%/0,5 s v: 2%/1 s
Markel–Davis 1979	labor	LP	Long term statistics (egyfajta távolságfüggvény)	független	17	i: 2%/39 s
Furui 1981	telefonos	normalizált kepsztrum	Mintaillesztés	függő	10	v: 0,2%/3 s
Schwartz et al. 1982	telefonos	LAR	Nem parametrikus (pdf) valószínűségeloszlás-függvény	független	21	i: 2,5%/2 s
Li–Wrench 1983	labor	LP, kepsztrum	Mintaillesztés	független	11	i: 21%/3 s i: 4%/10 s

Szerző	Korpusz	Jellemzők	Osztályozás	Szöveg	N	Hiba
Doddington 1985	labor	filterbank	DTW: dinamikus idővetemítés	függő	200	v: 0,8%/6 s
Soong et al. 1985	telefon	LP	VQ (64) Valószínűségi arány teszt	izolált szavas	100	i: 5%/1,5 s i: 1,5%/3 s
Higgins–Wohlford 1986	labor	kepsztrum	DTW Valószínűségi arány teszt	független	11	v: 10%/2,5 s v: 4,5%/10 s
Attili et al. 1988	labor	kepsztrum LP auto-korreláció	Projected long term statistics	függő	90	v: 1%/3 s
Higgins et al. 1991	irodai	LAR, LP-kepsztrum	DTW, Valószínűségi arány teszt	függő	186	v: 1,7%/10 s
Tishby 1991	telefonos	LP	HMM (AR mix)	független	100	v: 2,8%/1,5 s v: 0,8%/3,5 s
Reynolds 1995	irodai	Mel-kepsztrum	HMM (GMM)	függő	138	i: 0,8%/10 s v: 0,12%/10 s
Che–Lin 1995	irodai	kepsztrum	HMM	függő	138	i: 0,56%/2,5 s i: 0,14%/10 s v: 0,62%/2,5 s
Colombi et al. 1996	irodai	Kepsztrum, energia első két derivált	HMM (GMM)	független	416	v: 11%/3 s v: 6%/10 s v: 3%/30 s
Campbell et al. 2006	telefonos	LPCC MFCC	SVM (GLDS kernel) GMM GMM-SVM	független	356	v: 6,1%/30 s v: 4,8%/30 s v: 3,2%/30 s
Joshi et al. 2008	telefonos	LPCC	AANN-HMM	független	500	v: 6,69%
Avi–Yuval 2014	labor	LSF és MFCC első két derivált	skew-GMM	független	100	i: 1,5%

A későbbiekben a szabályalapú felismerőket felváltották a statisztikai alapú rendszerek, amelyek hatékonysága jóval meghaladta a szabályalapúakét.

A magyar nyelvre vonatkozóan számos kutatás jelent meg a beszélő személy azonosításának témakörében (Gósy–Nikléczy 1999; Nikléczy 2003; Beke 2008; Bőhm 2006). Igen kevés számban jelent meg azonban kifejezetten a beszélő személy gépi felismerésével foglalkozó tanulmány. Fék Márk (1997) TDK-, majd OTDK-dolgozatában foglalkozott a témával. Dolgozatában LPC-t (lineáris predikciós együtthatót) használt jellemzőként, a mintaillesztéshez pedig vektorkvantálót és különböző neurális hálózatokat (MLP: Multilayer Perceptron, RBF: Radial Basis Function). Eredményei azt mutatták, hogy a beszélőidentifikációs feladatban az MLP, míg a beszélőverifikációs feladatban a vektorkvantáló teljesített jobban.

A következőkben a leggyakrabban használt GMM-UBM beszélőfelismerőt mutatjuk be.

2.5.1. Kevert Gauss-beszélőmodell

2.5.1.1. Kevert Gauss-modell

Ebben a fejezetben bemutatjuk a kevert Gauss-modellt (GMM) és azt, hogy miért használható a beszélők modellezéséhez szövegfüggetlen beszélőfelismerésben. A GMM két okból használható a beszélőfelismeréshez. Egyrészt mert a GMM univerzális eloszlásbecslő (függvényapproximátor) jól kezelhető, nem igényel komplex számítást, mégis pontosan lehet vele becsülni a függvény paramétereit. A függvény paramétereinek becslése alatt a haranggörbe (Gauss-görbe) paramétereinek becslését értjük. A tanító minták alapján iteratív módon becslést végzünk a Gauss-görbe paramétereire; ezt a folyamatot nevezzük betanításnak.

Legyen $X = \{x_1, x_2, \ldots, x_T\}$ egy sor T vektor, amelynek mindegyike a beszédből kinyert d-dimenziós jellemzővektor. Mivel ezen jellemzővektorok eloszlása nem ismert, ezért kevert Gauss-modellekkel szokás modellezni őket, amely súlyozott összege az m komponensű eloszlásnak. A d-dimenziós jellemzővektor matematikai leírása m komponensű kevert Gauss-modellekkel:

$$p(x_i \mid \lambda) = \sum_{i=1}^{m} a_i N(X_t, \mu_i, \Sigma_i),$$

ahol $P(x|\lambda)$ a feltételes valószínűségi értékre vonatkozó becslés, ami azt jelenti, hogy az x jellemzővektor milyen valószínűséggel tartozik a kevert modellbe, λ-ba. A kevert modell m számú Gauss-görbe összege, amelyeket azok várható értékével, a mintaközéppel μ_i és a kovarianciamátrixszal $\sum_i$ parametrizáljuk. A koefficiensek a_i, a keveréket alkotó egyes normális (Gauss) komponensek súlyai. A súlyok pozitívak, és összegüknek 1-nek kell lennie.

A kevert Gauss-modell paramétereinek, a_i, μ_i és $\sum_i$, becslésére számos algoritmus létezik. Ezek közül a legtöbbször alkalmazott eljárás a legnagyobb valószínűség elve (maximum likelihood criterion), amelyet az iteratív Expectation-Maximization (EM) (Dempster et al.

1977; McLachlan–Krishnan 1997) algoritmussal szokás megvalósítani. Általában kevesebb mint 10 iteráció elégséges az EM-algoritmusnak ahhoz, hogy elérje a paraméterek elégséges konvergenciáját. A teljes kevert Gauss-modell valószínűségi eloszlását az összes komponens átlagvektorával, kovarianciamátrixával és a súlyukkal reprezentáljuk. Ezeket a paramétereket együttesen a következőképpen foglalhatjuk egybe:

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\left\{-\frac{1}{2}(x_t - \mu_i)' \sum_i^{-1} (x_t - \mu_i)\right\}}{(2\pi)^{d/2} \left|\sum_i\right|^{1/2}}.$$

A tanításkor az alapvető cél, hogy megbecsüljük a GMM paramétereit, amelyet a tanító adatbázisban lévő beszélő beszédéből kinyert jellemzővektorokból számítunk. Ennek számítására a legnagyobb valószínűség elvét (ML) alkalmaztuk. Az ML célja, hogy olyan modellparamétereket becsüljön, amely maximalizálja a GMM valószínűségét. Megadva a tanító jellemzővektorokat *(T),* a GMM valószínűsége a következőképpen írható le:

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda).$$

Az ML paraméter becslését optimalizálhatjuk iteratív módon az expectation-maximization (EM) algoritmussal. Az algoritmus kezdeti állapotként veszi a modell $\bar{\lambda}$ értékeit, majd kiszámítja az új modell $\bar{\lambda}$ értékeit úgy, hogy teljesüljön

$$p(X|\bar{\lambda}) \geq p(X|\lambda).$$

Az új modell számolása akkor kezdődik, amikor ismertté válik az előző modellből számolt λ értéke. Ez az eljárás akkor fejeződik be, amikor egy bizonyos konvergenciaküszöböt átlép a rendszer.

Minden egyes iterációval újrabecsüljük (reestimation) a GMM paramétereit. A komponensek súlyainak újraszámolása a követezőképpen történik:

$$\bar{a}_i = \frac{1}{T} \sum_{t=1}^{T} p(i|x_t, \lambda).$$

Az átlag, μ újraszámolása:

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)}.$$

A kovarianciamátrix, $\sum$ újraszámolása:

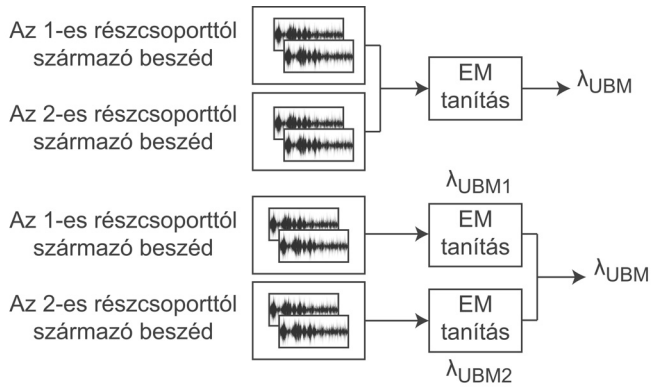
$$\bar{\Sigma}_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda)(x_t - \mu_i)(x_t - \mu_i)'}{\sum_{t=1}^{T} p(i|x_t, \lambda)}.$$

Az *a posteriori* valószínűségét az *i*-edik beszélőszemély-modellnek a következő egyenlettel számolhatjuk:

$$p(i|x_t,\lambda)=\frac{p_i b_i(x_t)}{\sum_{k=1}^{N} p_k b_k(x_t)}.$$

2.5.1.2. Univerzális háttérmodell

Számos kutatás kimutatta, hogy a beszélőfelismerés eredménye jelentősen javítható, hogyha a beszélő valószerűségi értékét normalizáljuk egy általános háttérmodellből származó valószerűségi értékkel (Higgins et al. 1991; Rosenberg et al. 1992; Reynolds 1995; Matsui–Furui 1995; Reynolds 1997; Reynolds–Rose 1995). A GMM-UBM-alapú beszélőfelismerő rendszer (kevert Gauss-modell és általánosított háttérmodell: Gaussian Mixture Model-Universal Background Model) egy beszélőfüggetlen háttérmodellt alkalmaz, amelyet a következőképpen reprezentálunk: $p(X|\lambda_{\overline{hyp}})$. A UBM egy nagy adatbázison tanított modell, amely a jellemzők beszélőfüggetlen eloszlását reprezentálja. A UBM használatával speciálisan meghatározunk olyan körülményeket a beszédre, amelyeket folyamatosan figyelembe veszünk a felismerés folyamán. Ez a beszélők fogalmazásától kezdve, a beszéd típusán át, a beszéd minőségére is vonatkozhat. Például a jelen kutatásban *a priori* tudjuk, hogy a beszédjel egy igen jó minőségben rögzített jel, és hogy az adatbázisban mindkét nem szerepel. Ezért a UBM tanításakor egy olyan univerzális modellt hozunk létre, amely jó minőségű beszédet és azonos eloszlású női és férfi populációt tartalmaz. A UBM létrehozásában azonban nincsenek egységes irányelvek, sem objektív mérőeszköz annak meghatározására, hogy hány beszélőre és milyen hosszú beszédre tanítsuk a UBM-et. Az adatok megadására a UBM tanításához sokféle módszer létezik. A legegyszerűbb az, amikor az összes tanító adatbázisban lévő beszélőt felhasználjuk a UBM kialakítására, amit EM-algoritmussal optimalizálunk. Ekkor azonban figyelni kell arra, hogy az egyes részcsoportok előfordulása egyenlő legyen, például a nők és férfiak száma. Egy másik megközelítésben a részcsoportokra külön-külön készítenek el egy-egy UBM-et, majd azt egyesítik (Reynolds et al. 2000) (vö. *2.6. ábra*).



2.6. ábra
A UBM kétféle tanítási módszere

2.5.1.3. A beszélőegyezés mérése

A beszélőazonosításra a valószerűségi arány tesztet (likelihood-ratio test) szokás alkalmazni az azonosítandó beszédekre. Ebben a részben a többgaussos valószerűségi arány tesztet (Multi-Gaussian log-likelihood test) írjuk le.

A $f(X|\lambda_C)$ az a valószínűség, ami egy azonosítandó megnyilatkozáskor fennáll, amelyből λ_C-t a következőképpen számoljuk:

$$\ln f(X|\lambda_C) = \frac{1}{N}\sum_{t=1}^{N} \ln f(x_i|\lambda_c) = \frac{1}{N}\sum_{t=1}^{N} \ln \sum_{i=1}^{M} \left(\frac{m_{Ci}}{2\pi^{d/2}\left|\sum_{Ci}\right|^{1/2}} \right) \exp^{-\frac{1}{2}(x_t-\mu_{Ci})^T \sum_{Ci}^{-1}(x_t-\mu_{Ci})},$$

ahol m_{Ci}: i-edik súly, μ_{Ci}: átlagvektor, $\sum_{Ci}$: a kovarianciamátrixa az azonosítandó beszélői modellnek, λ_C-nek.

A beszélő azonosításakor a háttérmodell egy részét a beszélőmodellekből kell előállítani. Ebben a vizsgálatban mind a 20 beszélőből készítettük el a háttérmodellt. Azt a valószínűséget, ami nem az egyes beszélőktől származik, hanem a tanító adatbázisban szereplő beszélőktől, általános háttérmodellnek hívjuk (Universal Background Model), és a következőképpen számoljuk:

$$\ln f(X|\lambda_C) = \frac{1}{B}\sum_{b=1}^{B} \ln f(X|\lambda_b) = \frac{1}{B}\sum_{t=1}^{B} \left(\frac{1}{N}\sum_{t=1}^{T} \ln \sum_{i=1}^{M} \left(\frac{m_{bi}}{2\pi^{d/2}\left|\sum_{bi}\right|^{1/2}} \right) \exp^{-\frac{1}{2}(x_t-\mu_{bi})^T \sum_{bi}^{-1}(x_t-\mu_{bi})} \right),$$

ahol m_{bi}: i-edik súly, μ_{bi}: átlag vektor, $\sum_{bi}$: kovarianciamátrixa a háttérmodellnek, λ_b-nek.

Legyen λ_k, $k = 1, \ldots, N$, ahol az N az egyes beszélők beszélői modellje. Adott a jellemzővektorok sorozata, X. Az osztályozó kialakításakor ezt a jellemzővektor-sorozatot feleltetjük meg az N beszélői modellnek, felhasználva N diszkriminanciafüggvényt, $g_k(X)$, kiszámítva a hasonlóságot az ismeretlen X és az összes beszélői modell között, λ_k. A λ_{k*} modell akkor kerül kiválasztásra, ha

$$k^* = \arg \max_{1 \le k \le N} g_k(X) .$$

A minimumhibaarány-osztályozóban a diszkriminanciafüggvény az *a posteriori* valószínűség:

$$g_k(X) = p(\lambda_k | X) .$$

Felhasználva a Bayes-tételt:

$$p(X | \lambda_k) = \frac{p(\lambda_k) p(X | \lambda_k)}{p(X)},$$

és feltételezve, hogy a beszélők valószínűsége egyenlő, más szóval $p(\lambda_k) = 1/N$. Megjegyezve, hogy $p(X)$ azonos minden beszélői modell esetében, így a fent leírt diszkriminanciafüggvény ugyanaz, mint a következő egyenlet:

$$g_k(X) = p(X | \lambda_k) .$$

Végül felhasználva a log-likelihood függvényt, a döntési szabály a beszélőfelismerésre a következő: azonosított beszélő k^*, ha

$$k^* = \arg \max_{1 \le k \le N} \sum_{t=1}^{T} \log p(x_x | \lambda_k) ,$$

ahol $p(x_t | \lambda_k)$-t a következő egyenlet adja:

$$p(x_t | \lambda_k) = \sum_{i=1}^{m} a_i N(x_t, \mu_i, \Sigma_i) .$$

2.6. Az egyszerre beszélés detektálása

Az olvasott beszédre (például újságfelolvasás, hírbemondás, időjárás-jelentés) már léteznek olyan felismerő rendszerek (speech-to-text), amelyek legalább 90%-os pontossággal alakítják át a beszédet folyamatos írott szöveggé. A beszédfelismerő rendszerek eredményei a monologikus spontán beszédben azonban már romlanak (Furui 2007; Mihajlik 2010). Az eredmények

romlását az okozza, hogy az akusztikai és a nyelvi modelleket általában az írott nyelvtan szabályaiból és a felolvasott szövegek nyelvéből építik ki. Az akusztikai modelleket gyakran olvasott anyagon készítik, mivel kevés a spontán korpusz. Továbbá a spontán beszéd akusztikuma igen heterogén, és a beszédfelismerést számos más tényező is nehezíti (megakadások, atipikus realizációk stb.). A társalgás a spontánbeszéd-technológia speciális esete, mivel a gépi beszédfelismerő rendszerek számára nehezebb az olyan beszédtípusok dekódolása, ahol több beszélő társalog egymással. Ezért megnőtt az igény a gépi beszélődetektálásra is. A társalgás során a monologikus beszédre jellemző akusztikai és nyelvtani szabályok nagyszámú varianciája mellett újabb nehézségek jelennek meg. Ezek lehetnek a társalgást jellemző egységek, mint például a beszédforduló, az egyszerre beszélés, a nonverbális jelek (nevetés) stb., ezért a beszélődetektáláskor valamennyiük modellezésére szükség van (Boakye et al. 2008, 2011; Zelenák et al. 2010).

Az egyszerre beszélések aránya a spontán társalgásokban meglehetősen nagynak mondható (Bata–Gráczi 2009). Beattie a beszélőváltásokat elemezve (1982, idézi Levelt 1989) kimutatta, hogy a két résztvevős angol társalgásban 11%-ban fordul elő egyszerre beszélés (azaz a beszédpartner közbevág), több beszélőnél ez az arány már 31%. Az újabb kutatások ezeket az arányokat igazolták. Çetin és Shriberg (2006a, 2006b) angol korpuszokat vizsgálva adatolta, hogy az átfedő beszéd átlagosan 10–13%-át teszi ki a társalgásoknak. A hazai kutatásokban Markó (2006) 6%-ot állapít meg a teljes beszéd és az átfedő beszéd arányaként négybeszélős spontán társalgásban. Bata (2009b) 1,7–3%-ot adatolt kutatásában, amit spontán társalgásokban elemzett. Ez a magas előfordulási szám az átfedő beszéd funkciójából adódik. A társalgás során ugyanis az egyszerre beszélés kettős funkciót tölt be. Egyrészt megerősítő szerepe van (háttércsatorna-jelzés, például *igen, aha, ühüm*), másrészt versengő funkciójú, amikor a társalgás egyik szereplője át kívánja venni a szót, és már azalatt elkezdi a beszédét, mialatt az aktuálisan beszélő még nem fejezte be a mondanivalóját (Iványi 2001; Hámori 2006; Bata 2009a).

Az egyszerre beszélések vizsgálata sokrétű (Çetin–Shriberg 2006a; 2006b). Az átfedő beszéd több szempontból is jelentős. A diskurzuselemzésben fontos kérdés, hogy mikor következik be az egyszerre beszélés a társalgó felek szociális viszonyaitól, ismertségi fokától és egyéb tényezőktől függően, és hogy ezek az átfedő részek milyen szintaktikai, pragmatikai, illetve fonetikai formában jelennek meg. Fontos szerepük van továbbá a spontán beszéd automatikus felismerésében is, hiszen az egyszerre beszélések a gépi beszédfelismerés számára korlátozottan feldolgozható szakaszai a beszédnek. A beszélődetektálásban a beszélői modell kialakítása során az átfedő beszédrészek mint zaj jelentkeznek. Ez azért lehetséges, mivel az átfedő részekben nem csak egy beszélő jelenik meg akusztikailag, ami az egyes beszélői modellek egységességét gyengítheti, csökkentve ezzel a végleges beszélődetektálási eredményt. Ezért elengedhetetlen, hogy az átfedő részek gépi úton automatikusan azonosíthatók legyenek.

Az elmúlt évtizedekben megnőtt a spontán társalgásokat tartalmazó korpuszok száma (Gósy 2012). Ezen korpuszok felvételi körülményeit tekintve kétfelé oszthatók: egycsatornás, illetve többcsatornás. Ez azt jelenti, hogy a spontán társalgásokban *a)* minden egyes beszélőtől

bejövő jelet külön csatornára vesznek fel, illetve *b)* minden egyes beszélő beszédét egy csatornára rögzítik. Ez az alapvető felépítés meghatározza az egyszerre beszélések automatikus osztályozásának beszédtechnológiai eszközeit. A legtöbb kutatásban többcsatornás felvételeket elemeznek (Yamamoto et al. 2006; Laskowski–Schultz 2006; Xiao et al. 2011). Lényegesen nehezebb feladat azonban, amikor egycsatornás felvételben kell osztályoznunk az egyszerre és a nem egyszerre beszéléseket.

Az egyszerre beszéléseket modellező munkák száma relatíve kevés, és azok közül is csak néhány kutatásban igazolták, hogy csökkenti a beszélődetektálási hiba arányát (DER) (Boakye et al. 2008; Boakye 2008; Trueba-Hornero 2008; Zelenak et al. 2010; Xiao et al. 2011).

Az automatikus beszélődetektálás során kimutatták, hogy a legtöbb hiba szignifikánsan azon részeken történik a felvételekben, ahol egyszerre beszélés található. Wooters és Huijberts (2007) munkájukban azt írták le, hogy a beszélődetektálási hiba arányának 17%-át a téves elutasítások száma adja, amit az átfedő beszédrészek okoznak.

Az egyszerre beszélések automatikus detektálására történt vizsgálatok közül Moattar és Homayounpour (2006) a társalgásban megjelenő egyszerre beszélést a hang periodicitásából ítélték meg. A vizsgálat során azt figyelték meg, hogy ahol a beszéd nem mutatott periodicitást a Fourier-spektrumban, ott jelent meg az egyszerre beszélés. Boakye és munkatársai (2008) kimutatták, hogy az átfedő beszédet MFCC és más akusztikai paraméterekkel GMM/HMM-mel modellezve 7,4%-ban csökkenteni lehetett a detektálási hiba arányát a beszélőazonosításban. Ugyancsak Boakye és munkatársai (2011) amerikai angol spontán társalgási korpuszban vizsgálták az átfedő beszédrészek automatikus osztályozhatóságát a beszélődetektáló rendszerek javítása érdekében. Akusztikai jellemzőként MFCC-t, RMS-energiát (beszédjel energiája), LPC-analízist (lineáris predikciós együttható) és még számos más, a zöngeminőséget jellemző eljárást alkalmaztak. Ezeket dimenziócsökkentették, és GMM-mel mintaillesztették. A hasonlóság méréséhez Kullback–Leibler-távolságot számoltak. Ezzel az eljárással kimutatták, hogy szignifikánsan csökkenthető a tévesztési arány a beszélődetektálás során a spontán társalgásokban.

Otterson és Ostendorf (2007) munkájukban elméleti megközelítésben kimutatták, hogy az átfedő beszéd osztályozásával javítani lehet a beszélődetektálás eredményét. Az általuk létrehozott osztályozót azonban nem tesztelték beszélődetektálóban. Trueba-Hornero (2008) munkájában már egy valós átfedőbeszéd-detektálót hozott létre, és tesztelt beszélődetektálóban. A legtöbb munka azonban nagyon magas hibaértékekről számol be, ami mutatja a feladat nehézségét (Boakye et al. 2008; Boakye 2008). Ezen alkalmazások HMM-GMM-et használnak, amelyben három modellt hoznak létre: nembeszéd, nem átfedő beszéd és átfedő beszéd. Az eredmények azt mutatták, hogy a legjobb eredményük alapján a pontosság (precision) 58%, míg a fedés (recall) 19% volt. Az alacsony pontossági és fedési értékek mellett is 10%-os relatív DER-csökkenést tudtak elérni az átfedő beszédrészek detektálásával.

Becslések szerint azonban az ideális egyszerre beszéléseket detektáló algoritmussal a DER 37%-kal lenne csökkenthető, ezért ezen a területen még igen sok fejlesztésre van szükség.

A jelen kutatás célja, hogy a spontán társalgásokban modellezze az egyszerre beszéléseket, és automatikus osztályozó algoritmussal különítse el azoktól a beszédszakaszoktól, ahol csak egy társalgó beszél. Hipotézisünk szerint az átfedő beszéd jellegzetes akusztikai szerkezettel rendelkezik, ezért létrehozható egy automatikus osztályozó algoritmus. Ugyanakkor feltételezzük, hogy a háttércsatorna-jelzések okozzák majd a legtöbb hibát az osztályozáskor.

Az egyszerre beszélések automatikus osztályozása jóllehet egyszerű feladatnak tűnik, megvalósítása korántsem triviális. Ez a beszélődetektálás egyik alapfeladata, mégis csak néhány olyan tanulmány ismert, amely megfelelő eredménnyel tudta megvalósítani az egyszerre beszélések automatikus osztályozását (vö. Boakye et al. 2008).

A jelen kutatásban egy ANN/SVM (Artificial Neural Network/Support Vector Machine, mesterséges neuronháló/szupport vektor gép) hibrid rendszert hoztunk létre az egyszerre beszélések automatikus osztályozásához.

Az osztályozás során az első lépés a lényegkiemelés, amelynek fő feladata, hogy a beszédjelből olyan információkat vonjunk ki, amelyekkel jól megragadhatók az egyszerre beszélések. Mivel nem ismert, hogy mely akusztikai paraméter mentén különülnek el az átfedő és a nem átfedő beszédrészek, több akusztikai jellemzőt is teszteltünk, mint például az FFT-spektrum, MFCC, Mel-skála szerinti logaritmikus szűrőbank (MSL), részsávenergia (subband-energy). A jellemzők jobb reprezentálásához főkomponens-analízist (PCA: Principal Component Analysis) használtunk, amely növeli az osztályozó eredményét.

3. A kutatás célja, kutatási kérdések és hipotézisek

3.1. Kutatási kérdések

A kutatás egyik fő kérdése az volt, hogy milyen eredménnyel tudjuk megvalósítani a beszélődetektálót magyar nyelvű spontán társalgásokra. Hogyan valósíthatók meg a beszélődetektálás egyes előfeldolgozó rendszerei, mint a beszéddetektálás, egyszerre beszélés detektálása, illetve hogy ezek milyen eredménnyel implementálhatók a beszélődetektáló rendszerbe. Arra is kerestük a választ, hogy melyek azok az akusztikai jellemzők, amelyek az egyénre jellemző akusztikai lenyomatokat tartalmazhatják. Vizsgáltuk, hogy milyen eredménnyel lehet az egyszerrebeszélés-detektálót megvalósítani. Elemeztük, hogy a beszélőszegmentálásban milyen beállítások mellett kapjuk a legjobb eredményt.

3.2. A kutatás célja

A kutatás fő célja, hogy elsőként nagy mennyiségű magyar nyelvű spontán társalgás felhasználásával hozzon létre egy felügyelet nélküli tanuláson alapuló beszélődetektáló algoritmust. A kutatás fő motivációja az volt, hogy spontán társalgásokra valósítsunk meg beszélődetektálót, mivel az eddigi beszélődetektálók híradós adásokra vagy telefonhívásokra készültek. A híradós adások tipikusan felolvasásokat vagy előre megtervezett beszédeket tartalmaznak, amelyekben a társalgó felek kerülik az egyszerre beszélést. A telefonos felvételek pedig általában dialogikus beszélgetéseket jelentenek, amelyek többsége csak két személy interakciójából áll. A beszélődetektálás megvalósítása igen nehéz feladat mind a híradós felvételekre, mind a telefonos hívásokra. A legnagyobb kihívást azonban a spontán társalgások beszélőkre való bontása jelenti. Ez abból fakad, hogy ebben a beszédstílusban fordul elő a legtöbb egyszerre beszélés, a beszédfordulók megvalósulási formái változatosak, sokszor igen rövidek, illetve számos más jelenséget is tartalmaz, mint a nevetés, köhögés, zaj stb. A kutatás célja az volt, *i)* hogy az automatikus gépi beszélődetektálóhoz szükséges algoritmusokat elkészítsük (beszélőszegmentáló és beszélőklaszterező algoritmus, egyszerrebeszélés-detektáló), illetve a már rendelkezésre állókat implementáljuk a rendszerbe (beszéddetektáló, beszélőfelismerő algoritmus). További célja az volt, *ii)* hogy vizsgáljuk, milyen sikerrel lehet implementálni a beszélődetektálóba a beszéddetektáló és az egyszerre beszélést detektáló algoritmusokat.

Célunk volt az is, *iii)* hogy megállapítsuk, a beszélődetektálóban milyen akusztikai paraméterekkel lehet a legjobb eredményt elérni. Mindezen algoritmusokat a MATLAB (2011a) szoftverben írtuk és futtattuk.

3.3. A kutatás hipotézisei

A kutatás elején a következő hipotéziseket fogalmaztuk meg:

1. A beszédfelismerésben a spektrum célzott részsávjára történő akusztikai jellemzőkinyerés jobb eredményeket adhat, mint a teljes spektrumot feldolgozó eljárások.
2. A beszélődetektálásban kikísérletezett akusztikai jellemzők jól alkalmazhatók a beszélőszegmentálásban, illetve a beszélőklaszterezésben.
3. A beszéddetektálás implementációjával a beszélődetektálás eredményei növelhetők.
4. Az egyszerrebeszélés-detektáló implementációjával a beszélődetektálás eredményei növelhetők.

4. Kísérleti személyek, általános anyag és módszer

4.1. Anyag és kísérleti személyek

A kutatáshoz a BEA adatbázist használtuk (Gósy 2012). A BEA adatbázis az MTA Nyelvtudományi Intézet Fonetikai Osztályának munkája. Az adatbázis fejlesztésének fő célja az, hogy nagyszámú magyar anyanyelvű adatközlőtől rögzítsen különféle beszédstílusban beszédfelvételeket. Az adatközlők egynyelvű budapesti felnőttek, életkoruk 20 és 70 év közötti. Minden adatközlőtől rögzítve vannak a következő beszédstílusok: mondatolvasás, szövegolvasás, mondatvisszamondás, tartalomösszegzés, spontán narratíva és társalgás. A felvételi körülmények állandók, mindig azonos helyen és körülmények között történnek, csendesített helyiségben. A rögzítés digitális, közvetlenül számítógépre történik 44,1 kHz-es mintavételezéssel (tárolás: 16 bit, monó).

A BEA adatbázisból 100 társalgást választottunk ki, amely 55 órányi hanganyagot jelent. A társalgásokban minden esetben három személy vett részt. Ebből két társalgó állandó volt (2 nő, életkoruk 32 év). A harmadik személy (adatközlő) 43 férfi és 67 nő közül került ki, átlagos életkoruk 35 év.

A felvétel minősége laboratóriumi körülményekhez hasonló. A felvételt egy Audio-Technica AT 4040 típusú mikrofonnal egy csatornára rögzítették 44,1 kHz-en, amelyet újra-mintavételeztünk 16 kHz-en. A BEA alapvető céljának megfelelően az adatközlőhöz volt legközelebb a mikrofon, így az ő beszédjele volt a legerősebb, míg a kísérletvezető, illetve egy másik bevont személy beszédjele gyengébb volt. Ez megnehezítette az egyes algoritmusok kialakítását. Lehetőség lett volna normalizációs eljárásokat használni, de ez feltehetően a zajt is felerősítette volna, ezért ilyen jellegű kompenzációt nem alkalmaztunk.

A társalgások annotációi a következőket tartalmazták:

i) Szünetek: minden olyan szünetet jelöltünk, amely meghaladta a 100 ms-ot. Nyilvánvalóan a zöngétlen zár- és zár-rés hangok artikulációjából adódó néma fázisokat nem jelöltük még akkor sem, ha azok ezen küszöböt átlépték.

ii) Beszélőváltások: a folytonos jelben bejelöltük, hogy mely időpillanatban van beszélőváltás, illetve hogy az egyes beszédszegmensek mely beszélőhöz tartoznak. A hátércsatorna-jelzéseket (például *ühüm, ja* stb.) nem tekintettük beszélőváltásnak, csak abban az esetben, ha tényleges szóátvételről volt szó.

iii) Egyszerre beszélések: bejelöltük a beszédnek azon részeit is, ahol egy időben két vagy három személy szólalt meg. Nem jelöltük azonban azon részeket, ahol az átfedő beszéd nem haladta meg az 50 ms-ot, mivel ezek detektálása nem megvalósítható.

4.1.1. A beszélődetektáló kiértékeléséhez használt korpusz

A BEA adatbázisból 12 társalgást választottunk ki random módszerrel. A 12 társalgás összidőtartama közel 2,8 óra. A 2,8 órányi társalgásban 490 beszélőváltás történt *(4.1. táblázat)*. Ezeket a felvételeket csak arra használtuk, hogy az ezen kívüli felvételeken elkészített rendszert teszteljük.

4.1. táblázat
A beszédfordulók száma és teljes időtartama az egyes tesztfájlokra

A felvétel sorszáma	A beszédfordulók száma (db)	A teljes időtartam (s)
bea071n037	55	919,5
bea072n038	46	1020,4
bea073n039	23	590,5
bea074n040	25	1053,3
bea075n041	16	887,6
bea094f039	31	799,5
bea150n091	32	769,7
bea166f066	50	982,4
bea174n105	46	773,0
bea184n111	48	599,4
bea189n114	68	973,1
bea192f077	50	816,2

4.1.2. A beszélőspecifikus jellemzők kialakításához használt korpusz

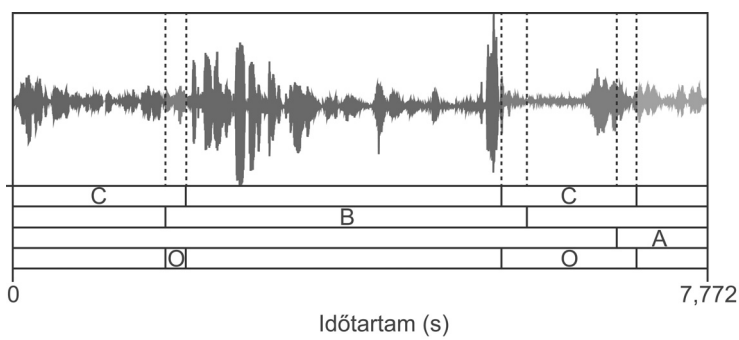
A BEA adatbázisból választottunk ki 100 középkorú beszélőt (42 férfi és 58 női adatközlő). A tanító adatbázishoz minden beszélő beszédéből kivágtunk egy 25 másodperces részt. A tanító adatbázison az algoritmus elsajátítja az adott minták tulajdonságát, amit majd a teszt-adatbázison tesztelünk, hogy ez mennyire volt sikeres. A tesztadatbázishoz minden beszélő beszédéből kivágtunk egy 13 másodperces részt. A rendszer tanításához 80 beszélő 25 s-os beszédmintáit használtuk. A tesztelést 13 s-os beszédmintán végeztük el. A tanítás során minden egyes beszélőre külön modellt hoztunk létre. Az általános háttérmodell (UBM) kialakításához a tanító adatbázisból másik 20 adatközlő 25 s-os beszédét használtuk fel.

4.1.3. A beszéddetektálóhoz használt korpusz

A társalgásokban manuálisan jelöltük azokat a részeket, ahol valamelyik adatközlő beszél, illetve azokat a részeket, ahol nincs beszédjel, vagyis néma szünet van. A korpusz 49 órányi beszédrészt és 6 órányi szünetet tartalmaz, vagyis a teljes korpusz 10,9%-át a szünetek teszik ki.

4.1.4. Az egyszerrebeszélés-detektálóhoz használt korpusz

A társalgásokban manuálisan jelöltük azokat a részeket, ahol egyszerre több adatközlő beszél, illetve azokat a részeket, ahol csak egy beszélő beszél *(4.1. ábra)*.



4.1. ábra
Az átfedő beszéd illusztrálása
(A, B, C: beszélők, O: egyszerre beszélés)

A 100 beszélő spontán társalgásaiban összesen 8056 olyan időintervallum található, ahol kettő vagy annál több résztvevő szólal meg egyszerre, vagyis ahol átfedő beszéd van. Ezen intervallumok összidőtartama közel 7 óra, ami a teljes korpusz 12%-a.

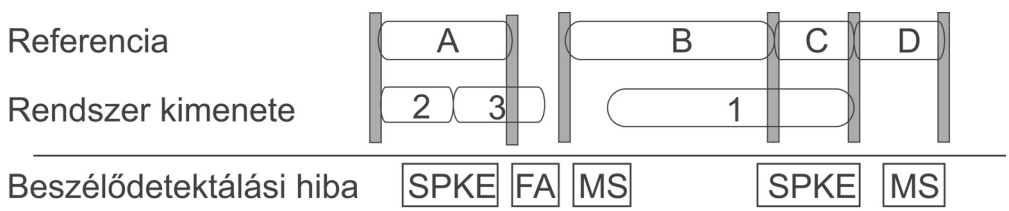
4.2. Kiértékelési módszer

A jelen kutatásban kétféle kiértékelő rendszert alkalmaztunk. A beszéddetektálás, beszélőszegmentálás és a beszélőklaszterezés kiértékeléséhez a NIST által javasolt DER (beszélődetektálási hibaarány, Diarization Error Rate) módszert használtuk. Az egyszerre beszélés kiértékeléséhez pedig a kétosztályos kiértékelési metrikát alkalmaztuk, amely a DET (Detection Error Tradeoff).

4.2.1. Beszélődetektálási hibaarány (DER, Diarization Error Rate)

A beszélődetektálás kiértékeléséhez a NIST munkatársai által fejlesztett DER-algoritmust használtuk. A DER-t tulajdonképpen úgy értelmezzük, mint azt a törölt időt, amely nem kategorizálható helyesen sem beszélőnek, sem nembeszédnek. Ennek mérésére az MD-eval-v12.pl-t (NIST MD-eval-v12 DER kiértékelő szkriptje 2006) használtuk.

Mivel a váltási pontok meghatározása a feladat, a rendszer hipotéziseként a beszélődetektálás kimenetében nem kell explicite meghatározni a beszélő nevét vagy identitását, ezért a beszélőkhöz rendelt azonosító címkéknek nem kell azonosnak lenniük a bemeneti (kézi) címkében és a kimeneti (automatikus) címkében. Ez a feladat tehát nem olyan, mint a beszéd/nembeszéd automatikus címkézése, amely során a szegmenset azonosító címkének egyeznie kell a bemeneti és a kimeneti címkében *(4.2. ábra).*



4.2. ábra

A DER kiértékelési metódus sematikus ábrázolása

(SPKE: beszélőhiba, FA: téves riasztások száma, MS: téves elutasítások száma)

A kiértékelő szkript először megtalálja az optimális egy az egyben átfedést az összes beszélői címke azonosítóira a referencia- és az automatikus címke között. Ez teszi lehetővé az egyezés mérését a különböző azonosítóval rendelkező két címkesor között. A DER értékét a következőképpen számoljuk:

$$DER = \frac{\sum_{s=1}^{S} dur(s) \cdot \left(\max\left(N_{ref}(s), N_{hyp}(s) \right) - N_{correct}(s) \right)}{\sum_{s=1}^{S} dur(s).\, N_{ref}},$$

ahol az *S* az összes beszélői szegmens száma, és ahol mind a hipotetikus, mind a referenciacímke tartalmazza ugyanazt a beszélőt. Ezt úgy kapjuk meg, hogy összehasonlítjuk a hipotetikus, illetve a referencia-beszédfordulókat. A N_{ref} és N_{hyp} kifejezések a beszélők számát jelölik a beszédszegmensben, *s* és $N_{correct}$ a beszélők számát mutatja, amely a helyes találatokat jelenti a referencia- és a hipotetikus címkesor között. A címkesorban a nembeszéd részeket 0 beszélőnek jelölik. Ha mind a beszélők, mind a nembeszéd szegmensek helyesen lettek azonosítva, akkor a hiba értéke 0. A DER tulajdonképpen különböző módon létrejött hibák összege:

1. A beszélőhiba *(E_{SPKR}):* a helytelenül azonosított beszélői azonosítók a teljes időtartam arányában. Ez a típusú hiba nem veszi figyelembe a beszélők átfedését vagy bármilyen

más hibát, ami a nembeszéd részek azonosításából fakad. Ezt a következőképpen írhatjuk fel:

$$E_{Spkr} = \frac{\sum_{s=1}^{S} dur(s) \cdot \left(\min\left(N_{ref}(s), N_{hyp}(s) \right) - N_{correct}(s) \right)}{T_{score}},$$

ahol a $T_{score} = \sum_{s=1}^{S} dur(s) \cdot N_{ref}$ a teljes időtartama a kiértékeléshez használt fájloknak.

2. A téves riasztások száma *(E_{FA}):* a teljes időtartamra vetítve a referenciacímkében a nembeszéd szerepel, de az automatikus címkesorban beszélőnek azonosított a szegmens. A következőképpen írhatjuk fel:

$$E_{FA} = \frac{\sum_{s=1}^{S} dur(s) \cdot \left(N_{ref}(s) - N_{hyp}(s) \right)}{T_{score}} \forall \left(N_{hyp}(s) - N_{ref}(s) \right) > 0,$$

amit csak azon szegmensekben mérünk, amely a referenciacímkében nembeszéd részként szerepel.

3. A téves elutasítások száma *(E_{MISS}):* a teljes időtartamra vetítve a referenciacímkében a beszélő szerepel, de az automatikus címkesorban nembeszédnek azonosított a szegmens. A következőképpen írhatjuk fel:

$$E_{MISS} = \frac{\sum_{s=1}^{S} dur(s) \cdot \left(N_{ref}(s) - N_{hyp}(s) \right)}{T_{score}} \forall \left(N_{ref}(s) - N_{hyp}(s) \right) > 0,$$

amit csak azon szegmensekben mérünk, amely a hipotetikus címkében nembeszéd részként szerepel.

4. Egyszerre beszélések *(E_{ovl}):* a teljes időtartamra vetítetve, amikor több beszélő beszél egy szegmensben, amely nem tartozik egy beszélőhöz sem. Ez a fajta hiba általában az E_{MISS}-hez vagy az E_{FA}-hoz tartozik. Ez a hiba attól függ, hogy a referencia- vagy a hipotetikus címkesorban szerepel-e az egyszerre beszélés. Ha mindkettőben, akkor E_{SPKR}-hez tartozik.

Felírva az összes lehetséges hibát, a DER a következőképpen áll össze:

$$DER = E_{Spkr} + E_{MISS} + E_{FA} + E_{ovl}.$$

Amikor a kiértékelést végezzük, egy olyan időbeli határsávot használunk minden referenciában lévő beszédfordulóra, amely bizonyos pontatlanságot enged meg az automatikus címkézésnek. A NIST ([Amerikai] Nemzeti Szabványügyi Hivatal) ezt az időbeli határsávot ±250 ms-ban határozta meg. A NIST DER szkript kiértékelő megadja minden egyes referenciahipotetikus szegmentációra a DER értékét, illetve az összes kiértékeléshez használt fájlra ad egy súlyozott átlagot.

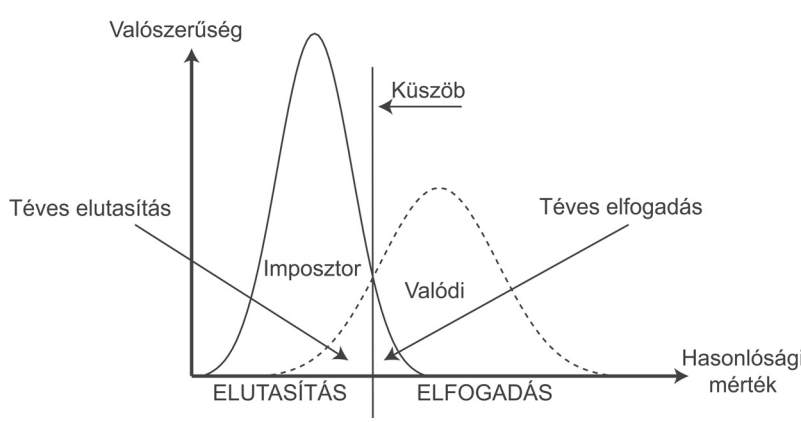
4.2.2. További kiértékelési technikák (DET: Detection Error Tradeoff)

Az osztályozásra alkalmazott algoritmusok működésének kiértékelésére és összehasonlítására a DET (Detection Error Tradeoff) kiértékelő algoritmust használtuk. A DET kiértékeléséhez először bemutatjuk a tévesztési mátrixot a bináris osztályozás esetén *(4.2. táblázat)*.

4.2. táblázat
A tévesztési mátrix a bináris osztályozás esetén

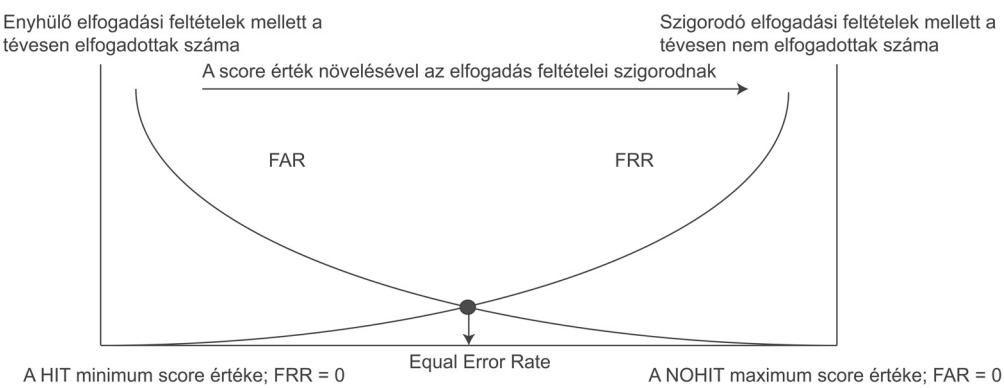
		Aktuális feltétel	
		Pozitív	**Negatív**
Teszt eredménye	**Pozitív**	A feltétel teljesül + pozitív teszt = TP (True Positives)	A feltétel nem teljesül + pozitív teszt = FP (False Positives)
	Negatív	A feltétel teljesül + negatív teszt = FN (False Negatives)	A feltétel nem teljesül + negatív teszt = TN (True Negatives)

A bináris osztályozáskor megkülönböztetünk első- és másodfajú hibát. Az elsőfajú hiba a téves elfogadás (False Acceptance Rate: FAR; False Positives). Jelen munka során téves elfogadásról akkor beszélnénk, ha a beérkező szegmens nem átfedő beszéd, de annak fogadja el a gép. A másodfajú hiba a téves elutasítás (False Rejection Rate: FRR; False Negatives) *(4.3. ábra)*. A jelen munka során téves elutasításról akkor beszélnénk, ha a beérkező szegmens átfedő beszéd, de nem annak minősíti a gép.



4.3. ábra
A bináris osztályozáskor fellépő hibák sematikus ábrázolása

Az osztályozó egy-egy összehasonlítás során a hangmodelleket összeveti az aktuális jellemzőkkel, és mintánként egy hasonlósági számot képez (score), aztán sorba állítja az eredményt a csökkenő score szerint, és döntést hoz, hogy az első helyen levő találat-e vagy sem. A küszöbérték (threshold) alapján döntenek a találatról: ha az első score (érték) alacsonyabb a küszöbértéknél, akkor nincs találat (NOHIT), ha magasabb, akkor van találat (HIT). Ekkor felmerül az a kérdés, hogy milyen küszöbértéket állítsunk be, hogy az osztályozás a lehető legjobb legyen. Ennek megoldására léteznek különböző technikák, mint a ROC (Receiver Operating Characteristic) vagy a DET (Detection Error Tradeoff). A DET-ben úgy választjuk meg a küszöbértéket, hogy az elsőfajú hiba és a másodfajú hiba egyenlő legyen. Ezt úgy hívják, hogy Equal Error Rate (EER) *(4.4. ábra)*.



4.4. ábra

Az EER sematikus ábrázolása

5. Beszélődetektálás társalgásokban

5.1. A korpusz általános statisztikai jellemzői

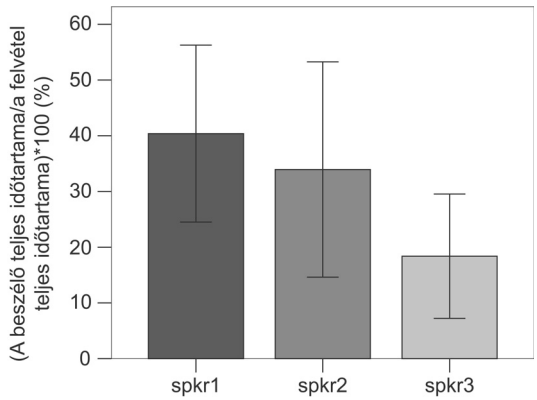
Ebben a részben bemutatjuk az általunk kidolgozott gépi beszélődetektálót, amelyet a BEA adatbázison hoztunk létre és teszteltünk. Magyar társalgásokra még nem történt ilyen jellegű munka, ezért a jelen dolgozat mindenképpen újnak mondható. A BEA adatbázisban lévő társalgások spontánok, amely közelít – noha természetéből fakadva soha nem érheti el azt – a természetes beszédhez. Az eddig létrehozott beszélődetektáló rendszereket rádiós műsorokon hozták létre különböző nyelveken, amelyek inkább félspontánnak minősülnek, hiszen a műsor vezetője előzetesen felkészül a beszélgetésre (ismeri a témát), és a műsor résztvevői is ismerik előzetesen a témát. A BEA adatbázis spontán társalgásainak témáit a résztvevők nem ismerik, ezért a beszédtervezés és -kivitelezés egyszerre zajlik ott helyben. Ezért azt mondhatjuk, hogy a jelenleg használt korpusz jobban közelít a spontán beszédhez, mint az eddig használt korpuszok. Ezért a jelen dolgozat szintén újszerűnek mondható, mivel ilyen jellegű spontán társalgásokon való beszélődetektáló kialakítása eddig még nem történt meg.

Ebben a fejezetben elsőként bemutatjuk a korpusz általános, leíró statisztikai jellemzőit a beszélődetektálásra vonatkozóan. Ezután ismertetjük az általunk javasolt beszélőszegmentálót, majd a beszélőklaszterező eljárásokat. Mindezek után bemutatjuk az általunk elért eredményeket a BEA korpuszon. Továbbá teszteljük, hogy az előző fejezetben létrehozott egyszerrebeszélés-detektáló implementációja milyen hatással van az eredményeinkre.

Minden egyes beszédfelvételt (100 társalgást) manuálisan annotáltunk a beszélőváltások szerint. Minden társalgásban hárman vettek részt, ezért a jelölés beszélőnként: *spkr1:* adatközlő; *spkr2:* felvételvezető; *spkr3:* harmadik résztvevő.

Az általunk random kiválasztott 100 társalgásban 7827 db beszédforduló volt. Egy felvételre átlagosan 70 db beszédforduló jut, amelynek szórása 41 db. A társalgásonkénti legtöbb beszédforduló 240 db volt, míg a legkevesebb 11 db. Megvizsgáltuk, hogy a nemek között van-e különbség a beszédfordulók gyakoriságának tekintetében. Azokban a társalgásokban, amelyekben férfi volt az adatközlő, átlagosan 79 db (szórás 45 db) beszédforduló volt, míg ahol nő, 65 db (szórás 37 db). Azonban ez a különbség nem szignifikáns (egy szempontos ANOVA).

Megvizsgáltuk, hogy a társalgásokon belül az egyes beszélők a teljes időtartamra nézve hány százalékban szólalnak meg. Az adatok szerint az adatközlők átlagosan 40,3%-ban tartják maguknál a szót. A felvételvezető átlagosan 33,9%-ban tartja magánál a szót, míg a harmadik résztvevő csupán átlagosan 18,3%-ban *(5.1. ábra).* Ezek az arányok azt mutatják, hogy a társalgások során a szerepek nem kiegyenlítettek, a harmadik személy sokszor háttérbe szorul (ennek oka többféle lehet, például a feladat jellege, az ismertségi fok).



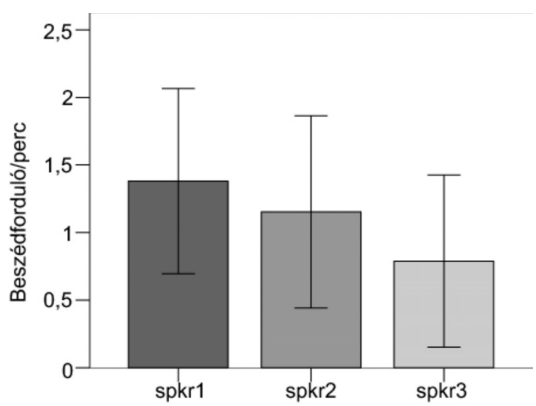
5.1. ábra

A beszélők beszédidejének teljes időtartama a felvétel teljes időtartamának függvényében
(*spkr1:* adatközlő; *spkr2:* felvételvezető; *spkr3:* harmadik résztvevő)

Ez a kiegyenlítetlenség statisztikai elemzéssekkel is alátámasztható, hiszen mind a felvételvezető, mind az adatközlő szignifikánsan többet beszél, mint a harmadik résztvevő (ismétléses ANOVA: spkr1*spkr3: $F(2, 200) = 39{,}833$; $p < 0{,}001$; spkr2*spkr3 $F(2, 200) = 39{,}833$; $p < 0{,}001$).

Elemeztük, hogy nemek tekintetében van-e különbség a beszédidőtartamban az adatközlők esetében. A férfiak átlagosan 37%-ban tartják maguknál a szót a teljes időtartamhoz képest, míg a nők 42%-ban. Azonban ez a különbség szintén nem szignifikáns (egyváltozós ANOVA).

Továbbá kiszámoltuk, hogy az egyes résztvevőkre hány beszédforduló jut percenként. Az adatközlőre átlagosan 1,38 beszédforduló jut percenként, a felvételvezetőre 1,15, míg a harmadik személy esetében 0,78 *(5.2. ábra)*. Ez szintén a társalgás résztvevőinek aszimmetriáját mutatja.



5.2. ábra

Az egy percre eső beszédfordulók száma a résztvevők függvényében
(spkr1: adatközlő; spkr2: felvételvezető; spkr3: harmadik résztvevő)

Megvizsgáltuk, hogy a beszédidőtartamok és a beszédforduló/perc hogyan függnek össze az egyes résztvevők függvényében. Az adatközlőnél nem lehet kimutatni semmilyen tendenciát, vagyis e két jelenség nem függ össze egymással; tehát nem lehet azt mondani, hogy aki sokat beszél, az többször kap vagy veszi át a szót. A kísérletvezető esetében azonban pozitív, közepesen erős függvénykapcsolatot tudtunk kimutatni (Pearson-korreláció: $r = 0{,}424$, $p < 0{,}001$). Ugyanilyen tendenciát találtunk a harmadik résztvevő esetében is (Pearson-korreláció: $r = 0{,}441$, $p < 0{,}001$). Mindez azt mutatja, hogy míg az adatközlőnek nem kell törekedni a szóátvételre, hiszen az adatbázis elsődleges célja, hogy ő beszéljen, addig a felvételvezetőnek és a harmadik személynek igen.

5.2. A beszélődetektáló felépítése

5.2.1. Beszélőszegmentálás

5.2.1.1. Jellemzőkinyerés a beszélőszegmentáláshoz

Az általunk javasolt beszélőszegmentáláshoz a *Beszélőspecifikus jellemzők a gépi beszélőfelismerésben* fejezetben bemutatott MFCC-eljárást használtuk mint jellemzőkinyerő algoritmust. Az MFCC-t kétféleképpen használtuk. Az első megközelítésben a teljes spektrumra kiszámoltuk. A másodikban pedig részsávra; kimutattuk ugyanis, hogy a 2,5 kHz és a 3,5 kHz közötti részsáv az, amelyik a beszélőre vonatkozó akusztikai lenyomatokat tartalmazza. Az MFCC-együtthatókat 32 ms-os ablakhosszra számoltuk, 10 ms-onként.

5.2.1.2. Bayes-féle információs kritérium (BIC: Bayesian Information Criterion)

A jelen munkában a Bayesian Information Criterion algoritmust használtuk a beszélők szegmentálásához. Azért választottuk ezt az eljárást, mert az egyik legtöbbet használt módszer, és mert számítása igen egyszerű és hatékony, illetve nem igényel előzetes komplex modelltanítást. A BIC feladata a beszélőszegmentálásban az, hogy két szomszédos ablakról eldöntse, hogy azonos beszélőtől származik-e vagy sem, vagyis beszélőváltás történik-e vagy sem. A BIC-et valójában tehát arra tudjuk használni, hogy vajon a váltás a két szegmens között megtörténik-e. Két előfeltételezést fogalmazhatunk meg: *i)* a BIC modell jobban illeszkedik az X akusztikai szegmensre, *ii)* mint a $X_i + X_j$ szomszédos ablakkeretekre összevonva.

Legyen M_1 és M_2 két modell:

– Az M_1 modell azt feltételezi, hogy X minden mintája független, és egyetlen többváltozós Gauss-szal leírható.

$$Z = Z_1, Z_2, \ldots, Z_N \sim N(\mu_Z, \Sigma_Z)$$

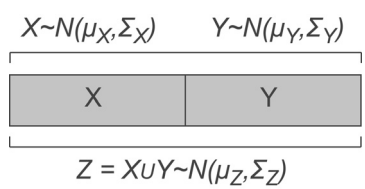
– Az M_2 modell esetében azt felételezzük, hogy X két többváltozós Gauss-szal írható le. Az egyik Gauss a szegmens elejétől b időkeretig, a másik Gauss a b időkerettől a szegmens végéig.

$$M_2 : Z = X + Y$$

$$Z = Z_1, Z_2, \ldots, Z_b \sim N(\mu_X, \Sigma_X)$$

$$Z = Z_{b+1}, Z_{b+2}, \ldots, Z_N \sim N(\mu_Y, \Sigma_Y)$$

Ezen hipotéziseket a következő ábra szemlélteti *(5.3. ábra):*



5.3. ábra

A szegmentálásban hipotetikus modellek egy keretre

A fent leírt BIC-számítások alapján előáll $BIC(M_1)$, $BIC(M_2)$. A BIC-szegmentálás során a büntetőfaktor λ értékét 0-ra vesszük. Ekkor teszteljük a két hipotézisünket:

$$\text{ha } \Delta BIC = BIC(M_2) - BIC(M_1) = 0,$$

akkor ez azt jelenti, hogy a modellre érkező score alapján az adatokra jobban illeszkedik a két többváltozós Gauss-modell (M_2), mint az egy többváltozós Gauss-modell (M_1). Mindez azt jelenti, hogy a szegmens nem homogén, vagyis a szegmensben váltási pont van.

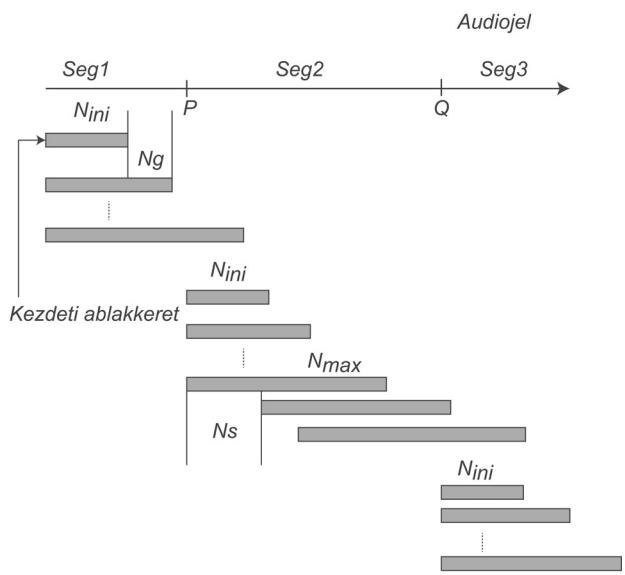
A ΔBIC a beszédben egy akusztikai váltási pont detektálására alkalmas. Nyilvánvalóan a társalgások során jóval több váltási pont létezik, ezért ennek megoldására szekvenciális detektáló algoritmust szokás használni (Cheng et al. 2010). Ennek alapján a ΔBIC-et fel tudjuk írni mint a b váltási pont függvényét. Ha a jellemzővektorok száma X vagy n_x egyenlő b-vel és n_y egyenlő $n - b$-vel, akkor felírhatjuk a következő egyenletet:

$$\Delta BIC_b\{x, y\} = \frac{n}{2}\log|\Sigma_x| - \frac{n-b}{2}\log|\Sigma_y| - \frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right)\log n.$$

A ΔBIC érték alapján akkor helyes a beszédszegmens két részre osztása, vagyis váltási pont feltételezése, ha $\Delta BIC(b) > 0$. A ΔBIC pozitív értéke azt jelenti, hogy az M_2 modell jobban leírja az adatokat, mint M_1 modell, és a váltási pont *(b)* tényleg valós.

5.2.1.2.1. Növekedő ablakhosszmetódus a Δ*BIC* számításához

Ezt az eljárást a beszélőváltási pont detektálására szokás alkalmazni. Az *5.4. ábra* szemlélteti ezt a növekedő ablakhosszú eljárást. Veszünk egy bizonyos hosszúságú ablakot, amelyben N_{ini} jellemzővektor létezik. Ezt az ablakot folyamatosan növeljük N_g mérettel mindaddig, amíg váltási pontot nem találunk a BIC feltétel alapján. Emellett meghatározunk egy nagyobb méretű ablakhosszt, amely N_{max}. Ha a váltási pont előbb detektálódik, mint ahogy elérne az algoritmus az N_{max} időpillanatig, a váltási pont kijelölődik, a folyamat ettől a ponttól kezdődik újra a kezdeti ablakmérettel. Ha N_{max} alatt az algoritmus nem talál váltási pontot, az ablakot eltoljuk *Ns* mintányival, és a folyamat megismétlődik (Cheng et al. 2010). A hátránya ennek a folyamatnak, hogy ahogyan növeljük az ablakhosszt, sokkal nagyobb számítási kapacitásra van szükség.



5.4. ábra
A növekedő ablakolási eljárás sematikus ábrája (Cheng et al. 2008)

5.2.1.2.2. A BIC paraméterei

A jelen kutatásban a BIC értékét a következő beállítások mellett végeztük el.

A BIC kiértékelési ideje: 10 másodperc.

A BIC durva kiértékelési ideje: 1 másodperc.

A BIC végleges kiértékelési ideje: 0,1 másodperc.

A BIC kiértékelésének zárópuffermérete: 1 másodperc.

A BIC értékét a kontextus figyelembevétele miatt 2 egymást követő ablakhosszra számoljuk, amely jelen esetben 2 másodperc.

5.2.1.3. Téves riasztások csökkentése (False Alarm Compensation)

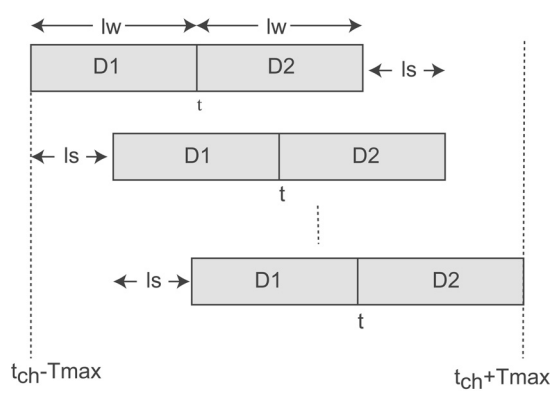
A legtöbb beszélőszegmentáló kétutas folyamatot tartalmaz. Az első „durva" szegmentálás után egy újraellenőrző (utófeldolgozás) szegmentálási folyamatot szokás végezni, az elsőként hipotetizált váltási pontok érvényességének vizsgálatához. A jelen munkában a BIC-szegmentálás után szimmetrikus Kullback–Leibler-távolság-alapú szegmentálási folyamattal vizsgáltuk felül a váltási pontokat (Siegler et al. 1997; Hung et al. 2000).

A KL2 szimmetrikus távolság esetén egy küszöbértékkel számolunk, amelynek megválasztása kísérleti úton történik. A küszöb megválasztása azonban történhet automatikus küszöb számításával. Az automatikus küszöb számítása úgy történik, hogy kiszámoljuk a KL2-különbséget minden egyes időkeretre *(t),* amelyet *ls* mérettel tolunk tovább, és amelyben a távolság $\pm T_{max}$ a hipotetizált váltási pont körül t_{ch}:

$$küszöb_{ch} = \alpha \cdot \frac{1}{2T_{\max} + 1} \sum_{i} KL2_{ch+i},$$

ahol $-T_{\max} / l_s < i < T_{\max} / l_s$, és az α előre definiált faktor, amelyeket kísérleti úton kell meghatároznunk (Ida 2011).

Mivel a beszélőszegmentálásból származó hibák többsége a téves riasztásból fakad, ezért KL2-folyamatot a téves riasztások csökkentésére alkalmazzuk (Ida 2011). A KL2-távolságot az utófeldolgozáskor a hipotetikus t_{ch} pont körül $\pm T_{max}$ időkeretben számoljuk ki *(5.5. ábra).*



5.5. ábra

A KL2 utófeldolgozásának folyamata

5.2.1.3.1. A KL2-alapú utófeldolgozás beállításai

A KL2-utófeldolgozást végző algoritmusnak két szabad paramétere van. Az egyik a keret hossza, l_s, amelyet 10 keretnyire, vagyis 0,1 másodpercre állítottunk. A másik a T_{max}, amelyet a jelen munkában 200 keretnyire, vagyis 2 másodpercre állítottunk. A harmadikat, az α paramétert pedig 0,5-re állítottuk be.

5.2.2. Beszélőklaszterezés

5.2.2.1. Jellemzőkinyerés a beszélőklaszterezéshez

Az általunk javasolt beszélőszegmentáláshoz a *Beszélőspecifikus jellemzők a gépi beszélőfelismerésben* fejezetben bemutatott MFCC-eljárást használtuk mint jellemzőkinyerő algoritmust. 12 dimenziós MFCC-vektort nyertünk ki 10 ms-onként 32 ms-os Hamming-ablakolófüggvénnyel. A határsávértékeket a Mel-szűrő-skálához először a teljes spektrumra, majd 2,5 kHz és a 3,5 kHz közöttire állítottuk, amelyek a beszélőre vonatkozó akusztikai lenyomatokat tartalmazzák. A kepsztrális együtthatók mellett hozzávettük a jel energiájának logaritmusát. Kiszámoltuk a jellemzők dinamikus információit is, azaz az első két deriváltat, így egy 39 dimenziós jellemzővektort kaptunk. Ezek után minden egyes jellemzővektort normalizáltunk az átlagához és a varianciájához.

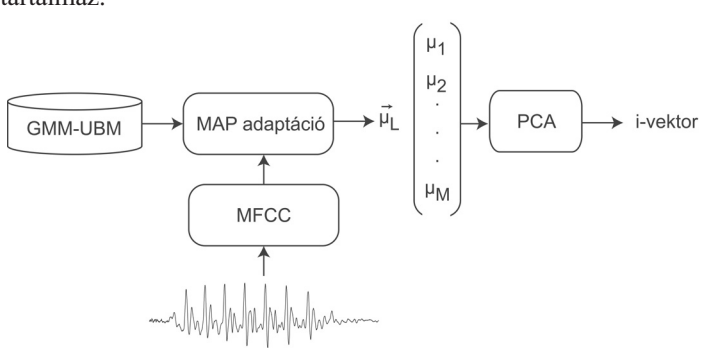
A beszélőklaszterezés során a beszélőszegmentálóból érkező szegmensek a bemenetek, vagyis a két beszélőváltás között lévő beszédjelek. Ezen beszédjelek modellezésére a kinyert akusztikai jellemzőkből GMM-szupervektorokat képeztünk.

5.2.2.2. GMM-szupervektor

Tegyük fel, hogy a kevert Gauss-modell általános háttérmodell (GMM-UBM):

$$g(x) = \sum_{i=1}^{N} \lambda_i N(x, m_i, \Sigma_i),$$

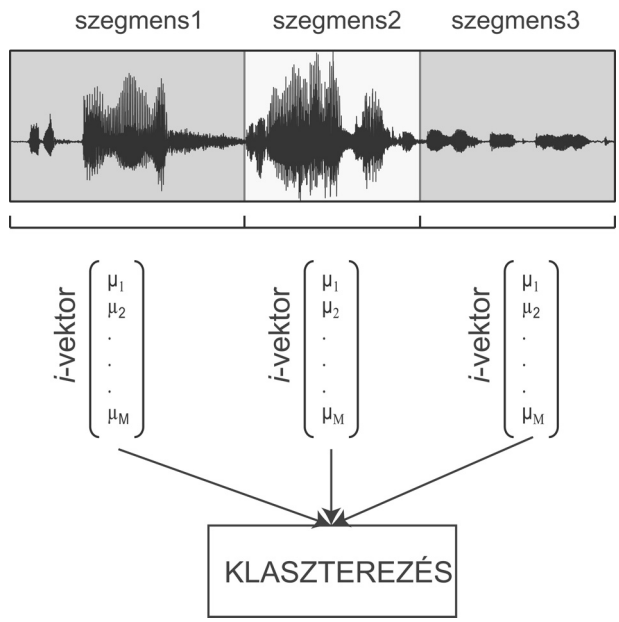
ahol λ_i a súlyok, *N()* Gauss-komponens, m_i és $\sum_i$ a középértéke és a kovarianciája a Gauss-eloszlásnak. A jelen kutatás során diagonális kovarianciájú GMM-et használtunk. Adott egy beszédszegmens a beszélőszegmentálásból, a GMM-UBM tanítása a középértékeknek MAP-adaptálásával hajtódik végre (Reynolds et al; 2000, Reynolds 2009). Ezen adaptált középértékeket fűzzük össze GMM-szupervektorrá *(5.6. ábra)*. A jelen kutatásban a GMM-UBM 256 kevert komponenst tartalmaz.



5.6. ábra

A jellemzőkinyerés sematikus blokkdiagramja

Mivel így egy magas dimenziószámú jellemzővektort kapunk, ezért a jelen dolgozatban a dimenziószámot lecsökkentettük PCA-val (főkomponens-analízis, Principal Component Analysis). A dimenziócsökkentett jellemzővektor jelen esetben az *i*-vektor. Az *i*-vektor a bemenete a felügyelet nélküli tanuláson alapuló beszélőklaszterezésnek *(5.7. ábra)*.



5.7. ábra

Az i-vektor mint bemenet a beszélőklaszterezéshez

5.2.2.3. BIC-alapú klaszterezés

A BIC-alapú klaszterezési eljárás az agglomeratív hierarchikus klaszterezési (AHC) eljárások közé tartozik. Az agglomeratív (vagy összevonó) klaszterezési eljárás az egyik legtöbbet alkalmazott eljárás a beszélőklaszterezésben. Az AHC alapvető működése, hogy progresszíven vonja össze az egyes klasztereket valamilyen egyezőségi mutató alapján. Az AHC alapvető két kérdése, hogy *i)* milyen metrikát használjunk az egyes klaszterek közelségének/azonosságának mérésére, *ii)* illetve hogy milyen mérőszám alapján állítsuk le az összevonást, vagyis hogy hány klasztert képezzünk (a beszélődetektálásban a beszélők számát jelenti). Számos eljárás létezik ezen kérdések megválaszolására. A jelen tanulmányban a BIC-algoritmust használjuk mindkét probléma megoldására.

Ahhoz, hogy kiválasszuk a közelebbi klaszterpárokat, majd összevonjuk őket a rekurziós lépés során, ki kell számolnunk az összes lehetséges klaszterpár közötti BIC-távolságot. Ezek után a legkisebb BIC-értékű klaszterpár összevonásra kerül.

Legyen egy klaszterpár C_x és C_y a rekurziós lépésben, amely n-dimenziójú adat (akusztikai jellemzővektor), $x = \{x_1, x_2, ..., x_M\}$ és $y=\{y_1, y_2, ..., y_M\}$. A ΔBIC-értéket a következőképpen számolhatjuk:

$$\Delta BIC(C_x, C_y) = BIC(C_x, C_y | H_1 - BIC(C_x, C_y) | H_2) =$$

$$= \ln p(x \cup y | H_1) - \frac{\lambda}{2} \cdot N_{H_1} \cdot \ln N_{total} - \left\{ \ln p(x \cup y | H_2) - \frac{\lambda}{2} \cdot N_{H_2} \cdot \ln N_{total} \right\} =$$

$$= \ln \frac{\ln p(x \cup y | H_1)}{\ln p(x \cup y | H_2)} - (N_{H_1} - N_{H_2}) \ln N_{total},$$

ahol

– H_0 (nincs összevonás hipotézis): C_x és C_y nem kerül összevonásra,
– H_1 (összevonás hipotézis): C_x és C_y összevonásra kerülnek, így egy új klasztert képeznek együtt, C_z, ahol $z = x \cup y$.

A fenti egyenletben a λ (teoretikusan 1) hangoló paraméter, N_{H_0} és N_{H_1} a két hipotézis paramétereinek száma a statisztikai eloszlások reprezentálásában, és N_{total} a teljes száma az adatoknak.

Az agglomeratív hierarchikus klaszterező algoritmus akkor áll le, ha a BIC értéke negatívvá válik.

5.3. A beszéddetektálás felépítése

Mivel a jelen kutatásnak nem alapvető célja, hogy új beszéddetektálót fejlesszen, ezért a Giannakopoulos (2009) által kidolgozott és MATLAB-ba implementált beszéddetektáló algoritmusát használtuk, illetve módosítottuk. Ez az algoritmus rövid idejű energiafüggvény (short-term energy), spektrális centroid (spectral centroid) akusztikai jellemzőket és adaptív küszöbölést alkalmaz a beszéd és nembeszéd szegmensek automatikus meghatározására. Az általunk ajánlott módszer annyiban tér el ettől (lásd részletesebben lent), hogy a küszöb meghatározását felügyelet nélküli tanulási metódussal végezzük el.

A jelen kutatás célja tehát az, hogy automatikusan meghatározzuk az egyes jelszegmensekre, hogy beszéd vagy nembeszéd szegmens-e, illetve hogy teszteljük, hogy az általunk javasolt felügyelet nélküli tanulási módszer javít-e az eredményeken.

5.3.1. Jellemzőkinyerés

A jellemzőkinyerés előtt a folytonos jelet rövid szegmensekre bontottuk, vagyis ablakoltuk (frame-ekre: keretekre). Az ablakok hossza 50 ms-os volt. Az ablakok között nem volt átfedés. Az ablakolást Hamming-típusú függvénnyel végeztük. Ezután minden egyes keretre kiszámoltuk a két akusztikai jellemzőt: rövid idejű energiafüggvény és spektrális centroid jellemzőket.

i) A rövid idejű energiafüggvény:

Legyen $x_i(n)$, $n = 1, \ldots, N$ az i-edik keret egy audiojelben, amelynek hossza N. Minden egyes i-edik keretre kiszámoljuk az energiát a következő egyenlettel:

$$E(i) = \frac{1}{N} \sum_{n=1}^{N} |x_i(n)|^2.$$

ii) Spektrális centroid:

A spektrális centroid C_i, az i-edik keretre számolt spektrum súlyközéppontját (center of gravity) jelenti, amelyet a következőképpen számolhatunk:

$$C_i = \frac{\sum_{n=1}^{N} (k+1) X_i(k)}{\sum_{k=1}^{N} X_i(k)},$$

ahol $k = 1, \ldots, N$ az i-edik keret diszkrét koszinusztranszformáció koefficiense, és ahol az N a keret hossza. Ez a jellemző frekvenciákat mutatja meg a spektrumban, amelynek magas értékkel való realizációja a beszédjelre utal (Saunders 1996; Theodoridis-Koutroumbas 2008).

Mindkét akusztikai paraméter kiszámolása után 5 pontos mediánszűrést alkalmaztunk a kiugró értékek simítása végett.

A jelen beszéddetektáló megvalósításához azért alkalmas ez a két jellemző, mert *i)* (ha az akusztikai jel nem terhelt nagy zajjal) az energia értéke magasabb a beszéd esetén, mint a szünet esetén, illetve *ii)* hasonlóképpen, a spektrális centroid értéke szintén magasabb értéken, vagyis frekvencián realizálódik beszéd esetén, mint szünet esetén.

5.3.2. A beszéddetektáló döntési metódusa

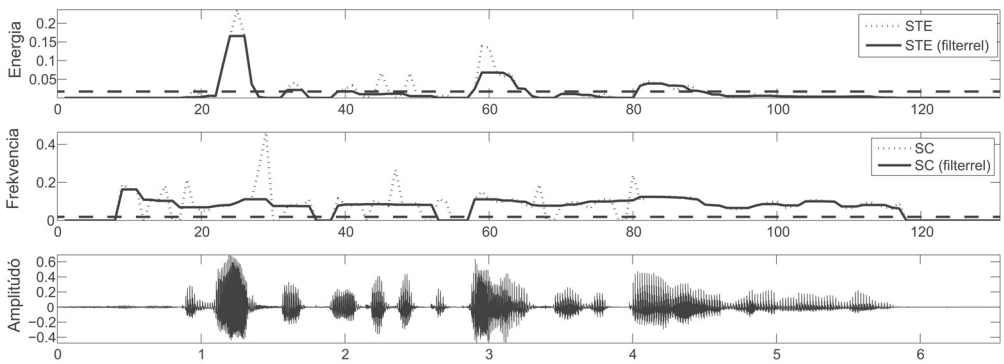
Az akusztikai jellemzők kinyerése után egy egyszerű küszöbalapú döntési eljárást alkalmaztunk. Első lépésként a két küszöb (mindkét jellemzőre egy-egy) kiszámolására kerül sor. A küszöb kiszámolásáig a következő folyamatok mennek végbe *(5.8. ábra):*

1. Az egyes jellemzők eloszlásának modellezése.
2. A hisztogram simítása.
3. A hisztogram lokális maximumainak detektálása.

4. A küszöb értékének kiszámítása: legyen M_1 és M_2 az első és második lokális maximum pozíciója. Ekkor a küszöbértéket a következő egyenlettel számolhatjuk ki:

$$T = \frac{W \cdot M_1 + M_2}{W + 1},$$

ahol W egy szabad paraméter. Ha a W értéke magas, akkor az M_1-hez lesz közelebb a küszöbérték.



5.8. ábra

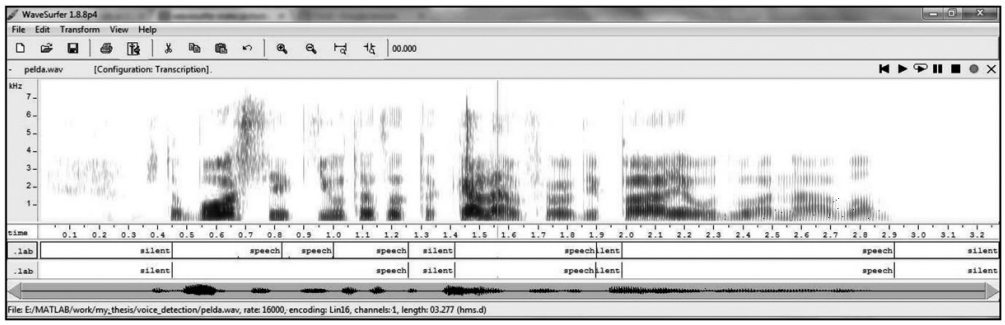
Az egyes jellemzőkre alkalmazott küszöbérték
(STE: rövid idejű energia; SC: spektrális centroid; a felső két ábrán lévő vízszintes szaggatott vonal a küszöbértéket jelenti)

Amikor a küszöbértékek rendelkezésre állnak mindkét akusztikai jellemzőre: T_1 és T_2, akkor az egyes ablakolt szegmensekre végrehajtódik a döntési feladat. Ebben az esetben a döntés a következő:

1. a rövid idejű energiafüggvényre: $M_1 > T_1$, akkor beszédrész,
2. a spektrális centroidra: $M_2 > T_2$, akkor beszédrész,
3. végső döntés: ha 1. és 2., akkor beszédrész.

5.3.3. A beszéddetektáló utófeldolgozása

A beszéddetektáló utófeldolgozásakor a detektált beszédszegmenseket meghosszabbítjuk y rövid távú ablakkal (ez a keret hossza szorozva y ms-os hosszt jelent) mindkét oldalon. Végül az egymást követő szegmenseket összevonjuk *(5.9. ábra)*.



5.9. ábra

A döntési metódus (felső annotációs sor) és az utófeldolgozás (alsó annotációs sor) utáni automatikus annotáció

5.3.4. Az általunk javasolt eljárás a küszöb meghatározására

A jelen kutatásban a hisztogram számolása és a csúcsok megtalálása helyett felügyelet nélküli módszert alkalmaztunk. A korábbi munkákban szintén használtak felügyelet nélküli tanuló algoritmusokat a beszéddetektáló megvalósításában. Ying és munkatársai (2011) szekvenciális kevert Gauss-modell-alapú beszéddetektálót javasoltak, amelynek a bemenete az energia eloszlása a Mel-szűrő frekvenciasávjaiban. Kezdeti lépésként az algoritmus a beérkezett keretekre felügyelet nélküli módon két Gauss-t illeszt, ahol az alacsonyabb középértékkel rendelkező klaszter nembeszédnek felel meg, míg a magasabb középértékkel rendelkező beszéd résznek. Ezt a metódust a küszöbérték meghatározásában is alkalmazzák.

Jelen munkában a középpontok (beszéd és nembeszéd) megtalálásához klaszteranalízist használtunk. A klaszteranalízis lényege, hogy egy adattömböt több homogén részcsoportra bontsunk úgy, hogy az azonos csoportba tartozó elemek között a hasonlóság mértéke nagyobb legyen, mint az azon kívüli elemek között. A hasonlóság mérésére többféle lehetőség van. Az egyik leggyakrabban használt hasonlóságot mérő eljárás az euklideszi távolság vagy annak négyzetes távolsága, illetve használatos még a Manhattan-távolság, Mahalabonis-távolság stb.

A jelen kutatásban a *k*-közép (*k*-means) algoritmust alkalmaztuk, amely egy változata a klaszteranalízisnek. A *k*-közép eljárás lépései a következők:

a) Véletlenszerűen vagy egy adott stratégia alapján létrehoz *k* számú klasztert, és meghatározza ezek középpontjait.
b) Minden egyes pontot abba a klaszterbe sorol, amelynek középpontjához a legközelebb helyezkedik el.
c) Kiszámolja a klaszterek középpontjait.
d) Addig ismételi az előző két lépést, amíg a reprezentánsok rendszere változik.

Ezek alapján meghatározunk egy hibafüggvényt:

$$Err = \sum_i \sum_j r_{ij} || \mu_i - x_t ||^2 \rightarrow \min,$$

ahol μ_i az *i*-edik klaszter középpontja. A $r_{ij} = 1$, ha az *i*-edik klaszterbe soroljuk a *j*-edik mintát, egyébként 0. A cél az, hogy minimalizálja a fent leírt hiba értékét, azaz az *i*-edik klaszterbe tartozó minták távolságnégyzet-összege az *i*-edik középponttól minimális legyen. Az optimum ott van, ahol az *Err* μ szerinti deriváltja 0, azaz ha a klaszterközéppontok egybeesnek a klaszterekhez tartozó pontokkal.

Az algoritmus előnye, hogy egyszerűen megvalósítható, és nem érzékeny az alappontok sorrendjére.

Mivel alapvetően két csoportot kívánunk létrehozni, ezért a klaszterközéppontok számát kettőben határozzuk meg: beszéd és szünet. Az így kialakított két csoport klaszterközéppontja M_1 és M_2 lesz.

5.3.5. A beszéddetektáló kiértékelése

A beszéddetektáló kiértékeléséhez a DER-módszert használtuk. A módszer ebben az esetben nem a beszélők szegmentálását és klaszterezését méri, hanem a beszéd és nembeszéd szegmens és nembeszéd részek szegmentálási pontosságát és helyes azonosítását adja meg. A beszéddetektáló működésének tesztelésekor tehát nem a Diarization Error Rate-et kapjuk eredményül, hanem a beszéddetektáló Error Rate-et, vagyis a beszéddetektálásból származó hibát.

Az alaprendszer és az általunk javasolt rendszer összehasonlításához nemparametrikus összetartozó mintás (Wilcoxon-próba) tesztet használtunk, Monte-Carlo-szimulációval megerősítve.

5.4. Az egyszerre beszélések automatikus osztályozása spontán magyar társalgásokban

Az osztályozás második fontos lépése a mintaillesztés, amelyben két fontos részfeladatot kell megoldani: *i)* osztályozás, vagyis melyik beszédrészletmodell a legvalószínűbb az adott időpillanatban; *ii)* időillesztés: melyik időszegmenst rendeljük az egyik vagy a másik modellhez. Ennek megvalósításához a beérkező mintát, vagyis a vektorsorozatot (statisztikai úton becsült) valószínűségimodell-struktúrához illesztjük. Az akusztikus modell létrehozásához legtöbbször a kevert Gauss-modellt (GMM: Gaussian Mixture Model) használják. Bár az akusztikus modell létrehozásában igen széles körben és kiválóan alkalmazható, mégis számos

hátránya létezik. Az egyik hátránya, hogy az adatoknak előzetes feltételeknek kell megfelelniük a becslést megelőzően – ilyen követelmény a normál eloszlás. A GMM alternatívájaként léteznek más megoldások, mint például a mesterséges neuronhálók (például a MLP: Multilayer Perceptron, Bishop 1996, 2006). Az elmúlt években a mesterséges neurális hálózat (ANN) egy új fajtája jelent meg: ún. mély neuronhálók, amely a vizsgálatok szerint igen jól alkalmazható többek között a beszédhang-felismerésben (Dahl et al. 2010; Grósz–Tóth 2013). A mély neuronhálók elsősorban abban különböznek az előző neuronhálóktól, hogy általában nem egy, hanem 3–9 rejtett réteget használnak. A több rejtett réteg tanításához újfajta tanulóalgoritmust is fejlesztettek. A jelen kutatásban a mély neuronhálókat az akusztikai jellemzők előfeldolgozásához használtuk. A tényleges osztályozást LS-SVM-mel végeztük el, amely az SVM egyik változata. Korábbi tanulmányok kimutatták, hogy az ANN és az SVM algoritmusok kombinációja jól alkalmazható automatikus osztályozáshoz (Bellili et al. 2001).

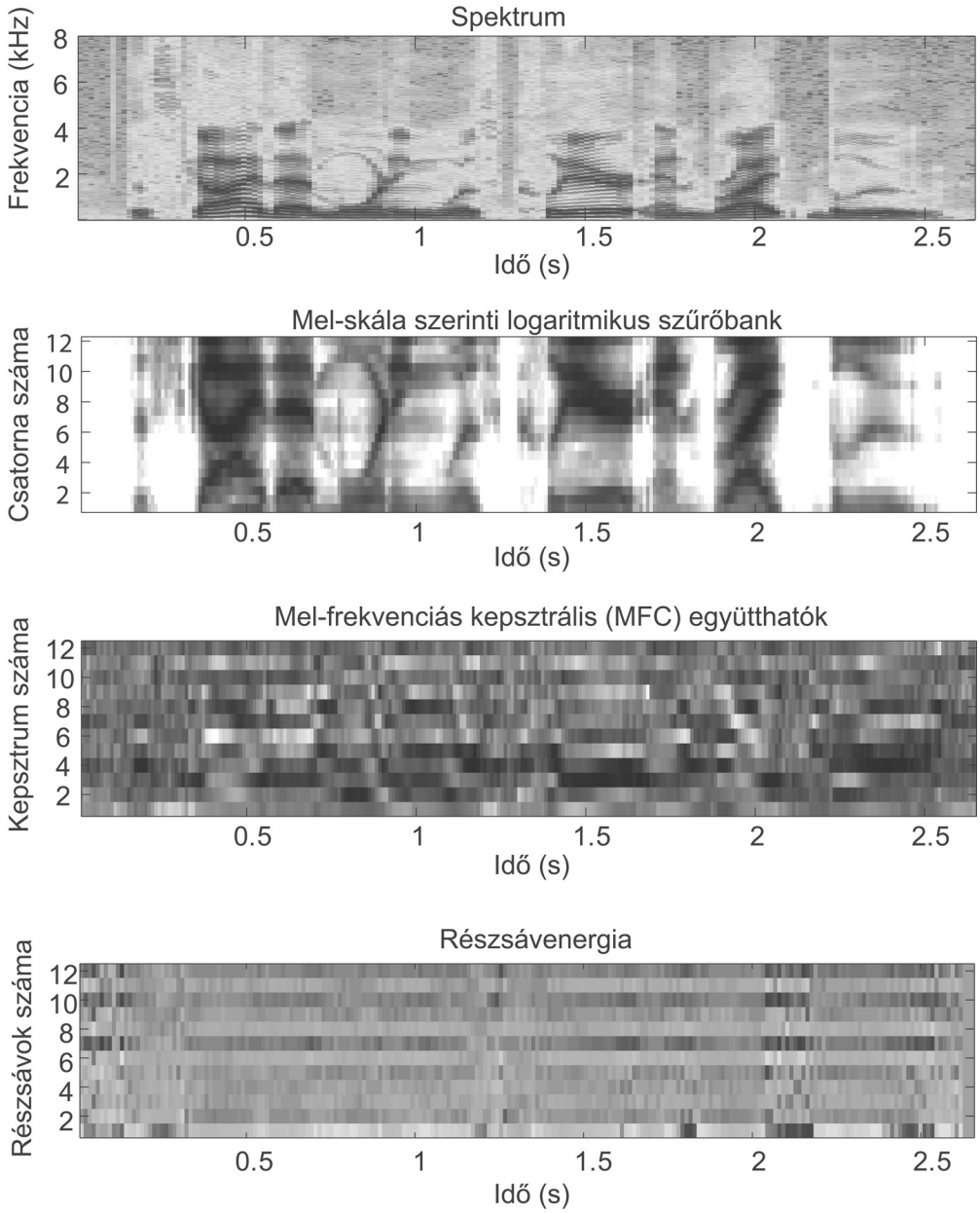
5.4.1. Jellemzőkinyerés

Az egyszerre beszélések jó megfeleltethetőségéhez az akusztikai beszédjelből különböző jellemzőket nyertünk ki. Az első kísérletben az akusztikai jelet FFT-analízissel felbontottuk spektrális jellemzőkké. A második kísérletben MFCC-együtthatókat használtuk. A harmadik kísérlet során a Mel-skála szerinti logaritmikus szűrőbankjellemzőt vizsgáltuk. A negyedik kísérletben a spektrumot részsávokra bontottuk, és az egyes részsávokban számoltuk ki a jel energiáját *(5.10. ábra)*.

i) A **spektrum (SP)** kiszámolásához 256-pontos FFT-analízist használtunk Hamming-ablakkal, (8000 Hz-es mintavételezés esetén) az ablak hossza 32 ms volt, amelyet 10 ms-onként léptettünk. A jellemzővektor hossza ebben az esetben 257 minden egyes 10 ms-os időkeretre. Mivel a 257 dimenzió igen nagy, ezért PCA-val (Principal Component Analysis, főkomponens-analízis) lecsökkentettük 80-ra.

ii) A **Mel-frekvenciás kepsztrális (MFCC) együtthatók** kinyeréséhez a PLP-RASTA csomagban található MATLAB szoftverkörnyezetre írt MFCCC-algoritmust használtuk (vö. Daniel 2005). A jellemzők száma egy-egy időkeretben 39: a szokásos 12 MFCC koefficiens + az energia logaritmusa + ezek első két deriváltja ($13 \times 2 = 26$). Ezt a 39 paramétert 10 ms-onként 25 ms-os, 50%-ban átlapolódó időkeretekben kimértük. A jellemzővektor hossza így 39, minden egyes 10 ms-os időkeretre.

iii) A **Mel-skála szerinti logaritmikus szűrőbank** *(MSL)* számítása ugyanúgy történik, ahogyan az MFCC kiszámítása. A különbség abban áll, hogy a Mel-frekvenciás szűrés után vesszük annak logaritmusát, de nem végezzük el a kepsztrális transzformációt. Ennek kiszámítása szintén: 12 koefficiens + az energia logaritmusa + ezek első két deriváltja ($13 \times 2 = 26$). Ezt a 39 paramétert 10 ms-onként 25 ms-os, 50%-ban átlapolódó időkeretekben kimértük. A jellemzővektor hossza így 39, minden egyes 10 ms-os időkeretre.



5.10. ábra

Az akusztikai jelből számolt különféle akusztikai reprezentációk

iv) A **részsávenergiát (RSE)** úgy számoltuk ki, hogy a spektrumot 20 részsávra bontottuk, majd mind a 20 részsávban kiszámoltuk a jel energiáját. A folyamat végén a 20 elemű vektort DCT-vel (Discrete Cosine Transformation, diszkrét koszinusztranszformáció) dimenziócsökkentettük 12-re.

Mindegyik jellemző esetén a különféle zajok – elsősorban a konvolúciós zajok (például a csatornatorzítás) – hatását mérséklendő további transzformációs lépést alkalmaztunk: kepsztrális átlagkivonást (CMS: Cepstral Mean Substraction).

Mivel a következő lépésben neurális hálózatot alkalmazunk, ezért az adatokat 0 és 1 közé normalizáltuk.

5.4.2. Lényegkiemelés

5.4.2.1. Korlátozott Boltzmann-gép

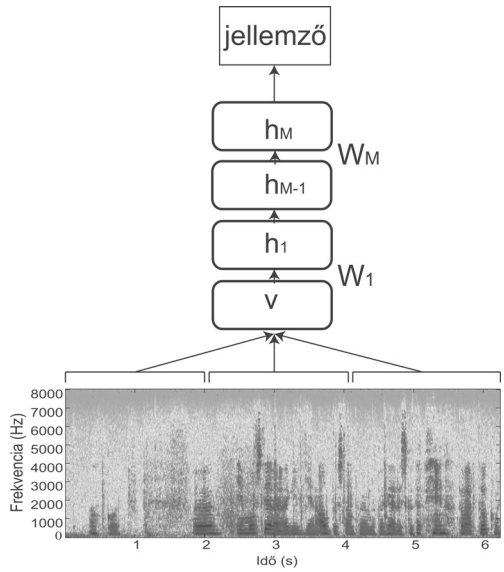
Az elmúlt években számos kísérlet bizonyította, hogy a gépi látásos módszerek jó eredménnyel alkalmazhatók beszéddel kapcsolatos problémák megoldására (Dahl et al. 2010). A gépi látásos módszerek egyik legtöbbet használt algoritmusa a konvolúciós hálózatok. A konvolúciós hálózatok hierarchiát alkotva több szintből épülnek fel, ahol az alsóbb szinteken csak egy kis részét látják a képnek, erről a részletről lokális jellemzőket nyernek ki, amelyeket a felsőbb szinteknek továbbítanak, és egyre feljebb jutva az egyes szinteken egyre általánosabb jellemzőket állapítanak meg. Ezt a módszert napjainkban egyre szélesebb körben használják beszédre. Ekkor a cél az akusztikai jelből valamilyen képi jellegű információ kinyerése. Az akusztikai jelfeldolgozásban ilyen eljárások a spektrogramok és a leggyakrabban használt Mel-skálázott spektrogramok, amelyek az emberi hallást modellezik.

Az RBM (Restricted Boltzmann Machine, korlátos Boltzmann-gép) két különböző réteget tartalmaz: látható és rejtett réteg. A korlátos jelző arra utal, hogy a neuronok között csak akkor van összeköttetés, ha az egyik a látható, a másik pedig a rejtett réteghez tartozik. Az azonos rétegbe tartozó neuronok között nincs összeköttetés.

A súlyok az egyes kapcsolatok között, illetve a neuronokhoz tartozó eltolásértékek (bias-ok) egy véletlen eloszlást definiálnak a látható réteg neuronjainak állapotait tartalmazó vektorok felett, amelyet egy energiafüggvény segítségével írhatunk le. Az alapenergia-függvény bináris adatok eloszlásának leírására alkalmas. Mivel a jelen kutatásban az RBM bemeneti vektora valós értékű, ezért az RBM-eknek a Gauss–Bernoulli RBM változatát használjuk.

A korlátos Boltzmann-gép tanító algoritmusa a CD-algoritmus (kontrasztív divergencia). A CD-algoritmus felügyelet nélküli tanulást végez, amely a „maximum likelihood" tanítás közelítését adja. Ezt a folyamatot az RBM előtanításának nevezzük (Grósz–Tóth 2013).

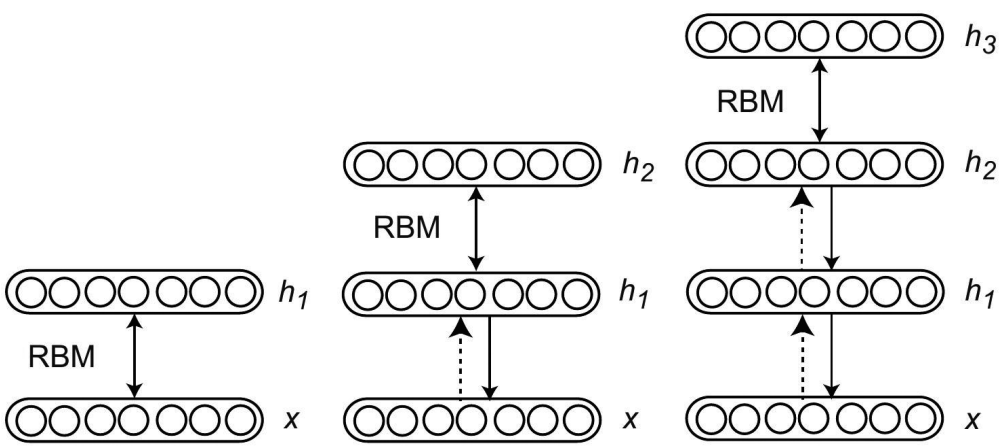
A jellemzők kinyerése után korlátozott Boltzmann-géppel emeltük ki a lényeget az akusztikai jellemzőkből. A korlátozott Boltzmann-gépet szokás jellemzőkinyerésre is alkalmazni – főként a képfeldolgozásban –, amely ebben az esetben felügyelet nélküli tanulási eljárással működik *(5.11. ábra)*.



5.11. ábra

Jellemzőkinyerés korlátozott Boltzmann-géppel

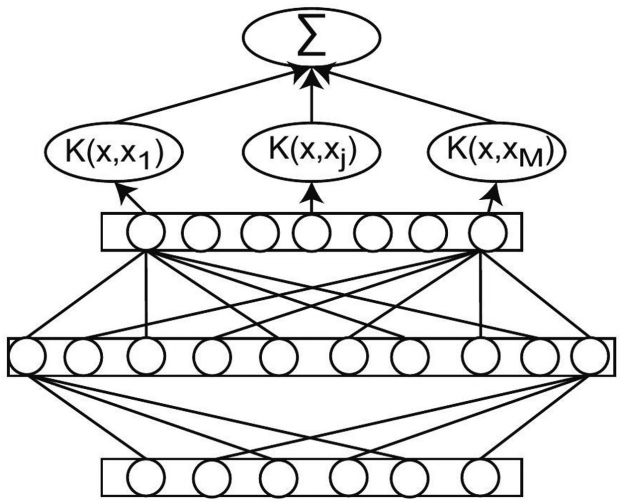
Az RBM előnye, hogy könnyedén mély neuronná lehet alakítani, ha az egyes RBM-eket összekötjük, előállítva ezzel egy hierarchikus tanulási láncot, így segítve a magasabb szintű struktúrák kinyerését az adatokból *(5.12. ábra)*.



5.12. ábra

A korlátozott Boltzmann-gép és a belőle felépített mély neuronháló (Deep Neural Network)

Az RBF tanítása után a rejtett rétegek aktivációs értékeit használtuk fel az átfedő beszédrészek és nem átfedő beszédrészek automatikus osztályozásához, amelyet szupport vektor géppel valósítottunk meg *(5.13. ábra)*.



5.13. ábra
Szupport vektor gép mély neuronhálóval előtanítva

5.4.2.2. Az RBM előtanítási paraméterei

Az RBM előtanításához az akusztikai paramétereket 9 keret hosszúságú csúszóablakkal nyerjük ki. Mindegyik összefüggő ablakot felhasználjuk az RBM tanításához. Az RBM látható egységeinek száma: a jellemzővektor dimenziószámának a keret hosszával képzett szorzata. Minden egyes audioszegmensre az érvényes konvolúcióval kifejezve *m–n+1* összefüggő ablak adódik, ahol *m* a keretek száma, *n* a csúszóablak hossza. A mélyrétegű neurális hálózatok létrehozásához 1–3 RBM-et kapcsoltunk össze úgy, hogy a megelőző rejtett réteg aktivációja a következő látható réteg bemenete.

Az első RBM-ben (H_1) a unitok száma 300.

A második RBM-ben (H_2) a unitok száma 600.

A harmadik RBM-ben (H_3) a unitok számát 300–900-ig növeltük 100 unitonként.

Minden egyes rétegben energiafüggvényként a Gauss–Bernoulli-algoritmust használtuk. A batch mérete 100 volt, amely a kötegelt tanítás mérete. Az első rétegben 50 iteráció, a többi rétegben 25 volt.

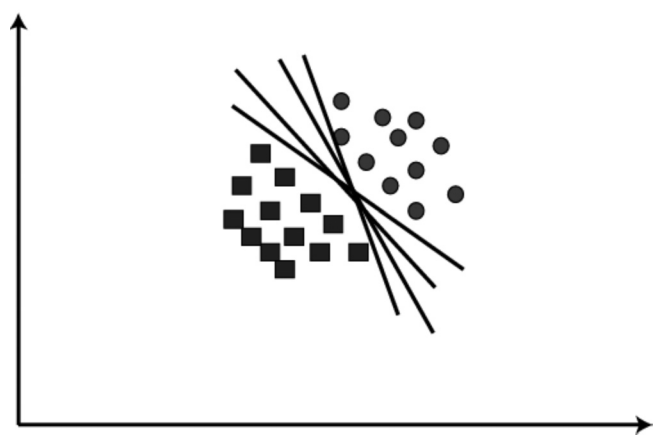
Az RBM megvalósításához Kyung Hyun Cho MATLAB-ban írt GitHub toolbox-át használtuk (Cho 2013).

5.4.3. Osztályozás

Az átfedő és nem átfedő beszédrészeket szupport vektor géppel (SVM) kernelfüggvényként radiális bázisfüggvényt (RBF) alkalmazva osztályoztuk.

5.4.3.1. Szupport vektor gép (SVM: Support Vector Machine)

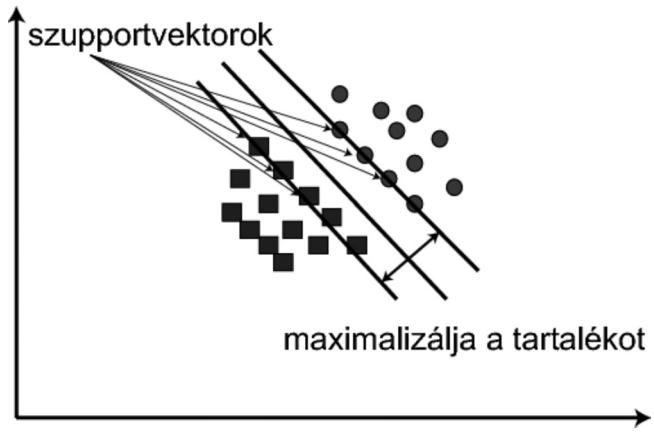
Az SVM olyan matematikai konstrukció, amelyet döntési problémák megoldásához szoktak alkalmazni. Alapverziója a lineáris osztályozók családjába tartozik, de bináris osztályozási problémák megoldására is alkalmas. A többi lineáris osztályozóhoz képest az a fő ismérve, hogy nemcsak egyszerűen olyan hipersíkot (más néven vágási síkot) keres, amely elválasztja a pozitív és a negatív tanító mintákat, hanem ezek közül a legjobbat kutatja, vagyis intuitíve azt, amelyik a két osztály mintái között éppen „középen" fekszik *(5.14. ábra).*



5.14. ábra

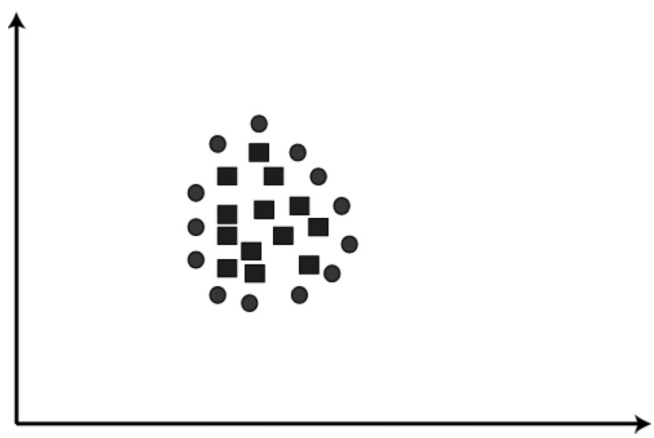
Lehetséges hipersíkok lineárisan szeparálható adatok esetében

Az SVM tehát olyan döntési hipersíkot határoz meg, amely maximalizálja a tartalékot, azaz a hipersík és a hozzá legközelebbi pozitív és negatív tanítóadatok közti eltérést. Ezeket a tanítóadatokat szupportvektoroknak nevezzük. A hipersík meghatározásában a tanítóadatok közül csak a szupportvektorok játszanak szerepet. Ennek az eljárásnak az előnye egyrészt az, hogy a hipersíkhoz közel álló események osztályba sorolása a legbizonytalanabb; így minél kevesebb pont esik erre a területre, annál kevesebb bizonytalan döntést hoz az osztályozó. Másrészt a maximális tartalék által meghatározott szélességű szeparálósáv elhelyezésére sokkal kevesebb lehetőség van, mint egy tetszőleges szeparáló hipersík esetén. Így kevésbé függ a konkrét adatoktól, ezért az osztályozási modell nagyobb általánosító képességgel rendelkezik. Az SVM-et alapvetően lineárisan szeparálható esetekre találták ki *(5.15. ábra).*



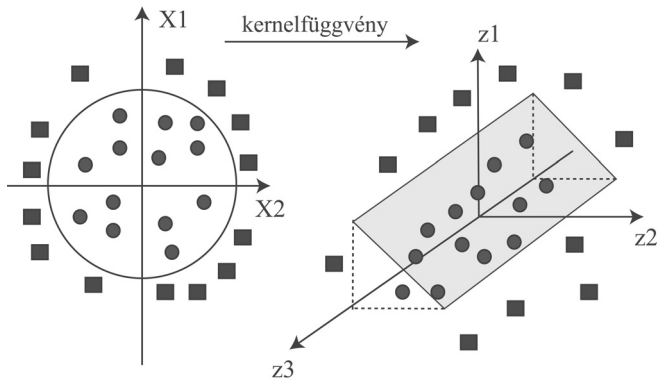
5.15. ábra
Két osztály, amelyek egy hipersíkkal elkülöníthetők: lineárisan szeparálható eset

A valóságban azonban a legtöbb probléma nemlinearitása olyan nagyságrendű, hogy az osztályozó nem lesz hatékony *(5.16. ábra)*.



5.16. ábra
Két osztály, amelyek egy hipersíkkal nem különíthetők el: nemlineárisan szeparálható eset

Ennek a problémának a megoldására az adatokat nagyobb dimenziójú térbe transzformáljuk, ahol az adathalmaz már lineárisan szeparálható. Az erre képes matematikai függvényeket kernel- vagy magfüggvényeknek nevezzük. A magfüggvények segítségével a lineárisan nem szeparálható feladatok lineárisan szeparálhatóvá tehetők azzal, hogy az adatokat jobban reprezentálható problématérbe transzformáljuk *(5.17. ábra)*.



5.17. ábra

A lineárisan nem szeparálható adatok kernelfüggvénnyel való transzformációja egy olyan térbe, ahol lineárisan szeparálhatóvá válnak

A gyakorlatban a következő magfüggvényeket szokták alkalmazni: polinomiális, radiális bázisfüggvény, kétrétegű perceptron.

A jelen kutatásban az SVM egy változatát használtuk, ez az LS-SVM (Least Square Support Vector Machine, Suykens et al. 2002). Ez a típus abban tér el az alap SVM-től, hogy az idő- és energiaigényes kvadratikus programozás helyett lineáris egyenletrendszerre vezeti vissza a megoldandó problémát. Ezáltal a számítási idő jelentősen csökken.

A kész osztályozó kiértékeléséhez teszthalmazt használtunk. Vizsgálatunkban az osztályozáshoz az LS-SVM függvénykészletet használtuk (MATLAB-implementáció; Chang–Lin 2012) az úgynevezett radiális bázis (RBF – Radial Basis Function) kernelfüggvénnyel. Így a szupport vektor gépnek két szabadon állítható paramétere van: C a hibázási paraméter (penalty parameter) és γ az RBF-kernelfüggvény (Gauss-függvény) szórásparamétere. Érdemes először egy úgynevezett keresztvalidációs eljárással (cross-validation) és egy optimalizáló eljárással (simplex method) kizárólag a tanítóhalmazon beállítani az SVM-tanítás említett paramétereit (Hsu et al. 2003). Az SVM számos lehetséges C és γ paraméterpárjára kimerítő kereséssel találhatjuk meg az optimális beállítást, vagyis amikor az SVM a legnagyobb felismerési arányt éri el. Hsu és munkatársai (2003) szerint a C és γ értékeket az alábbi tartományokban érdemes keresni:

$$C: \{2^{-5}; 2^{-3}; \ldots; 2^{13}; 2^{15}\}$$
$$\gamma: \{2^{-15}; 2^{-13}; \ldots; 2^{1}; 2^{3}\}.$$

5.4.3.1.1. Az SVM tanítási paraméterei

Az SVM az átfedő és nem átfedő beszédrészek osztályozására úgy alkalmazható, hogy a korpusz minden beszédszegmensére kinyerjük az akusztikai jellemzőket, majd a tanító halmaz értékeivel tanítjuk be az osztályozót.

Az SVM tanításához a 8056 átfedő beszédszegmens 2/3-át, vagyis 5370-et használtunk fel, míg a teszteléshez az 1/3-át, amely 2386 szegmenst jelent. A korpuszban az átfedő beszédszegmensek előfordulása alacsonyabb volt, ezért a nem átfedő beszédrészek számát ehhez igazítottuk a tanító adatbázisban (random kiválasztási módszerrel). Erre azért volt szükség, hogy az algoritmus ne tanuljon rá jobban az egyik csoportra.

Az SVM bemeneti vektora tehát *i)* a spektrumra: 80×9; *ii)* az MFCC-re: 39×9; *iii)* az MSL-re 39×9; *iv)* a részsávenergiára: 12×9.

Az SVM RBF-függvényének két szabad paraméterét, a *C*-t és a *gammát* háromszoros keresztvalidációval és softmax függvénnyel optimalizáltuk.

6. Eredmények

6.1. A beszélőszegmentálás eredménye az alapbeállítások mellett

Ebben a fejezetben az általunk kialakított beszélődetektáló rendszert teszteljük különféle beállítások mellett. A beszélődetektálás működésének kiértékelésekor nagy különbségek lehetnek a tesztelésre használt korpusz függvényében. Ennek kiküszöbölésére hoztak létre standard korpuszokat, így az új beszélődetektáló algoritmusokat azonos korpuszon lehet tesztelni, ezzel összehasonlíthatóvá válnak az egyes eredmények, algoritmusok. Mivel ezen korpuszok többsége nem ingyenes, és hozzáférésük nem állt rendelkezésünkre, ezért az általunk használt algoritmust csak a BEA adatbázison teszteltük, így az eredményeink pusztán erre a korpuszra korlátozódnak.

A kiértékeléshez a NIST által javasolt md-eval-21.pl algoritmust használtuk (lásd a *Kiértékelési módszer* című fejezetet), amellyel minden tesztfájlra meghatároztuk a DER-értéket (Diarization Error Rate).

A standard BIC beszélődetektáló rendszerben MFCC teljes spektrumot lekódoló jellemzőt használunk, a BIC λ paraméterét 0-ra állítottuk, és nem használtunk sem szünetmodellt, sem egyszerrebeszélés-modellt a beszélődetektáláshoz.

A standard BIC beszélődetektáló átlagos eredménye 39,43%-os DER. Ez azt jelenti, hogy 60,56%-ban helyesen szegmentál és klaszterez a kiinduló algoritmusunk *(6.1. táblázat)*.

6.1. táblázat
A standard BIC beszélődetektálóval elért eredmények

A felvétel sorszáma	**A beszédfordulók száma**	**Teljes időtartam (s)**	**DER BIC (λ=0)**
			MFCC
bea071n	55	919,5	22,84%
bea072n	46	1020,4	35,92%
bea073n	23	590,5	43,38%
bea074n	25	1053,3	30,02%
bea075n	16	887,6	41,25%
bea094f	31	799,5	39,25%
bea150n	32	769,7	44,55%
bea166f	50	982,4	36,38%
bea174n	46	773,0	49,45%

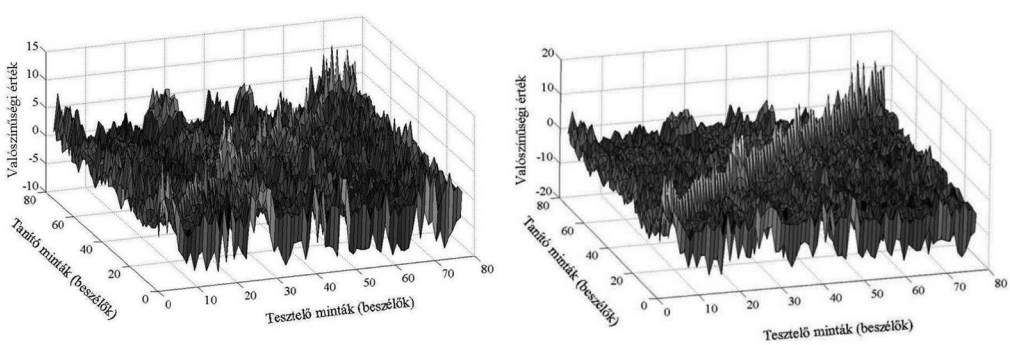
A felvétel sorszáma	A beszédfordulók száma	Teljes időtartam (s)	DER BIC (λ=0)
			MFCC
bea184n	48	599,4	46,21%
bea189n	68	973,1	48,37%
bea192f	50	816,2	42,60%
Átlag	**40,83**	**848,71**	**39,43%**

6.2. A BIC beszélődetektáló beszélőspecifikus akusztikai jellemzővel

6.2.1. Beszélőspecifikus jellemzők

Az egyes beszélők automatikus felismeréséhez különböző számú Gauss-komponenst tartalmazó GMM-eket használtunk a tanítás és a tesztelés során. A kutatás során teszteltük azt is, hogy mely MFCC-együtthatóval a legsikeresebb az osztályozás: *i)* teljes spektrumot kódoló MFCC; *ii)* 1,5–2,5 kHz közötti MFCC; *iii)* 2,5–3,5 kHz közötti MFCC; *iv)* 3,5–4,5 kHz közötti MFCC.

A Gauss-komponensek függvényében a valószínűségi érték az azonos beszélők esetében egyre magasabb, míg a különböző beszélők esetében ez csökken *(6.1. ábra)*.



6.1. ábra

A valószínűségi érték 2 komponensű Gauss (balra) és 256 komponensű Gauss esetén (jobbra)

Az ábrán jól látható, hogy a magasabb komponensszámú GMM esetében a felismerési mátrixban hogyan emelkedik ki az átló (ahol az azonos beszélők vannak), míg a körülötte lévő területek lecsökkennek.

Az eredmények azt mutatják *(6.2. táblázat),* hogy ha a GMM-et általános háttérmodellel (UBM) használjuk, akkor átlagosan jobb eredményeket kaptunk, mint a GMM általános háttérmodell nélkül.

6.2. táblázat
A felismerés pontossága (%) az osztályozónak a UBM megléte vagy hiánya, a Gauss-komponensek száma és az akusztikai jellemző függvényében

		A Gauss-komponensek száma					
Osztályozó	**Jellemző**	**8**	**16**	**32**	**64**	**128**	**256**
GMM	$MFCC_{(full\text{-}band)}$	28,81	34,01	56,08	69,46	72,46	75,66
	$MFCC_{(1,5–2,5)}$	26,32	32,58	54,11	67,61	69,39	72,01
	$MFCC_{(2,5–3,5)}$	*33,72*	*39,71*	*60,20*	*72,89*	*76,44*	*77,12*
	$MFCC_{(3,5–4,5)}$	**26,85**	**30,01**	**56,08**	**67,46**	**70,46**	**70,66**
GMM-UBM	$MFCC_{(full\text{-}band)}$	**29,15**	**34,35**	**55,42**	**74,8**	**77,81**	**76,76**
	$MFCC_{(1,5–2,5)}$	**30,81**	**32,01**	**61,08**	**71,46**	**75,78**	**72,60**
	$MFCC_{(2,5–3,5)}$	**34,05**	**35,21**	**66,53**	**74,91**	**76,11**	**79,76**
	$MFCC_{(3,5–4,5)}$	31,15	33,35	57,42	74,55	75,81	74,25

Megvizsgáltuk, hogy az $MFCC_{(2,5–3,5)}$ jellemzővel elért teljesítmény szignifikánsan jobbnak mondható-e. A statisztikai elemzés szerint ez a különbség szignifikáns: Wilcoxon-próba: $Z = -2{,}944$; $p = 0{,}003$. Megvizsgáltuk azt is, hogy mely akusztikai jellemzővel tanított osztályozó adja a legjobb eredményt. A *6.2. táblázat*ból látható, hogy a legjobb osztályozási arányt a 2,5–3,5 kHz részsávra számolt MFCC-együtthatókkal értük el mind a GMM, mind a GMM-UBM esetében. Ez azonban statisztikailag csak részben igazolható. Az $MFCC_{(2,5–3,5)}$ jellemzővel elért eredmények szignifikánsan különböznek az $MFCC_{(1,5–2,5)}$-vel ($Z = -2{,}201$; $p = 0{,}028$) és az $MFCC_{(3,5–4,5)}$-vel ($Z = -2{,}201$; $p = 0{,}028$) elért eredményektől, azonban a teljes spektrumot lekódoló eljárástól nem. Az adatokból azonban látszik, hogy szisztematikusan jobban teljesít az $MFCC_{(2,5–3,5)}$ jellemző, mint az $MFCC_{(full\text{-}band)}$. Ez az eredmény megerősíti a nemzetközi kutatások eredményeit, miszerint valóban a spektrum ezen régiója (2,5 kHz és 3,5 kHz) hordozza az egyéni beszédjellemzőket.

Elemeztük, hogy a felismerés pontossága hogyan alakul a Gauss-komponensek számának függvényében. Az eredményekből az látszik (vö. *6.2. táblázat*), hogy a Gauss-komponensek számának növekedésével javul a pontosság értéke is.

Összességében tehát elmondható, hogy a legjobb eredményt az $MFCC_{(2,5–3,5)}$ jellemzőt használó 256 Gauss-komponenst tartalmazó GMM-UBM osztályozóval értük el.

6.2.2. A beszélőspecifikus jellemzők implementálása a beszélődetektálóba

Előzetes kísérleteink szerint, ha az MFCC-jellemzőt specifikusan a 2,5–3,5 kHz-es részsávra számoljuk, akkor az eredmények javíthatók, hiszen eredményeink szerint ezen frekvenciatartomány tartalmazhatja a beszélőre specifikus akusztikai lenyomatokat. Ezért a standard BIC beszélődetektálóban ezt az akusztikai jellemzőt használtuk mint a standard beszélődetektáló módosítását. Az eredmények *(6.3. táblázat)* szintén igazolták, hogy az $MFCC_{(2,5–3,5)}$ akusztikai jellemző átlagosan jobban teljesít, mint az MFCC. A $MFCC_{(2,5–3,5)}$ jellemzővel 38,56% DER-értéket kaptunk, amely átlagosan 0,869%-os DER-javulást okozott. Csupán két esetben hozott az MFCC jobb eredményt (bea074n; bea94f). Jóllehet az átlagos javulás mértéke csupán 0,8%, ez a különbség szignifikáns (Wilcoxon-teszt Monte-Carlo-szimulációval kiegészítve: $Z = -2{,}824$; $p = 0{,}005$).

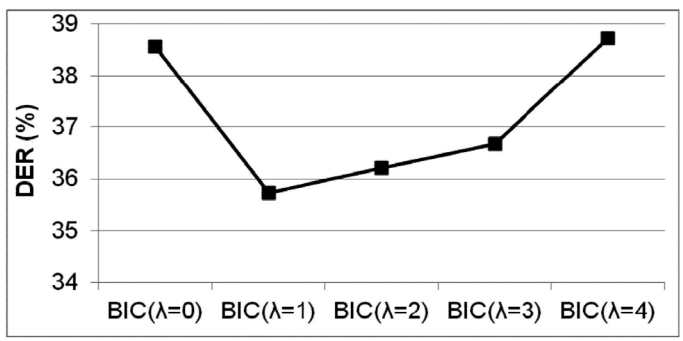
6.3. táblázat
A standard BIC beszélődetektálóval elért eredmények

A felvétel sorszáma	A beszédfordulók száma	Teljes időtartam (s)	DER BIC (λ=0)		ΔDER
			MFCC	$MFCC_{(2,5–3,5)}$	
bea071n	55	919,5	22,84%	21,21%	–1,63
bea072n	46	1020,4	35,92%	34,28%	–1,64
bea073n	23	590,5	43,38%	41,53%	–1,85
bea074n	25	1053,3	30,02%	30,33%	0,31
bea075n	16	887,6	41,25%	39,81%	–1,44
bea094f	31	799,5	39,25%	39,59%	0,34
bea150n	32	769,7	44,55%	42,63%	–1,92
bea166f	50	982,4	36,38%	34,14%	–2,24
bea174n	46	773,0	49,45%	47,48%	–1,97
bea184n	48	599,4	46,21%	44,46%	–1,75
bea189n	68	973,1	48,37%	46,03%	–2,34
bea192f	50	816,2	42,6%	41,24%	–1,36
Átlag	**40,81**	**848,71**	**39,43%**	**38,56%**	**–0,869**

A legjobb eredményt a bea071-es felvételére kaptunk, amelynek időtartama az egyik leghosszabb, így a beszélőfordulók száma is soknak mondható. A legrosszabb eredményt pedig a bea174n-es felvételre kaptuk. A bea071n-es felvételében egy idős nő az adatközlő, így hangja hallható módon különbözik a felvételvezetőétől, illetve a harmadik személyétől. A bea174n-es felvételen pedig egy fiatal felnőtt nő az adatközlő, akinek hangszínezete igen közeli a felvételvezetőéhez.

6.3. A BIC λ paraméterének optimális megválasztása

Teoretikusan a λ büntetőfaktor értéke zéró, amelyet a gyakorlatban sokszor 1-re szokás állítani (Ajmera et al. 2004). A jelen dolgozatban 0-tól 4-ig növeltük a λ paraméter értékét, és megvizsgáltuk, hogy hogyan változik a DER értéke. Akusztikai jellemzőként az $MFCC_{(2,5–3,5)}$-t használtuk. A tesztelés során a legjobb eredményt, vagyis a legkisebb DER-értéket akkor kaptuk, ha a BIC λ paraméterét 1-re állítottuk. Ekkor az átlagos beszélődetektálási hiba aránya 35,731% *(6.2. ábra)*.



6.2. ábra
A DER értékének alakulása a BIC λ paraméterének függvényében

6.4. A beszéddetektálás implementálása

6.4.1. A beszéddetektáló eredményei spontán társalgásban

A BEA adatbázisban (Gósy 2012) a nembeszéd részek átlagos időtartama 413 ms volt, a szórása pedig 438 ms.

Az általunk létrehozott beszéddetektáló algoritmust 5 órányi tanító adatbázison készítettük el. A tanító adatbázist az algoritmus által használt szabad paraméterek beállítására alkalmaztuk. A beszéddetektáló kialakítása után 36 órányi spontán beszéden futtattuk az algoritmust a teszteléshez.

A beszéddetektáló által használt paraméterek a következők voltak:

a) Jellemzőkinyeréshez: mindkét jellemzőt 10 ms-onként 50 ms hosszú ablakokon számoltuk, ahol az ablakok között nem volt átfedés.
b) 5 pontos mediánszűrést végeztünk a jellemzőkön kétszer (~250 ms).
c) A szegmentáláshoz 5 keretet, vagyis 250 ms hosszúságú ablakot használtunk.
d) Az utófeldolgozáshoz szintén 250 ms-os ablakot használtunk.

Az első kísérletben azt teszteltük, hogy milyen hosszú ablakhosszt kell optimálisan választani ahhoz, hogy a legjobb felismerési eredményt kapjuk. Az ablakhosszt (vö. *c)* pont) 1 kerettől 5 keretig növeltük, vagyis 25 ms-tól 250 ms-ig *(6.3. ábra)*. Ezzel egy időben azt is teszteltük, hogy melyik módszerrel tudunk elérni jobb detektálási eredményt (az alappal vagy az általunk javasolttal).

6.3. ábra

A DER értéke a keretek számának és a küszöböt meghatározó módszernek a függvényében

Az eredményekből az látszik, hogy a legkisebb hibát akkor kaptuk, hogyha a szegmentáláshoz 5 keretet, vagyis 250 ms-os hosszúságú ablakot használtunk. Ekkor a szegmentálási hiba értéke 9,51% volt. Mindemellett az eredményekből az is látszik, hogy az általunk javasolt *k*-középpel működő szegmentáló 3 keret hosszúságú ablaktól jobb eredményt ad, azonban ez a különbség nem szignifikáns.

A beszélődetektálási hibát (diarization error) felbontva láthatjuk *(6.4. táblázat),* hogy a legtöbb hiba abból adódik, hogy a gépi annotálásban sok helyen helytelen a beszéd vagy a szünet címkéje, vagy a szegmens helye megfelelő a beszédben, csak azonosítója téves.

6.4. táblázat

Az általunk javasolt algoritmus teljesítménye 250 ms-os ablakhosszúságú ablakozással

	MISS	***FA***	***SPKR***	***DER***
Az általunk ajánlott *k*-közép eljárás	0,1%	0,0%	9,4%	9,51%

6.4.2. A beszéddetektáló implementációja a beszélődetektálóba

Számos kutatás kimutatta, hogy a beszéddetektáló implementációja a beszélődetektálásba jelentősen csökkenti a DER értékét (Wei 2008). Ezért a jelen kutatásban az általunk létrehozott beszéddetektáló algoritmust implementáltuk a beszélődetektálóba. Az eljárás lényege, hogy a beszéddetektáló által detektált szünetrészeket már nem továbbítottuk a beszélődetektáló felé, vagyis töröltük a felvételből. Tehát jelen esetben a beszéddetektálót mint előfeldolgozó egységet csatoltuk a beszélődetektáló elé.

Az eredmények azt mutatják *(6.5. táblázat),* hogy a beszéddetektáló előfeldolgozásával a DER értékét átlagosan 4,535%-kal tudtuk csökkenteni, ami azt jelenti, hogy a beszélődetektáló DER-értéke 31,196%-os. Ez az átlagos javulás statisztikailag igazolható (Wilcoxon-teszt Monte-Carlo-szimulációval kiegészítve: $Z = -3{,}059$; $p < 0{,}001$).

6.5. táblázat
A DER értéke beszéddetektáló nélkül és beszéddetektálóval

A felvétel sorszáma	**DER**		**ΔDER**
	Beszéddetektáló nélkül BIC (λ=1)	**Beszéddetektálót használva BIC (λ=1)**	
bea071n037	18,60%	14,98%	–3,62
bea072n038	31,86%	27,83%	–4,03
bea073n039	39,94%	35,64%	–4,3
bea074n040	26,69%	23,89%	–2,80
bea075n041	38,14%	34,21%	–3,93
bea094f039	36,19%	33,63%	–2,56
bea150n091	39,63%	36,26%	–3,37
bea166f066	31,70%	27,74%	–3,96
bea174n105	43,72%	36,37%	–7,35
bea184n111	41,03%	35,07%	–5,96
bea189n114	43,53%	37,11%	–6,42
bea192f077	37,92%	31,8%	–6,12
Átlag	**35,73%**	**31,21%**	**–4,535**

6.5. Az egyszerrebeszélés-detektáló eredménye

6.5.1. Az egyszerrebeszélés-detektáló eredménye spontán társalgásban

Az egyszerre beszélések 12%-át teszik ki a teljes korpusznak, míg a szünetek 10,9%-át, így a beszédrészek 77,1%-át adják a teljes korpusznak.

A jelen kutatásban teszteltük, hogy a vizsgált négy akusztikai paraméter közül melyikkel lehet elérni a legjobb eredményt. Továbbá teszteltük azt is, hogy hogyan változik az eredményünk annak függvényében, hogy a mélyrétegű neuronhálózat harmadik rétegében hány neuront használunk.

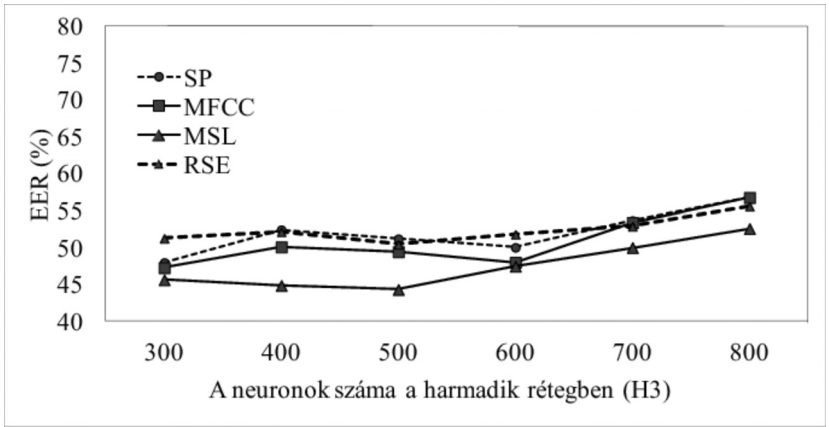
Az eredmények azt mutatják *(6.6. táblázat),* hogy a négy akusztikai paraméter: FFT-spektrum (SP); Mel-frekvenciás kepsztrális (MFCC) együtthatók; Mel-skála szerinti logaritmikus szűrőbank (MSL); részsávenergia (RSE) közül a legjobb teljesítményt akkor kaptuk, ha jellemzőként a Mel-skála szerinti logaritmikus szűrőbankot alkalmaztuk. Ekkor az Equal Error Rate (EER) átlagos értéke 47,49%, vagyis a helyesen felismert szegmensek aránya átlagosan 52,51%.

6.6. táblázat
Az átlagos EER értéke az akusztikai paraméterek függvényében

Származtatott jellemzők	Átlagos EER (%)
SP	52,03
MFCC	50,84
MSL	**47,49**
RSE	52,36

A második legjobban teljesítő jellemző az MFCC volt. Ennek átlagos EER-értéke 50,84% volt. Elmondható tehát az, hogy átlagosan 3,35%-os hibacsökkenést tudtunk elérni az MSL jellemző alkalmazásával az MFCC-vel elért eredményhez képest. Ez a javulás szignifikáns (Wilcoxon-próba: $Z = -2{,}211$; $p = 0{,}023$).

Megvizsgáltuk, hogy az EER értéke hogyan függ a jellemzők és a harmadik rétegben használt neuronok számától. Az eredmények azt mutatják, hogy a legjobb eredményt akkor kapjuk, ha MSL jellemzőt és 500 neuront használunk a H_3-ban *(6.4. ábra).*



6.4. ábra

Az EER értéke a jellemzők és a H_3-ban lévő neuronok számának függvényében

A statisztikai elemzések alátámasztják, hogy az MSL szignifikánsan jobban teljesít attól függetlenül, hogy hány neuront használunk a harmadik rétegben *(6.7. táblázat):* MSL-MFCC: $Z = -2{,}201$; $p = 0{,}028$; MSL-SP: $Z = -2{,}201$; $p = 0{,}028$; MSL-RSE: $Z = -2{,}201$; $p = 0{,}028$.

6.7. táblázat

Az EER értéke a jellemzők és a H_3-ban alkalmazott neuronok számának függvényében

	Neuronok száma a H_3 rétegben	**Akusztikai jellemzők**			
		SP	**MFCC**	**MSL**	**RSE**
EER (%)	300	**48,00**	**47,31**	45,65	51,27
	400	52,45	50,12	44,87	52,15
	500	51,22	49,44	**44,33**	**50,45**
	600	50,05	48,02	47,48	51,81
	700	53,68	53,41	50,01	52,91
	800	56,77	56,76	52,59	55,58
	900	52,03	50,84	47,49	52,36

Az EER-értékekből az látszik, hogy két esetben (SP és MFCC) akkor volt a legkisebb a hiba értéke, ha a harmadik rétegben 300 neuront használtunk. Az MSL és az RSE esetében pedig a legkisebb hibát akkor kaptuk, ha a neuronok száma 500 volt a harmadik rétegben. Általánosságban azonban az mondható el, hogy 500 neuron felett mindegyik jellemző esetében nőtt az EER értéke.

Az elért eredményeinket visszaellenőrizve elemeztük a hibák tulajdonságait. Az első és legnagyobb hibaforrás maga a kézi címkézés volt. Az egyszerre beszélések címkézése ugyanis sokszor igen nehéz feladat. A második hibaforrás a háttércsatorna-jelzésekre vezethető

vissza, a legtöbb hibát, 38,28%-ot ezek okozták. Ez a nagyszámú hiba annak tudható be, hogy a háttércsatorna-jelzések időtartama igen rövid, akár 60 ms-os is lehet, amely nem teszi lehetővé az elégséges számú jellemző kinyerését, így a belőlük származtatott statisztikai mutatók sem megbízhatók.

A háttércsatorna-jelzések után a nevetés volt az a jelenség, ami rontotta az osztályozás eredményét. Az ilyen típusú hibák aránya 10,34% volt. Ennél a hibánál is jól látható, hogy a nevetés közben az akusztikumban igen erős torzulás jelenik meg, a felvétel sokszor túlvezéreltté válik, így az akusztikai jellemző kinyerése nehezített.

6.5.2. Az egyszerrebeszélés-detektáló implementációja a beszélődetektálóba

Az eddigi kutatások alapján, noha az egyszerre beszélés detektálásának az eredménye jóval elmarad a kívánttól, a beszélődetektálóba való integráció során a DER értéke csökkenthető. Például Jin (2007) disszertációjában közel felére tudta csökkenteni a DER értékét, ha az audiofájlokból kivette az egyszerre beszéléseket tartalmazó részeket.

A jelen alfejezetben ennek a lehetőségét kívánjuk megvizsgálni, ezért az egyszerrebeszélés-detektálót implementáltuk az általunk létrehozott beszélődetektálóba. Hasonlóan a beszéddetektálóhoz, az egyszerre beszélések detektálóját úgy alkalmaztuk, hogy az általa generált kimenet alapján a társalgásból kivágtuk azon részeket, ahol egyszerre több beszélő szólalt meg. Tehát jelen esetben az egyszerrebeszélés-detektálót mint előfeldolgozó egységet csatoltuk a beszélődetektáló elé, a beszéddetektáló egység után.

Az egyszerre beszélés automatikus detektációjával átlagosan 2,49%-os relatív javulást tudtunk elérni, vagyis a DER értékét 31,21%-ról le tudtuk csökkenteni 28,71%-ra *(6.8. táblázat)*. Ez a javulás szignifikáns (Wilcoxon-teszt Monte-Carlo-szimulációval kiegészítve: $Z = -3{,}06$; $p = 0{,}002$).

6.8. táblázat

A DER értéke egyszerrebeszélés-detektáló nélkül és egyszerrebeszélés-detektálóval

A felvétel sorszáma	DER		ΔDER	Az egyszerre beszélés és a társalgás hosszának aránya
	Egyszerre beszélést			
	tartalmaz	nem tartalmaz		
bea071n037	14,98%	12,35%	–2,623%	21,52%
bea072n038	27,83%	24,85%	–2,98%	38,62%
bea073n039	35,64%	33,7%	–1,94%	15,68%
bea074n040	23,89%	20,71%	–3,18%	44,28%
bea075n041	34,21%	32,79%	–1,42%	6,46%
bea094f039	33,63%	31,88%	–1,75%	13,39%

A felvétel sorszáma	DER		ΔDER	Az egyszerre beszélés és a társalgás hosszának aránya
	Egyszerre beszélést			
	tartalmaz	nem tartalmaz		
bea150n091	36,26%	34,6%	–1,66%	28,96%
bea166f066	27,74%	25,59%	–2,15%	31,67%
bea174n105	36,37%	33,31%	–3,06%	40,26%
bea184n111	35,07%	30,69%	–4,38%	38,99%
bea189n114	37,11%	33,55%	–3,56%	42,53%
bea192f077	31,8%	30,54%	–1,26%	40,66%
Átlag	**31,21%**	**28,71%**	**–2,49%**	**30,94%**

Elemeztük, hogy a teszteléskor használt társalgásokban milyen arányban fordulnak elő egyszerre beszélések *(6.8. táblázat)*. A táblázatban látható, hogy elég gyakoriak az egyszerre beszélések ezeken a felvételeken. Jóllehet az egyszerre beszéléseket detektáló algoritmus eredményei nem voltak túl magasak, mégis statisztikailag igazolható relatív javulást tudtunk elérni a beszélődetektálóba való implementációval.

7. Következtetések

A jelen kutatás fő célja az volt, hogy magyar nyelvre elsőként hozzon létre spontán társalgásokra felügyelet nélküli tanuláson alapuló beszélődetektáló algoritmust. A kutatás egyik fő kérdése az volt, hogy milyen eredménnyel tudjuk megvalósítani a beszélődetektálót a spontán társalgásokra. Hogyan valósíthatók meg a beszélődetektálás egyes előfeldolgozó rendszerei, mint a beszéddetektálás, egyszerrebeszélés-detektálás, illetve hogy ezek milyen eredménnyel implementálhatók a beszélődetektáló rendszerbe. Arra is kerestük a választ, hogy melyek azok az akusztikai jellemzők, amelyek az egyénre jellemző akusztikai lenyomatokat tartalmazhatják. Vizsgáltuk, hogy milyen eredménnyel lehet az egyszerrebeszélés-detektálót implementálni a beszélődetektálóba. Elemeztük, hogy a beszélőszegmentálásban milyen beállítások mellett kapjuk a legjobb eredményt.

7.1. Beszéddetektáló

Ebben a vizsgálatban a Giannakopoulos (2009) által kidolgozott és MATLAB-ba implementált beszéddetektáló algoritmust használtuk, illetve módosítottuk. Ez az algoritmus rövid idejű energiafüggvény (short-term energy), spektrális centroid (spectral centroid) akusztikai jellemzőket és adaptív küszöbölést alkalmaz a beszéd és nembeszéd szegmensek automatikus meghatározására. Az általunk ajánlott módszer annyiban tér el ettől, hogy a küszöb meghatározását (beszéd és nembeszéd) felügyelet nélküli tanulási metódussal végezzük el, *k*-közép algoritmussal.

A cél az volt, hogy automatikusan meghatározzuk az egyes jelszegmensekre, hogy beszéd- vagy nembeszéd szegmens-e, illetve hogy teszteljük, hogy az általunk javasolt felügyelet nélküli tanulási módszer javít-e az eredményeken.

100 társalgásban (ami 5 órányi anyagot jelent) manuálisan jelöltük azokat a részeket, ahol valamelyik adatközlő beszél, illetve azokat a részeket, ahol nincs beszédjel, vagyis néma szünet van. A korpusz 49 órányi beszédrészt és 6 órányi szünetet tartalmaz, vagyis a teljes korpusz 10,9%-át a szünetek teszik ki. A beszéddetektáló kiértékelése a NIST által javasolt DER-metódussal történt.

Az eredmények azt mutatták, hogy az általunk javasolt módszerrel a felismerési hiba csökkenthető, statisztikailag azonban a javulás nem igazolható. Feltételezzük, hogy más klaszterező eljárással, például fuzzy klaszterezéssel az eredményeken javítani lehet.

Az általunk javasolt rendszer jó minőségű felvételen 90,49%-os eredménnyel működik.

Az elkészített beszéddetektálót az általunk fejlesztett beszélődetektálóba integráltuk.

7.2. Beszélőspecifikus jellemzők a gépi beszélőfelismerésen keresztül

A kutatás egyik célja az volt, hogy megvizsgálja, a magyar nyelvű beszédben mely spektrális régiók beszélőspecifikusak. Második célja az volt, hogy a beszélőket MFCC-vel előfeldolgozva GMM-ekkel, illetve GMM-UBM-ekkel modellezze és osztályozza a spontán beszédük alapján.

A kutatás célja, hogy olyan beszélőosztályozót hozzunk létre, amely szövegfüggetlen, és spontán beszédben képes a beszélőket automatikusan osztályozni. A kapott eredményeket (főként az akusztikai jellemzőkre vonatkozókat) az általunk fejlesztett beszélődetektálóba integráltuk.

A kutatásban a BEA adatbázisból választottunk ki 100 középkorú beszélőt (42 férfi és 58 női adatközlő). A tanító adatbázishoz minden adatközlő beszédéből kivágtunk egy 25 másodperces részt. A tesztadatbázishoz minden beszélő beszédéből kivágtunk egy 13 másodperces részt. A beszélőfelismeréshez MFCC jellemzőket (Mel Frequency Cepstral Coefficients) és GMM-UBM (Gaussian Mixture Model – Universal Background Model) algoritmust alkalmaztunk. A beszélőfelismerőt MATLAB szoftverben valósítottuk meg. Az MFCC kinyerését kétféleképpen végeztük el. Az egyik eljárásban az MFCC-t a beszédjel teljes spektrumára számoltuk ki (full-band spectral based MFCC). A másik akusztikai jellemző a spektrumból egy-egy tartományra koncentrálódik; részsávú kódolás (sub-band coding – SBC). Három részsávra számoltuk ki a Mel-frekvenciás kepsztrális együtthatókat: 1,5–2,5 kHz, 2,5–3,5 kHz, 3,5–4,5 kHz. Ezt úgy állítottuk elő, hogy a Mel-skála szerinti kritikus sávszélességű szűrősor karakterisztikáját ezekre a tartományokra állítottuk.

A beszélőszemély-felismerésben az eredmények azt mutatják, hogy a spektrumban a 2,5 kHz és a 3,5 kHz közé eső frekvenciatartomány őrzi a beszélő személyre utaló akusztikai jegyeket. Ez az eredmény megerősíti a nemzetközi kutatások eredményeit.

Az eredmények továbbá azt is igazolták, hogy a hagyományos GMM algoritmussal elért eredmények, a külföldi szakirodalomban leírtakkal összhangban, javíthatók az univerzális háttérmodell (UBM) használatával. A legjobb eredményt akkor értük el, ha 256 komponenst tartalmazó GMM-UBM-et használtunk, aminek értéke 79,76% volt. Eredményeink azt is mutatják, hogy a Nikléczy–Gósy (2008) által megállapított 16 s-nál rövidebb, 13 s-os rész is elégséges ahhoz, hogy a beszélőket alacsony hibaaránnyal tudjuk automatikusan felismerni a beszédhang alapján.

A kutatás eredményei felhasználhatók a kriminalisztikai fonetikában, illetve a beszélőfelismerés gyakorlatában.

Eredményeink javítására újabb kísérletet tervezünk, amely több adatközlővel történik, más akusztikai jellemzőket és más mintaillesztési eljárást használ.

7.3. Az egyszerre beszélések automatikus osztályozása spontán magyar társalgásokban

A kutatás célja az volt, hogy a spontán társalgásokban modellezze az egyszerre beszéléseket, és automatikus osztályozó algoritmussal különítse el azoktól a beszédszakaszoktól, ahol csak egy társalgó beszél. 100 társalgást (55 órányi társalgást) manuálisan egyszerre beszélésekre annotáltunk. A társalgásokban minden esetben három személy vett rész. Ebből két társalgó állandó volt (2 nő, életkoruk 33 év). A harmadik személy 43 férfi és 67 nő közül került ki, átlagos életkoruk 35 év. Összesen 8056 olyan időintervallum található, ahol kettő vagy több résztvevő szólal meg egyszerre, vagyis ahol átfedő beszéd van. Az egyszerre beszélések magas, 12%-os előfordulása a korpuszban indokolja, hogy a beszélődetektálásban foglalkozzunk ezen jelenség automatikus osztályozásának lehetőségével. Jóllehet az egyszerre beszélések automatikus osztályozása igen fontos feladat a beszélődetektálásban, mégis csak néhány tanulmány foglalkozik ezzel a kérdéssel (például Mowlaee et al. 2010; Saeidi et al. 2010). Boakye és munkatársai (2008) az AMI korpuszon (amely 18%-ban tartalmaz átfedő beszédet) 38%-os F-értéketet értek el az átfedő beszéd detektálására. Yella és Valente (2012) munkájukban azt a jelenséget igyekeztek modellezni, hogy a társalgásokban az átfedő beszédek előtt rövidebb a szünet (szüneteloszlás modellezése), mint a beszélőváltáskor. Az ezt modellező (HMM/GMM) metódussal a beszélődetektálás DER-értékét 8%-kal tudták csökkenteni. Prozódiai jellemzőket is tartalmazó eljárással Zelenak és Hernando (2011) hasonló F-score-t tudtak elérni az átfedőbeszéd-detektálásra, amely közel 40%-os volt. Vipperla és munkatársai (2012) konvolúciós nemnegatív ritka kódolással (convolutive non-negative sparse coding) az átfedőbeszéd-detektálásra 16,1%-os fedést és 28%-os pontosságot tudtak elérni a NIST RT korpuszon telefonbeszélgetésekre. Ben-Harush és munkatársai (2010) az időtartományban adott entrópiajellemzők becslésével próbálták meg detektálni az egyszerre beszéléseket (ez a munka csak kétbeszélős társalgásokat elemzett).

Yella és Bourlard (2013) Shriberg 2001-es kutatási eredményeiből indultak ki, amely azt a megfigyelést írta le, hogy az átfedő beszédrészek előfordulása jóval gyakoribb a társalgások egy bizonyos részén. A megfigyelés arra is kiterjedt, hogy az átfedő beszéd megjelenése összefügg a beszédfordulók számával. Ezt a jelenséget kihasználva Yella és Bourlard (2013) egy olyan algoritmust fejlesztettek, amely ezt a jelenséget modellezi. Az általuk javasolt egyszerrebeszélés-detektálót beépítették a beszélődetektálóba, amellyel 5%-os relatív DER-javulást tudtak elérni.

A fent leírt eredményekből látszik, hogy habár az egyszerre beszélések detektálásának eredménye jóval elmarad a kívánttól, a beszélődetektálóba való integráció során a DER értéke csökkenthető.

Mivel sem az akusztikai jellemzőben, sem a detektáló algoritmus típusában nincs megegyezés, hogy melyik alkalmas az egyszerre beszélések detektálására, ezért a jelen kutatásban több akusztikai jellemzőt is teszteltünk, illetve egy olyan hibrid osztályozót hoztunk létre

(DBN/SVM, Deep Belief Nets/Support Vector Machine, mély belief háló/szupport vektor gép), amelyet igen hatékonyan alkalmaztak már más típusú problémák megoldására (Tang 2008).

Jelen kutatás során a legjobb eredményt a Mel-skála szerinti logaritmikus szűrőbankjellemző adta. Ez korrelál más kutatásokban is ezt a jellemzőt használó algoritmusok által elért eredménnyel, például beszédhang-felismerésben (Li et al. 2012; Mohamed et al. 2012). Ezen tanulmányok arról számoltak be, hogy a Mel-skála szerinti logaritmikus szűrőbankjellemző jobban teljesített, mint az MFCC.

Teszteltük azt is, hogy hány neuront kell alkalmazni a harmadik rétegben. Az eredmények ebben a tekintetben azt mutatták, hogy 500 neuron után az EER értéke növekszik. A legjobb eredményt akkor kaptuk, ha Mel-skála szerinti logaritmikus szűrőbankjellemzőt és $H_1(300)$–$H_2(600)$–$H_3(500)$ topológiájú DBN-t használtunk előfeldolgozásként, valamint SVM-RBF-et osztályozóként.

A jelen kutatás során feltételeztük, hogy automatikusan osztályozhatók az átfedő beszédrészek, vagyis azon részek a spontán beszédben, amikor egynél több résztvevő beszél. Az átfedő részek tehát MSL-lel jellemzőkinyerve, DBN-nel előfeldolgozva és SVM-mel osztályozva azonosíthatók a spontán társalgásokban. Az EER értéke 44,33%.

Eredményeink alapján kimutattuk, hogy ebben a feladatban nehézségeket okoznak a háttércsatorna-jelzések és a nevetések, mivel ezek eredményezték a hibák többségét. Megjegyezzük viszont, hogy számos gyakorlati alkalmazás szempontjából – például ha az egyszerrebeszéd-detektálót beszédfelismerő előtt alkalmazzuk szűrőként a beszéddetektáló kiegészítésére – kifejezetten előnyös lehet, ha az egyszerre beszélések mellett más, a felismerés kivitelezését lehetetlenné tévő események – így például a nevetés, bizonyos háttércsatorna-jelzések – is detektálhatók (Neuberger–Beke 2013). Ezen beszédesemények törlésével az EER értéke jóval alacsonyabb lehet. Az egyszerre beszélés és egyéb események esetleges elkülönítése további osztályozással is megvalósítható, erre azonban a jelen munkában nem tértünk ki.

7.4. Beszélődetektálás

A beszélődetektáláshoz először megvizsgáltuk a kiválasztott részkorpusz jellemzőit: a beszédfordulók számát és időtartamát tekintve. Elemeztük továbbá, hogy van-e valamilyen különbség a társalgásban betöltött szerep vagy a nemek tekintetében.

Az általunk random módon kiválasztott 100 társalgásban 7827 db beszédfordulót adatoltunk. Egy felvételre átlagosan 70 db beszédforduló jut, amelynek szórása 41 db. A legtöbb beszédforduló 240 db volt, míg a legkevesebb 11 db. Nemek tekintetében nem találtunk szignifikáns különbséget a beszédfordulók számában (a férfi adatközlők átlagosan 79 db beszédfordulót produkáltak, míg a női adatközlők 65 db-ot). A társalgásban betöltött szerepek szerint

az adatközlők átlagosan 40,3%-ban veszik magukhoz a szót. A felvételvezető átlagosan 33,9%-ban veszi magához a szót, míg a harmadik résztvevő csupán átlagosan 18,3%-ban. Ezek az arányok azt mutatják, hogy a társalgások során a szerepek nem kiegyenlítettek, a harmadik személy sokszor háttérbe szorul (ennek oka többféle lehet, például ismertségi fok). A beszédidőtartamban sem tudtunk szignifikáns különbséget kimutatni a nemek között (a férfiak 36%-ban, a nők 42%-ban tartják maguknál a szót a teljes időtartamhoz képest). Megvizsgáltuk, hogy a beszédidőtartamok és a beszédforduló/perc értékek hogyan függnek össze az egyes résztvevők függvényében. Az adatközlőknél nem lehet kimutatni semmilyen tendenciát. A kísérletvezető esetében azonban pozitív közepesen erős függvénykapcsolatot tudtunk kimutatni (Pearson-korreláció: $r = 0{,}424$; $p < 0{,}001$), s ugyanilyen tendenciát találtunk a harmadik résztvevő esetében is (Pearson-korreláció: $r = 0{,}441$; $p < 0{,}001$). Mindez azt mutatja, hogy az adatközlőnek nem kell törekednie a szóátvételre, hiszen a beszédkorpusz alapvető célja, hogy az adatközlőtől minél több mintát rögzítsen, míg a felvételvezetőnek és a harmadik személynek ahhoz, hogy minél több közlést hozzanak létre, annál többször kell magukhoz venniük a szót.

7.4.1. A beszélődetektáló alaprendszere

A beszélődetektálón belül a beszélőszegmentáláshoz a Bayesian Information Criterion (BIC: Bayes-féle információs kritérium) algoritmust használtuk. Akusztikai jellemzőként az MFCC-t kétféleképpen használtuk. Az MFCC együtthatókat 32 ms-os ablakhosszra számoltuk, 10 ms-onként. A téves riasztások kezelésére egy utófeldolgozó lépést használtunk, amely Kullback–Leibler-távolságon alapul. A beszélőklaszterezéshez szintén a BIC algoritmust alkalmaztuk mind a klaszterek közötti hasonlóság mérésére, mind megállási feltételként. A beszélőklaszterezésben a GMM-szupervektor PCA transzformáltját használtuk mint a beszélőklaszterezés bemeneti jellemzőjét.

Kísérletileg igazoltuk, hogy magyar nyelvű spontán társalgásokra alapvetően felügyelet nélküli tanulási eljárásokat felhasználva létre lehet hozni olyan minőségű beszélődetektáló rendszert, amely 39,43%-os DER-értékkel működik.

A jelen munka elsőként készített magyar nyelvű spontán társalgásokban alkalmazható beszélődetektálót, amely a standard BIC-beszélődetektálóval, MFCC teljes spektrumot lekódoló jellemzőt használva, a λ paraméterét 0-ra állítva, sem szünetmodellt, sem egyszerre-beszélés-modellt nem használva **39,43%**-os DER-eredménnyel működik.

7.4.2. Beszélőspecifikus akusztikai jellemzők implementálása

A *Beszélőspecifikus jellemzők a gépi beszélőfelismerésben* című fejezetben bemutattuk, hogy ha az MFCC jellemzőkinyerést 2,5 és 3,5 kHz-es részsávban valósítjuk meg, akkor a beszélőszemély-felismerés eredménye javítható. Ezt az akusztikai paramétert teszteltük

a beszélődetektálóban is. A beszélődetektálóban elért eredmények szintén igazolták, hogy az $MFCC_{(2,5–3,5)}$ akusztikai jellemző átlagosan jobban teljesít, mint az MFCC. A $MFCC_{(2,5–3,5)}$ jellemzővel **38,56%** DER-értéket kaptunk, amely átlagosan 0,869%-os DER-javulást okozott (39,43%-ról 38,56%-ra).

7.4.3. A BIC λ paraméterének beállítása

Bemutattuk, hogy hogyan lehet optimálisan megválasztani a BIC λ szabad paraméterét. A tesztelés során a legjobb eredményt, vagyis a legkisebb DER-értéket akkor kaptuk, ha a BIC λ paraméterét 1-re állítottuk. Ekkor az átlagos beszélődetektálás hibaaránya **35,73%** volt. Tehát a BIC λ paraméter megfelelő beállítása 2,83%-os DER-javulást okozott.

7.4.4. A beszéddetektálás implementálása

A *Beszéddetektálás* című fejezetben létrehozott beszéddetektálót implementáltuk a beszélődetektálóba. Az eredmények azt mutatták, hogy a beszéddetektáló előfeldolgozásával az DER értéke átlagosan 4,535%-kal csökkenthető. Tehát a beszéddetektáló implementálásával a rendszer **31,196%**-os DER-eredménnyel működik.

7.4.5. Az egyszerrebeszélés-detektáló implementálása

Az egyszerre beszélések automatikus osztályozása spontán magyar társalgásokban című fejezetben létrehozott algoritmust implementáltuk beszélődetektáló rendszerünkbe. Az átfedő beszédek automatikus detektálásával átlagosan 2,49%-os relatív javulást tudtunk elérni, vagyis a DER értékét 31,21%-ról le tudtuk csökkenteni **28,71%**-ra.

7.4.6. A kifejlesztett rendszer végső eredménye

Összességében elmondható, hogy a legjobb eredményt akkor kaptuk, ha a BIC λ paraméterét 1-re állítottuk, $MFCC_{(2,5–3,5)}$ akusztikai jellemzőt alkalmaztunk, és előfeldolgozásként implementáltuk mind a VAD, mind az egyszerrebeszélés-detektáló algoritmusokat. Ekkor a DER értéke **28,71%** volt.

8. Összegzés

A beszédtudomány alapvető kutatási célja a beszédkommunikáció komplex körfolyamatának leírása. A beszédtechnológiában a beszédkommunikáció egyes moduljainak mesterséges eszközökkel történő helyettesítése a cél: a beszédprodukcióra a beszédszintézis, a beszédészlelésre a beszédfelismerés (beszédmegértésről gépi oldalról még nincs szó). Az ember-gép kommunikáció megteremtésében nyilvánvaló a dialogikus forma, ahol az ember és a gép váltakozva nyilatkoznak meg. Ezt a dinamikus váltakozást modellező modul a beszélődetektálás.

A napjainkban egyre nagyobb figyelmet kapó beszélődetektálás megvalósítására számos lehetőség létezik. Több nyelven, de főként angol korpuszokra történtek kísérletek. Magyar nyelvű spontán társalgásokra azonban ez idáig még nem készült ilyen jellegű munka. A meglehetősen szerteágazó megoldások mellett még igen sok lehetőség van a beszélődetektálók fejlesztésére, eredményeik javítására. Ehhez szükség van az olyan szorosan kapcsolódó tudományterületek eredményeire, gyakorlati tapasztalataira, mint a fonetika, a pszicholingvisztika, a diskurzuselemezés stb. Az értekezés ezt a sokszínűséget kívánta bemutatni, rendezni és felhasználni a beszélődetektálás megvalósításában.

Eredményeink hozzájárulhatnak a beszédkommunikáció több szempontú vizsgálatához, amelyben a beszélőváltakozás automatikus detektálását igyekeztünk megvalósítani mesterséges eszközökkel.

9. Kitekintés

A további terveinkben szerepel, hogy az általunk létrehozott nevetésdetektálót (Neuberger–Beke 2013) is integráljuk a beszélődetektálóba, hogy ezzel is csökkentsük a hiba arányát.

Véleményünk szerint a beszédtechnológiai eszközök mellett igen hasznos lehet bevonni nyelvtechnológiai eszközöket is. Tervezzük egy automatikus diskurzusjelölő-detektáló létrehozását, amellyel a beszédfordulók egy része egyértelműsíthető lenne, csökkentve ezzel a téves riasztások számát.

Továbbá tervezzük, hogy az általunk kidolgozott rendszert más standard korpuszokon teszteljük, így összevethető lenne más, már létező beszélődetektáló algoritmusok eredményeivel.

A társalgások gépi feldolgozásának elengedhetetlen szerepe lehet a napjainkban egyre növekvő adatmennyiség automatikus feldolgozásában, újrarendszerezésében, amelyeknek nagy része beszélők szerint strukturálható. A társalgások gépi feldolgozásával számos új kérdést válaszolhatunk meg: a társalgások alapvető felépítéséről, mikro- és makrostruktúrájáról; a társalgás alatt mutatott viselkedések és beszélői szerepek vizsgálatával jobban megérthetjük a beszélők közötti kapcsolatokat. Ezek elemzésével megalkothatók a beszélői profilok. A beszélői szerepek és viselkedés által feltárható az interakciós szekvenciák természete. Mindezek mellett számos új algoritmus fejlesztésére van lehetőség, mint a napjainkban egyre nagyobb figyelmet kapó topikváltás-detektáló, információkinyerő algoritmus és a beszédstílus-detektáló. A kutatásban fontos szerepet kap a beszélt nyelv szintaxisának kérdése, illetve annak automatikus elemzésének a lehetősége.

Mindezek mellett a beszélődetektálás fontos szerepet játszhat a dokumentum-visszakeresésben, a tartalomkinyerésben vagy a kérdés-válasz rendszerekben. Az ilyenfajta megközelítések új ismereteket nyújthatnak a társalgások felépítéséről és a társas viszonyokról.

Ezek a kutatások a valós nyelvhasználatot írják le valós kommunikációs helyzetekben, így új megközelítések válnak lehetővé és újabb kérdések fogalmazhatók meg a szélesebb nyelv- és beszédtechnológiai kutatásokban is (például a beszédfelismerés eredményének javítása, a spontán beszéd grammatikája, nyelvtipológia, univerzálék).

A beszélődetektálással foglalkozó kutatások eredményei hozzájárulnak az emberi viselkedés megértéséhez, illetve tovább mutatnak az ember-gép kommunikáció gépi modellezése felé.

10. Irodalom

Adami, André G. – Kajarekar, Sachin S. – Hermansky, Hynek 2002. A new speaker change detection method for two-speaker segmentation. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. USA, Florida, 3908–3911.

Ajmera, Jitendra 2004. *Robust audio segmentation*. PhD thesis. Ecole Polytechnique Federale de Lausanne, Lausanne. http://publications.idiap.ch/downloads/reports/2004/rr04–35.pdf (A letöltés ideje: 2013. szeptember 1.)

Ajmera, Jitendra – Bourlard, Hervé – Lapidot, Itshak 2002. *Unknown-multiple speaker clustering using HMM*. Technical report. IDIAP. http://publications.idiap.ch/downloads/reports/2002/ajmera2002icslp.pdf (A letöltés ideje: 2013. szeptember 1.)

Ajmera, Jitendra – McCowan, Iain – Bourlard, Hervé 2003. *Robust speaker change detection*. Technical report. IDIAP. http://publications.idiap.ch/downloads/reports/2002/rr02–39.pdf (A letöltés ideje: 2013. szeptember 1.)

Ajmera, Jitendra – McCowan, Iain – Bourlard, Hervé 2004. Robust speaker change detection. *IEEE Signal Processing Letters* 11/8. 649–651.

Ajmera, Jitendra – Wooters, Charles 2003. A robust speaker clustering algorithm. In: *Automatic Speech Recognition and Understanding Workshop, IEEE*. St. Thomas, US Virgin Islands, 411–416.

Anguera, Xavier 2005. *Xbic: Real-time cross probabilities measure for speaker segmentation.* Technical report. ICSI. http://www.xavieranguera.com/papers/techreport_xbic.pdf (A letöltés ideje: 2013. szeptember 1.)

Anguera, Xavier 2006. *Robust speaker diarization for meetings.* PhD thesis. Universitat Politecnica De Catalunya. http://nlp.lsi.upc.edu/papers/thesis_xanguera.pdf (A letöltés ideje: 2013. szeptember 1.)

Anguera, Xavier – Aguilo, Mateu – Wooters, Charles – Nadeu, Climen – Hernando, Javier 2006a. Hybrid speech/nonspeech detector applied to speaker diarization of meetings. In: *Proceedings of Speaker Odyssey Workshop*. Puerto Rico, USA, 1–6.

Anguera, Xavier – Hernando, Javier 2004. Evolutive speaker segmentation using a repository system. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Jeju Island, Korea. http://www.cs.upc.edu/~nlp/papers/anguera04.pdf (A letöltés ideje: 2013. szeptember 1.)

Anguera, Xavier – Wooters, Charles – Pardo, Jose M. 2006b. Robust speaker diarization for meetings: ICSI RT06s evaluation system. In: *Proceedings of International Conference on Speech and Language Processing 2006.* Pittsburgh, USA, 346–358.

Appel, Ulrich – Brandt, Achim V. 1982. Adaptive sequential segmentation of piecewise stationary time series. *Information Sciences* 29/1. 27–56.

Armani, Luca – Matassoni, Marco – Omologo, Maurizio – Svaizer, Piergiorgio 2003. Use of a CSP-based voice activity detector for distant-talking ASR. In: *Proceedings of European Conference on Speech Communication and Technology 2003.* Geneva, Switzerland, 501–504.

Atal, Bishnu S. 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America* 55/6. 1304–1312.

Attili, Joseph B. – Savic, Michael I. – Campbell, Joseph P. 1988. A TMS32020-based real time, text-independent, automatic speaker verification system. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. New York, USA, 599–602.

Auer, Peter 1996. On the prosody and syntax of turn-continuations. In Couper-Kuhlen, Elizabeth – Selting, Margret (eds.): *Prosody in Conversation: Interactional studies*. Cambridge University Press, Cambridge, UK, 57–100.

Avi, Matza – Yuval, Bistritz 2014. Skew Gaussian mixture models for speaker recognition. *IET Signal Processing*. Volume 8, Issue 8, October 2014, 860–867.

Bakis, Raimo – Chen, Scott – Gopalakrishnan, Ponani – Gopinath, Ramesh 1997. Transcription of broadcast news shows with the IBM large vocabulary speech recognition system. In: *Proceedings of the Speech Recognition Workshop*. 67–72.

Barras, Claude – Zhu, Xuan – Meignier, Sylvain – Gauvain, Jean-Lluc 2004. Improving speaker diarization. In: *Proceedings of Fall 2004 Rich Transcription Workshop (RT04),* Palisades, New York, USA, 67–72.

Basseville, Michèle – Nikiforov, Igor V. 1993. *Detection of abrupt changes: Theory and application*. Prentice-Hall, Englewood Cliffs, New Jersey.

Bata Sarolta 2009a. Beszélőváltások a beszédpartnerek személyes kapcsolatának függvényében. *Beszédkutatás* 2009. 107–120.

Bata Sarolta 2009b. A társalgás fonetikai jellemzőinek alakulása a beszédpartnerek életkorának függvényében. In Váradi Tamás (szerk.): *III. Alkalmazott Nyelvészeti Doktorandusz Konferencia*. Budapest, MTA Nyelvtudományi Intézet, 3–13.

Bata Sarolta – Gráczi Tekla Etelka 2009. Hatással van-e a beszédpartner életkora a beszélő beszédének szupraszegmentális jellegzetességeire? In Keszler Borbála – Tátrai Szilárd (szerk.): *Diskurzus a grammatikában, grammatika a diskurzusban*. Budapest, Tinta Kiadó, 74–83.

Beach, Wayne A. – Lindstrom, Anna K. 1992. Conversational universals and comparative theory: Turning to Swedish and American acknowledgement tokens in interaction. *Communication Theory* 2/1. 24–49.

Beattie, Geoffrey W. 1977. The dynamics of interruption and the filled pause. *The British Journal of Social and Clinical Psychology* 16/3. 283–284.

Beattie, Geoffrey W. 1982. Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica* 39. 93–114.

Beigi, Homayoon S. M. – Maes, Stéphane H. – Sorensen, Jeffrey S. 1998. A distance measure between collections of distributions and its application to speaker recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Detroit, USA, 12–15.

Beke András 2008. Az alapfrekvencia-eloszlás modellezése a beszélőfelismeréshez. *Alkalmazott Nyelvtudomány* 2008/1–2. 121–132.

Belin, Pascal – Fecteau, Shirley – bédard, Catherine 2004. Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences* 8/3. 129–135.

Bellili, Abdel – Gilloux, Michel – Gallinari, Patrick 2001. An hybrid MLP-SVM handwritten digit recognizer. In: *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*. 2001, Seattle, WA, USA, 28–32.

Ben, Mathieu – Betser, Michaël – Bimbot, Frédéric – Gravier, Guillaume 2004. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In: *Proceedings of International Conference on Speech and Language Processing*. Jeju Island, Korea. http://www.irisa.fr/metiss/guig/biblio/04/ben-interspeech-2004.pdf (A letöltés ideje: 2013. szeptember 1.)

Ben-Harush, Oshry – Guterman, Hugo – Lapidot, Itshak 2009. Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization. In: *Proceeding of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. Grenoble, France, 1–6.

Bimbot, Frédéric – Bonastre, Jean-François – Fredouille, Corinne – Gravier, Guillaume – Chagnolleau, Magrin Ivan – Meignier, Sylvain – Merlin, Teva – Garcia, Ortega Javier – Petrovska-Delacrétaz, Dijana – Reynolds, Douglas A. 2004. Tutorial on text-independent speaker verification. In: *Proceeding of EURASIP, Journal on Applied Signal Processing*. Vol. 4. New York, USA, 430–451.

Bishop, Christopher M. 1996. Theoretical foundations of neural networks. In Borcherds, Peter – Bubak, Marian – Maksymowicz, Andrzej (eds.): *Proceedings of Physics Computing, Academic Computer Centre*. Krakow, Poland, 500–507.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Springer, New York.

Boakye, Kofi A. 2008. *Audio segmentation for meetings speech processing*. PhD thesis. University of California at Berkeley. http://www.icsi.berkeley.edu/pubs/speech/boakye08_phd_thesis.pdf (A letöltés ideje: 2013. szeptember 1.)

Boakye, Kofi A. – Trueba-Hornero, Beatriz – Vinyals, Oriol – Friedland, Gerald 2008a. Overlapped speech detection for improved speaker diarization in multiparty meetings. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Las Vegas, USA, 4353–4356.

Boakye, Kofi A. – Vinyals, Oriol – Friedland, Gerald 2008b. Two'sa crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech. In: *Proceedings of International Conference on Speech and Language Processing 2008*. Brisbane, Australia, 32–35.

Boakye, Kofi A. – Vinyals, Oriol – Friedland, Gerald 2011. Improved overlapped speech handling for speaker diarization. In: *Proceedings of International Conference on Speech and Language Processing 2011*. Florence, Italy, 941–944.

Bonastre, Jean-François – Delacourt, Perrine – Fredouille, Corinne – Merlin, Teva – Wellekens, Christian J. 2000. A speakertracking system based on speaker turn detection for NIST evaluation. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Istanbul, Turkey, 1177–1180.

Boronkai Dóra 2008. Konverzációelemzés és anyanyelvtanítás I–II. *Anyanyelv-pedagógia* I/2. és I/3–4. szám. http://www.anyanyelv-pedagogia.hu/cikkek.php?id=60, http://www.anyanyelv-pedagogia.hu/cikkek.php?id=115 (A letöltés ideje: 2013. szeptember 1.)

Boronkai Dóra 2009. *Bevezetés a társalgáselemzésbe.* Ad Librum, Budapest.

Bőhm Tamás 2006. A glottalizáció szerepe a beszélő személy felismerésében. *Beszédkutatás 2006*. 197–207.

Bőhm Tamás 2007. Beszélőfelismerés – neurológiai háttér és pszichológiai modellek. *Magyar Pszichológiai Szemle* 62/4. 541–563.

Brown, Gillian – Yule, George 1989. *Discourse analysis.* Cambridge University Press. Cambridge – New York – Port Chester – Melbourne – Sydney.

Campbell, Joseph P. 1997. Speaker recognition: A tutorial. In: *Proceedings of the Institute of Electrical and Electronic Engineers*. Vol. 85, No. 9. 1437–1462.

Campbell, William M. – Campbell, Joseph P. – Reynolds, Douglas A. – Singer, Elliot – Torres-Carrasquillo, Pedro A. 2006. Support Vector Machines for speaker and language recognition. *Computer Speech and Language* 20/2–3. 10–29.

Çetin, Özgür – Shriberg, Elizabeth 2006a. Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Toulouse, France, 357–360.

Çetin, Özgür – Shriberg, Elizabeth 2006b. Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition. In: *Proceedings of International Conference on Speech and Language Processing 2006*. Pittsburgh, USA, 293–296.

Cettolo, Mauro – Vescovi, Michele 2003. Efficient audio segmentation algorithms based on the BIC. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 6. Hong Kong, China, 537–540.

Chafe, Wallace 1994. *Discourse, consciousness and time*. University of Chicago Press, Chicago.

Chang, Chih-Chung – Lin, Chih-Jen 2013. *LIBSVM: A library for support vector machines.* ACM Transactions on Intelligent Systems and Technology. http://www.csie.ntu.edu.tw/~cjlin/libsvm. (A letöltés ideje: 2013. június 5.)

Che, ChiWei – Lin, Qiguang 1995. Speaker recognition using HMM with experiments on the YOHO database. In: *Proceedings of European Conference on Speech Communication and Technology 1995.* Madrid, 625–628.

Chen, Scott S. – Gales, Mark J. F.– Gopinath, Ramesh A. – Kanvesky, Dimitri – Olsen, Peder A. 2002. Automatic transcription of broadcast news. *Speech Communication* 37. 69–87.

Chen, Scott S. – Gopalakrishnan, Ponani 1998a. Clustering via the Bayesian information criterion with applications in speech recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Seattle, USA, 645–648.

Chen, Scott S. – Gopalakrishnan, Ponani 1998b. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In: *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*. Virginia, USA, 127–132.

Cheng, Shi-sian – Wang, Hsin-min 2003. A sequential metric-based audio segmentation method via the Bayesian information criterion. In: *Proceedings of European Conference on Speech Communication and Technology 2003*. Geneva, Switzerland, 945–948.

Cheng, Shi-sian – Wang, Hsin-min 2004. METRIC-SEQDAC: A hybrid approach for audio segmentation. In: *Proceedings of International Conference on Speech and Language Processing*. Jeju Island, Korea, 1–5.

Cheng, Shih-Sian – Wang, Hsin-Min – Fu, Hsin-Chia 2008. BIC-based audio segmentation by divide-andconquer. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2008.* Las Vegas, USA, 4841–4844.

Cheng, Shih-Sian – Wang, Hsin-Min – Fu, Hsin-Chia 2010. BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* 18/1. Dallas, USA, 141–157.

Chengalvarayan, Rathinavelu 1999. Robust energy normalization using speech/non-speech discriminator for German connected digit recognition. In: *Proceedings of European Conference on Speech Communication and Technology 1999*. Budapest, Hungary, 61–64.

Chickering, Max – Heckerman, David 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning* 29. 181–212.

Cho, KyungHyun. The RBM code for Matlab developed by Cho, KyungHyun is used from http://users.ics.tkk.fi/kcho/ (A letöltés ideje: 2013. szeptember 1.)

Cho, Yong D. – Kondo, Ahmet 2001. Analysis and improvement of a statistical model-based voice activity detector. *IEEE Signal Processing Letters* 8/10. 276–278.

Cutler, Anne – Pearson, Mark 1986. On the analysis of prosodic turn-taking cues. In Johns-Lewis, Catherine (ed.): *Intonation in discourse*. College Hill, San Diego, California, 139–156.

Dahl, George E. – Ranzato, Marc'Aurelio – Mohamed, Abdel-Rahman – Hinton, Geoffrey E. 2010. Phone recognition with the mean-covariance restricted Boltzmann machine. In: *Proceeding of 24th Annual Conference on Neural Information Processing Systems 2010*. Vancouver, British Columbia, Canada (NIPS 2010). 469–477.

Daniel, Ellis P. W. 2005. *PLP and RASTA and MFCC, and inversion in Matlab.* http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/ (A letöltés ideje: 2013. június 5.)

Delacourt, Perrine – Kryze, David – Wellekens, Christian J. 1999a. Detection of speaker changes in an audio document. In: *Proceedings of European Conference on Speech Communication and Technology 1999*. 1195–1198.

Delacourt, Perrine – Kryze, David – Wellekens, Christian J. 1999b. Speaker-based segmentation for audio data indexing. In: *Proceedings of the ESCA Workshop Accessing Information in Spoken Audio*. Cambridge, UK, 78–83.

Delacourt, Perrine – Wellekens, Christian J. 2000. DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Communication: Special Issue in Accessing Information in Spoken Audio* 32. 111–126.

Dempster, Arthur P. – Laird, Nan M. – Rubin, Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39/1. 1–38.

Dér Csilla Ilona 2010. „Töltelékelem" vagy új nyelvi változó? A *hát, úgyhogy, így* és *ilyen* újabb funkciójáról a spontán beszédben. *Beszédkutatás 2010*. 159–170.

Dér Csilla Ilona 2012. Beszélőváltások során használt diskurzusjelölők a magyar spontán beszédben. *Beszédkutatás 2012*. 130–141.

Deshayes, Jean – Picard, Dominique 1986. Off-line statistical analysis of change-point models using non-parametric and likelihood methods. In Basseville, Michele – Benveniste, Albert (eds.): *Detection of abrupt changes in signals and dynamical systems.* Lecture notes in Control and Information Sciences 77. Springer-Verlag, Berlin, 103–168.

Dittmann, Allen T. – Llewellyn, Lynn G. 1968. Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology* 9. 79–84.

Doddington, George R. 1985. Speaker recognition – Identifying people by their voices. In: *Proceedings of the Institute of Electrical and Electronic Engineers* 73. 1651–1664.

Dommelen van, Wim A. – Moxness, Bente H. 1995. Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. *Language and Speech* 38. 267–287.

Drummond, Kent – Hopper, Robert 1993. Backchannels revisited: Acknowledgement tokens and speakership incipiency. *Research on Language and Social Interaction* 26. 157– 177.

Duncan, Starkey 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and social Psychology* 23. 283–292.

Duncan, Starkey – Fiske, Donald 1985. *Interaction structure and strategy*. Cambridge University Press, Cambridge.

Duncan, Starkey – Niederehe, George 1974. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology* 10. 234–247.

Fék Márk 1997. *Beszélőfelismerés neurális hálózatokkal és vektorkvantálóval*. OTDK-dolgozat. BME, Budapest.

Ferrer, Luciana – Shriberg, Elizabeth – Kajarekar, Sachin – Sönmez, Kemal 2007. Parametrization of prosodic feature distributions for SVM modeling in speaker recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Honolulu, Hawaii, 233–236.

Ford, Cecilia E. – Thompson, Sandra A. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In Ochs, Elinor – Schegloff, Emanuel A. – Thompson, Sandra A. (eds.): *Interaction and grammar.* Cambridge University Press, Cambridge, 134–184.

Friedland, Gerald – Vinyals, Oriol – Huang, Yan – Mueller, Christian 2009. Prosodic and other long-term features for speaker diarization. In: *Proceedings of IEEE Transactions on Speech and Audio Processing* 17/5. 985–993.

Furui, Sadaoki 1981. Cepstral analysis technique for automatic speaker verification. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* 29/2. Atlanta, USA, 254–272.

Furui, Sadaoki 2007. Recent advances in automatic speech summarization. In: *Proceedings of the IEEE/ACL Workshop on Spoken Language Technology.* Los Alamitos, USA, 115–122.

Gangadharaiah, Rashmi – Narayanaswamy, Balakrishnan – Balakrishnan, Narayanaswamy 2004. A novel method for two-speaker segmentation. In: *Proceedings of the International Conference on Speech and Language Processing*. Jeju Island, Korea, 65–78.

Gardner, Rod 2001. *When listeners talk*. John Benjamins Publishing Co., Amsterdam.

Garfinkel, Harold 1967. *Studies in ethnomethodology.* Prentice Hall, Englewood Cliffs, NJ.

Gauvain, Jean-Luc – Lamel, Lori – Adda, Gilles 1998. Partitioning and transcription of broadcast news data. In: *Proceedings of the International Conference on Speech and Language Processing.* Sidney, Australia, 1335–1338.

Gazor, Saeed – Zhang, Wei 2003. Speech probability distribution. *Signal Processing Letters, IEEE* 10/7. 204–207.

Giannakopoulos, Theodoros 2009. *Study and application of acoustic information for the detection of harmful content, and fusion with visual information*. PhD thesis. University of Athens, Greece. http://cgi.di.uoa.gr/~phdsbook/files/giannakopoulos.pdf (A letöltés ideje: 2013. szeptember 1.)

Gish, Herbert – Schmidt, Michael 1994. Text-independent speaker identification. *Signal Processing Magazine, IEEE* 11/4. 18–32.

Gish, Herbert – Siu, Man-Hung – Rohlicek, Robin 1991. Segregation of speakers for speech recognition and speaker identification. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Toronto, Canada, 873–876.

Gocsál Ákos 1998. Életkorbecslés a beszélő hangja alapján. *Beszédkutatás 1998*. 122–134.

Goffman, Erving 1983. The interaction order. *American Sociological Review* 48. 1–17.

Goodwin, Charles 1979. The interactive construction of a sentence in natural conversation. In Psathas, George (ed.): *Everyday language: Studies in ethnomethodology.* Irvington Publishers, New York, 97–121.

Górriz, Juan M. – Ramírez, Javier – Segura, José C. – Puntonet, Carios G. 2006. An effective cluster-based model for robust speech detection and speech recognition in noisy environments. *Journal of the Acoustical Society of America* 120/1. 470–481.

Gósy Mária 2001. A testalkat és az életkor becslése a beszéd alapján. *Magyar Nyelvőr* 125/4. 478–487.

Gósy Mária 2005. *Pszicholingvisztika.* Osiris Kiadó, Budapest.

Gósy Mária 2012. Multifunkcionális beszélt nyelvi adatbázis – BEA. In Prószéky Gábor – Váradi Tamás (szerk.): *Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások*. Akadémiai Kiadó, Budapest, 329–349.

Gósy Mária – Nikléczy Péter 1999. A beszélő felismerése a beszéde alapján: elméleti háttér és módszertani megközelítések. *Beszédkutatás 1999*. 1–19.

Grósz, Tamás – Tóth, László 2013. A comparison of Deep Neural Network training methods for large vocabulary speech recognition. In: *Proceedings of Text, Speech and Dialogue 2013.* Plzeň, Czech Republic, 36–43.

Hámori Ágnes 2006. A társalgási műfajokról. In Tolcsvai Nagy Gábor (szerk.): *Szöveg és típus. Szövegtipológiai tanulmányok.* Tinta Kiadó, Budapest, 157–181.

Hayashi, Reiko 1991. Floor structure of English and Japanese conversation. *Journal of Pragmatics* 16. 1–30.

Heck, Larry – Sankar, Ananth 1997. Acoustic clustering and adaptation for robust speech recognition. In: *Proceedings of European Conference on Speech Communication and Technology 1997*. Rhodes, Greece, 1867–1870.

Heritage, John 1984. A change-of-state token and aspects of its sequential placement. In Atkinson, J. Maxwell – Heritage, John (eds.): *Structures of social action: Studies in conversation analysis*. Cambridge University Press, Cambridge, 299–345.

Hermansky, Hynek 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* 87/4. 1738–1752.

Higgins, Allan L. – Bahler, Lawrence – Porter, Jack 1991. Speaker verification using randomized phrase prompting. *Digital Signal Processing* 1/2. 89–106.

Higgins, Allan L. – Wohlford, Robert E. 1986. A new method of text-independent speaker recognition. In: *Proceedings of the Institute of Electrical and Electronic Engineers, International Conference on Acoustics, Speech and Signal Processing*. Tokyo, Japan, 869–872.

Horváth Viktória 2009. *Funkció és kivitelezés a megakadásjelenségekben*. PhD-disszertáció. ELTE, Budapest.

Hsu, Chih-Wei – Chang, Lin 2003. *A practical guide to support vector classification.* Technical report. Department of Computer Science, National Taiwan University, Taiwan. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (A letöltés ideje: 2013. szeptember 1.)

Hung, Jeih-Weih – Wang, Hsin-min – Lee, Lin-shan 2000. Automatic metric based speech segmentation for broadcast news via principal component analysis. In: *Proceedings of the International Conference on Speech and Language Processing*. Beijing, China. 121–124.

Ida, Onshus 2011. *Indexing of audio databases: Event log of broadcast news*. PhD thesis. Norwegian University of Science and Technology, Trondheim.

ITU-T 1996. *Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear prediction (CS-ACELP).* Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70. International Telecommunication Union.

Iványi Zsuzsanna 2001. A nyelvészeti konverzációelemzés. *Magyar Nyelvőr* 125. 74–93.

Janin, Adam – Baron, Don – Edwards, Jane A. – Ellis, Dan – Gelbart, David – Morgan, Nelson – Peskin, Barbara – Pfau, Thilo – Shriberg, Elizabeth – Stolcke, Andreas – Wooters, Chuck 2003. The ICSI meeting corpus. In: *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing*. Hong Kong, China, 364–367.

Jefferson, Gail 1984. Notes on a systematic deployment of the ackowledgement tokens '*yeah*' and '*mm hm*'. *Papers in Linguistics* 17. 197–216.

Jin, Hubert – Kubala, Francis – Schwartz, Richard 1997. Automatic speaker clustering. In: *DARPA Speech Recognition Workshop*. Chantilly, USA, 108–111.

Jin, Qin 2007. *Robust speaker recognition*. PhD thesis. Carnegie Mellon University, Pittsburgh.

Jin, Qin – Laskowski, Kornel – Schultz, Tanja – Waibel, Alex 2004. Speaker segmentation and clustering in meetings. In: *Proceedings of NIST 2004 Spring Rich Transcrition Evaluation Workshop*. Montreal, Canada, 112–117.

Johnson, Sue E. 1999. Who spoke when? Automatic segmentation and clustering for determining speaker turns. In: *Proceedings of European Conference on Speech Communication and Technology 1999*. Budapest, Hungary, 2211–2214.

Johnson, Sue E. – Woodland, Phil C. 1998. Speaker clustering using direct maximization of the MLLR adapted likelihood. In: *Proceedings of International Conference on Speech and Language Processing* 5. Sydney, Australia, 1775–1779.

Joshi, Sachin – Prahallad, Kishore – Yegnanarayana, *Bayya* 2008. AANN-HMM models for speaker verification and speech recognition. In: *Proceedings of International Joint Conference on Neural Networks 2008 (IJCNN)*. Hong Kong, China, 2681–2688.

Juang, Biing-Hwang – Rabiner, Lawrence 1985. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal* 64/2. 391–408.

Kajarekar, Sachin S. – Ferrer, Luciana – Sönmez, Kemal – Zheng, Jing – Shriberg, Elizabeth – Stolcke, Andreas 2004. Modeling NERFs for speaker recognition. In: *Proceedings of Speaker Odyssey Workshop*. Toledo, Spain, 51–56.

Kass, Robert E. – Raftery, Adrian E. 1995. Bayes factors. *Journal of the American Statistics Association* 90. 773–795.

Kemp, Thomas – Schmidt, Michael – Westphal, Martin – Waibel, Alex 2000. Strategies for automatic segmentation of audio data. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Istanbul, Turkey, 1423–1426.

Kendon, Adam 1967. Some functions of gaze direction in social interaction. *Acta Psychologica* 26. 22–63. (Reprinted in Kendon, Adam 1990. *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge University Press, Cambridge, 51–89).

Kendon, Adam 2002. Some uses of the head shake. *Gesture* 2/2. 147–182.

Kim, Hyoung-Gook – Ertelt, Daniel – Sikora, Thomas 2005. Hybrid speaker-based segmentation system using model-level clustering. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Philadelphia, USA, 745–748.

Kiss Jenő 1995. *Társadalom és nyelvhasználat. Szociolingvisztikai alapfogalmak*. Nemzeti Tankönyvkiadó, Budapest.

Kotti, Margarita – Benetos, Emmanouil – Kotropoulos, Constantine 2006. Automatic speaker change detection with the Bayesian Information Criterion using MPEG-7 features and a fusion scheme. In: *Proceedings of IEEE International Symposium Circuits & Systems*. Island of Kos, Greece, 21–24.

Kotti, Margarita – Moschou, Vassiliki – Kotropoulos, Constantine 2008. Speaker segmentation and clustering. *Signal Processing* 88/5. 1091–1124.

Kubala, Francis – Jin, Hubert – Matsoukas, Spyros – Gnuyen, Long – Schwartz, Richard – Machoul, John 1997. The 1996 BBN byblos HUB-4 transcription system. In: *Proceedings of Speech Recognition Workshop*. 90–93.

Laskowski, Kornel – Schultz, Tanja 2006. Unsupervised learning of overlap speech model parameters for multichannel speech activity detection in meetings. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Toulouse, France, 993–996.

Lass, Norman J. – Hughes, Karen R. – Bowyer, Melanie D – Waters, Lucille T. – Bourne, Victoria T. 1976. Speaker sex identification from voiced, whispered and filtered isolated vowels. *Journal of the Acoustical Society of America* 59. 675–678.

Lerner, Gene H. 2003. Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society* 32/2. 177–201.

Levelt, Willem J. M. 1989. *Speaking: From intention to articulation*. A Bradford Book. The MIT Press, Cambridge (Massachusetts) – London (England).

Li, Jinyu – Yu, Dong – Huang, Jui-Ting – Gong, Yifan 2012. Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM. In: *Proceedings of IEEE Workshop on Spoken Language Technology 2012*. 131–136.

Li, K-Po – Wrench, Jr. Edwin H. 1983. An approach to text-independent speaker recognition with short utterances. In: *Proceedings of the Institute of Electrical and Electronic Engineers, International Conference on Acoustics, Speech and Signal Processing*. Boston, MA, 555–558.

Liu, Daben – Kubala, Francis 1999. Fast speaker change detection for broadcast news transcription and indexing. In: *Proceedings of European Conference on Speech Communication and Technology 1999*. Budapest, 1031–1034.

Local, John – Kelly, John 1986. Projection and "silences": Notes on phonetic and conversational structure. *Human Studies* 9. 185–204.

Lopez, Ferreiros – Ellis, Daniel P. W. 2000. Using acoustic condition clustering to improve acoustic change detection on broadcast news. In: *Proceedings of International Conference on Speech and Language Processing*. Beijing, China, 568–571.

Lu, Lie – Li, Stan Z. – Zhang, Hong-Jiang 2001. Content-based audio segmentation using support vector machines. In: *Proceedings of ACM International Conference on Multimedia*. Ottawa, Canada, 203–211.

Lu, Lie – Zhang, Hong-Jiang 2002a. Real-time unsupervised speaker change detection. In: *Proceedings of the International Conference on Pattern Recognition*. Quebec City, Canada, 358–361.

Lu, Lie – Zhang, Hong-Jiang 2002b. Speaker change detection and tracking in real-time news broadcasting analysis. In: *Proceedings of ACM International Conference on Multimedia*. uan les Pins, France, 602–610.

Lu, Lie – Zhang, Hong-Jiang – Jiang, Hao 2002. Content analysis for audio classification and segmentation. In: *IEEE Transactions on Speech and Audio Processing* 10/7. 504–516.

Maclay, Howard – Osgood, Charles E. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15. 19–44.

Malegaonkar, Amit S.– Ariyaeeinia, Aladdin M. – Sivakumaran, Perasiriyan – Fortuna, J. 2006. Unsupervised speaker change detection using probabilistic pattern matching. In: *Proceedings of IEEE Signal Processing Letters* 13/8. 509–512.

Markel, John D. – Davis, Steven B. 1979. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* Washington, USA, 74–82.

Markó Alexandra 2006. *Beszélőváltások a társalgásban*. http://fonetika.nytud.hu/letolt/ma_2.pdf (A letöltés ideje: 2011. október 1.)

Markó Alexandra – Dér Csilla Ilona 2011. Diskurzusjelölők használatának életkori sajátosságai. In Navracsics Judit – Lengyel Zsolt (szerk.): *Lexikai folyamatok egy- és kétnyelvű közegben.* Pszicholingvisztikai tanulmányok II. Tinta Kiadó, Budapest, 49–61.

Marzinzik, Mark – Kollmeier, Birger 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. In: *Proceedings of IEEE Transactions on Speech and Audio Processing* 10/6. 341–351.

Matsui, Tomoko – Furui, Sadaoki 1995. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. *Speech Communication* 17. 109–116.

McLachlan, Geoffrey J. – Krishnan, Thriyambakam 1997. *The EM Algorithm and its Extensions*. Wiley, New York.

Meignier, Sylvain – Bonastre, Jean-Francois – Igournet, Stephane 2001. E-HMM approach for learning and adapting sound models for speaker indexing. In: *Proceedings of Speaker Odyssey.* Chiana, Crete, 175–180.

Meinedo, Hugo – Neto, Joao. 2003. Audio segmentation, classification and clustering in a broadcast news task. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Hong Kong, China, 5–8.

Metze, Florian – Fugen, Christian – Pan, Yue – Schultz, Tanja – Yu, Hua 2004. The ISL RT-04S meetings transcription system. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Montreal, Canada, 464–474.

Mihajlik Péter 2010. *Spontán magyar nyelvű beszéd gépi felismerése nyelvspecifikus szabályok nélkül*. PhD-disszertáció. BME, Budapest.

Moattar, Mohammad H. – Homayounpour, Mohammad M. 2006. Speech overlap detection using spectral features and its application in speech indexing. In: *Information and Communication Technologies,* Damascus, Syria, 1270–1274.

Moh, Yvonne – Nguyen, Patrick – Junqua, Jean-Claude 2003. Towards domain independent speaker clustering. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Hong Kong, China, 85–88.

Mohamed, G. Hinton, – Penn, Gerald 2012. Understanding how deep belief networks perform acoustic modelling. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Kyoto, Japan, 4273–4276.

Moraru, Daniel – Ben, Mathieu – Gravier, Guillaume 2005. Experiments on speaker tracking and segmentation in radio broadcast news. In: *Proceedings of International Conference on Acoustics, Speech and Language Processing.* Lisbon, Portugal, 3049–3052.

Mori, Kazumasa – Nakagawa, Seiichi 2001. Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Salt Lake City, USA, 413–416.

Mowlaee, Pejman – Christensen, Mads G. – Tan, Zheng-Hua – Jensen, Soren H. 2010. A MAP criterion for detecting the number of speakers at frame level in model-based single-channel speech separation. In: *Signals, Systems, and Computers Website, 2010*. Asilomar, USA, 538–541.

Nakagawa, Seiichi – Suzuki, Hideyuki 1993. A new speech recognition method based on VQ-distortion and HMM. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Minneapolis, USA, 676–679.

Nemer, Elias – Goubran, Rafik – Mahmoud, Samy 2001. Robust voice activity detection using higherorder statistics in the LPC residual domain. In: *Proceedings of IEEE Trans. Speech Audio Processing*. 9/3. 217–231.

Németh Géza – Olaszy Gábor (szerk.) 2010. *A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek*. Akadémiai Kiadó, Budapest.

Németh Zsuzsanna 2007–2008. A forduló (beszédlépés) kiterjesztésének grammatikája a magyarban. *Nyelvtudomány* III–IV. 149–184.

Neuberger Tilda – Beke András 2013. Automatic laughter detection in spontaneous speech using GMM-SVM method. In: *Proceeding of Text, Speech and Dialogue 2013*. Plzeň, Czech Republic, 113–120.

Nguyen, Patrick 2003. SWAMP: An isometric frontend for speaker clustering. In: *Proceedings of NIST 2003 Rich Transcription Workshop*. USA, Boston.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.2420&rep=rep1&type=pdf
(A letöltés ideje: 2013. szeptember 1.)

Nikléczy Péter 2001. A műszeres személyazonosítás lehetőségei rövid időtartamú beszédminták alapján. *Beszédkutatás 2000*. 154–172.

Nikléczy Péter 2003. A zönge periódusidejének funkciója a hangszínezetben. *Beszédkutatás 2003*. 101–113.

Nikléczy Péter – Gósy Mária 2008. A személyazonosítás lehetősége a beszédanyag időtartamának függvényében. *Beszédkutatás 2008*. 172–181.

Nishida, Masafumi – Kawahara, Tatsuya 2003. Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Hong Kong, China, 172–175.

Otterson, Scott – Ostendorf, Mari 2007. Efficient use of overlap information in speaker diarization. In: *Proceeding of ASRU*. Kyoto, Japan, 683–686.

Parthasarathi, Sree H. K. – Magimai-Doss, Mathew – Gatica-Perez, Daniel – Bourlard, Hervé 2009. Speaker change detection with privacy-preserving audio cues. In: *Proceedings of International Conference on Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction*. Cambridge, USA, 343–346.

Perez-Freire, Luis – Garcia-Mateo, Carmen 2004. A multimedia approach for audio segmentation in TV broadcast news. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Montreal, Canada, 369–372.

Peskin, Barbara – Navratil, Jiri – Abramson, Joy – Jones, Douglas – Klusacek, David – Reynolds, Douglas A. 2003. Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02. In: *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing*. 729–795.

Placencia, Maria E. 1997. Opening up closings. The Ecuadorian way. *Text. An interdisciplinary journal for the study of discourse* 17/1. 53–81.

Pléh Csaba 1998. *Mondatmegértés a magyar nyelvben*. Osiris Kiadó, Budapest.

Ptacek, Paul H. – Sander, Eric K. 1966. Age recognition from voice. *Journal of Speech and Hearing Research* 9/2. 273–277.

Ramírez, Javier – Górriz, Juan M. – Segura, José C. 2007. Statistical voice activity detection based on integrated bispectrum likelihood ratio tests. *Journal of the Acoustical Society of America* 121. 2946–2958.

Ramírez, Javier – Górriz, Juan M. – Segura, José C. – Puntonet, Carlos G. – Rubio, Antonio J. 2006. Speech/non-speech discrimination based on contextual information integrated bispectrum LRT. In: *Proceeding of IEEE Signal Processing Letters* 13/8. 497–500.

Ramírez, Javier – Segura, José C. – Benítez, Carmen – de la Torre, Ángel – Rubio, Antonio J. 2004. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication* 42/3–4. 271–287.

Ramírez, Javier – Segura, José C. – Benítez, Carmen – de la Torre, Ángel – Rubio, Antonio J. 2005. An effective OSF-based VAD with noise suppression for robust speech recognition. In: *Proceeding of IEEE Transactions on Speech and Audio Processing*. 13/6. 1119–1129.

Reynolds, Douglas A. 1995. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* 17. 91–108.

Reynolds, Douglas A. 1996. MIT Lincoln Laboratory site presentation. In: *Speaker Recognition Workshop.* ftp://jaguar.ncsl.nist.gov/speaker/ (A letöltés ideje: 2013. szeptember 1.)

Reynolds, Douglas A. 1997. Comparison of background normalization methods for text-independent speaker verification. In: *Proceedings of European Conference on Speech Communication and Technology 1997.* 963–966.

Reynolds, Douglas A. 2009. Universal Background Models. In Li, Stan Z.– Jain, Anil (eds.): *Encyclopedia of biometrics.* Springer, Journal Article, February *2008.* 1349–1352.

Reynolds, Douglas A. – Rose, Richard C. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. In: *Proceedings of IEEE Transactions on Speech and Audio Processing* 3/1. 72–83.

Reynolds, Dougles A. – Singer, Elliot – Carlson, Beth A. – O'Leary, Gerald C. – McLaughlin, Jack J. – Zissman, Marc A. 1998. Blind clustering of speech utterances based on speaker and language characteristics. In: *Proceedings IEEE International Conference on Speech and Language Processing*. Sidney, Australia, 3193–3196.

Reynolds, Douglas A. – Quatieri, Thomas F. – Dunn, Robert 2000. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10/1–3. 19–41.

Reynolds, Douglas – Andrews, Walter – Campbell, Joshep – Navratil, Jiri – Peskin, Barbara – Adami, André – Jin, Qin – Klusacek, David – Abramson, Joy – Mihaescu, Radu – Godfrey, Jack – Jones, Doug – Bing, Xiang 2003. The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. In: *Proceedings of International Conference on Acoustics, Speech, Signal Processing.* 784–787.

Reynolds, Douglas A. – Torres-Carrasquillo, Pedro A. 2004. The MIT Lincoln Laboratories RT-04F diarization systems: Applications to broadcast audio and telephone conversations. In: *Proceedings of Fall 2004 Rich Transcription Workshop*. Palisades, NY, USA, 1065–1103.

Roch, Marie – Cheng, Yanliang 2004. Speaker segmentation using the MAP-adapted Bayesian information criterion. In: *Proceedings of Speaker Odyssey Workshop.* Toledo, Spain, 349–354.

Rosenberg, Aaron E. – DeLong, Joel – Lee, Chin-Hui – Juang, Biing-Hwang – Soong, Frank K. 1992. The use of cohort normalized scores for speaker verification. In: *Proceedings of International Conference on Spoken Language Processing.* 599–602.

Rougui, Jamal – Rziza, Mohammed – Aboutajdine, Driss – Gelgon, Marc – Martinez, Jean 2006. Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Toulouse, France, 521–524.

Sacks, Harvey 1992. *Lectures on conversation*. Blackwell, Oxford.

Sacks, Harvey – Schegloff, Emanuel A. – Jefferson, Gail 1974. A simplest systematics for the organization of turntaking for conversation. *Language* 50. 696–735.

Saeidi, Rahim – Mowlaee, Pejman – Kinnunen, Tomi – Tan, Zheng-Hua – Christensen, Mads G. – Jensen, Soren H. – Franti, Pasi 2010. Improving monaural speaker identification by double-talk detection. In: *Proceedings of International Conference on Speech and Language Processing 2010.* Makuhari, Chiba, Japan, 1069–1072.

Sahidullah, Md. – Saha, Goutam 2012. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Communication* 54/4. 543–565.

Salembier, Phillipe – Sikora, Thomas – Manjunath, B. S. (eds.) 2002. *Introduction to MPEG-7: Multimedia content description interface*. Wiley & Sons, New York.

Sankar, Ananth – Beaufays, Francoise – Digalakis, Vassilios 1995. Training data clustering for improved speech recognition. In: *Proceedings of European Conference on Speech Communication and Technology 1995*. Madrid, Spain, 503–506.

Sankar, Ananth – Weng, Fuliang – Stolcke, Andreas – Grande, Ramana R. 1998. Development of SRI's 1997 broadcast news transcription system. In: *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*. Landsdowne, USA, 91–96.

Saunders, John 1996. Real-time discrimination of broadcast speech/music. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.* Atlanta, Georgia, USA, 993–996.

Schegloff, Emanuel A. 1982. Discourse as an interactional achievement: Some uses of *uh huh* and other things that come between sentences. In Tannen, Deborah (ed.): *Analyzing discourse: Text and talk.* Georgetown University Press, Washington D. C., 71–93.

Schegloff, Emanuel A. 1992. Introduction. In Sacks, Harvey (ed.): *Lectures on conversation*. Vol. 1. Blackwell, Oxford, 9–12.

Scherer, Klaus R. – Banse, Rainer – Wallbott, Harald 2001. Emotional inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology* 32/1. 76–92.

Schiffrin, Deborah 1987. *Discourse markers*. Cambridge University Press, Cambridge.

Schwartz, Richard – Roucos, Salim – Berouti, Michael 1982. The application of probability density estimation to text independent speaker identification. In: *Proceedings of International Conference Acoustics, Speech, and Signal Processing,* Paris, France, 1649–1652.

Schwarz, Gideon 1971. A sequential student test. *The Annals of Mathematical Statistics* Volume 42, Number 3 (1971), 42/3.1003–1009.

Schwarz, Gideon 1978. Estimating the dimension of a model. *The Annals of Statistics 6*. 461–464.

Selting, Margret 2000. TCUs and TRPs: The construction of units in conversational talk. *Language in Society* 29. 477–517.

Shriberg, Elizabeth – Stolcke, Andreas – Baron, Don 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In: *Proceedings of European Conference on Speech Communication and Technology 2001*. Aalborg, Denmark, 1359–1362.

Shriberg, Elizabeth – Ferrer, Luciana – Kajarekar, Sachin S. – Venkataraman, Anand – Stolcke, Andreas 2005. Modeling prosodic feature sequences for speaker recognition. *Speech Communication* 46/3–4. 455–472.

Siegler, Matthew A. – Jain, Uday – Raj, Bhiksha – Stern, Richard M. 1997. Automatic segmentation, classification and clustering of broadcast news audio. In: *Proceedings of DARPA Speech Recognition Workshop*. 97–99.

Sinha, Rohit – Tranter, Sue E. – Gales, Mark J. F. – Woodland, Phil C. 2005. The Cambridge University March 2005 speaker diarisation system. In: *Proceedings of European Conference on Speech Communication and Technology (Interspeech)*. Lisbon, Portugal, 2437–2440.

Siu, Man-Hung – Yu, George – Gish, Herbert 1992. An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. San Francisco, USA, 189–192.

Sivakumaran, Perasiriyan – Fortuna, J. – Ariyaeeinia, Aladdin M. 2001. On the use of the Bayesian information criterion in multiple speaker detection. In: *Proceedings of European Conference on Speech Communication and Technology 2001*. Aalborg, Denmark, 795–798.

Sohn, Jongseo – Kim, Nam Soo – Sung, Wonyong 1999. A statistical model-based voice activity detection. *IEEE Signal Processing Letters* 6/1. 1–3.

Solomonov, Alex – Mielke, Angela – Schmidt, Michael – Gish, Herbert 1998. Clustering speakers by their voices. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Seattle, USA, 757–760.

Soong, Frank K.– Rosenberg, Aaron E. – Rabiner, Lawrence – Juang, Biing-Hwang 1985. A vector quantization approach to speaker recognition. In: *Proceedings of International Conference Acoustics, Speech, and Signal Processing*. Tampa, Florida, 387–390.

Sönmez, Kemal – Shriberg, Elizabeth – Heck, Larry – Weintraub, Mitchel 1998. Modeling dynamic prosodic variation for speaker verification. In: *Proceedings of ICSLP.* Sydney, Australia, 3189–3192.

Stephens, Jane – Beattie, Geoffrey W. 1986. On judging the ends of speaker turns in conversation. *Journal of Language and Social Psychology* 5/2. 119–134.

Stokoe, Elizabeth 2006. On ethnomethodology, feminism, and the analysis of categorial reference to gender in talk-in-interaction. *Sociological Review* 54. 467–94.

Suykens, Johan A. K. – Gestel, Tony Van – de Brabanter, Jos – De Moor, Bart – Vandewalle, Joos 2002. *Least Squares Support Vector Machines*. World Scientific, Singapore.

Tanaka, Hiroko 2001. Adverbials for turn projection in Japanese: Toward a demystification of the "telepathic" mode of communication. *Language in Society* 30/4. 559–587.

Tang, Huixuan 2008. *A comparative evaluation of deep belief nets in semi-supervised learning.* Report for CSC2515. http://www.cs.toronto.edu/~hxtang/projects/dbn_eval/dbn_eval.pdf (A letöltés ideje: 2013. szeptember 1.)

Tang, Yichuan 2013. Deep learning using linear support vector machines. In: *International Conference on Machine Learning.* http://deeplearning.net/wp-content/uploads/2013/03/dlsvm.pdf (A letöltés ideje: 2014. január 12.)

Theodoridis, Sergios – Koutroumbas, Konstantinos 2008. *Pattern recognition*. Third Edition. Academic Press, Orlando, Florida, USA.

Tishby, Naftali Z. 1991. On the application of mixture AR hidden Markov models to text independent speaker recognition. In: *Proceedings of the IEEE Transactions on Acoustics, Speech, Signal Processing* 39/3. Toronto, Canada, 563–570.

Tranter, Sue – Reynolds, Douglas A. 2004. Speaker diarization for broadcast news. In: *Proceeding of Odyssey Speaker and Language Recognition Workshop*. Toledo, Spain, 337–344.

Tritschler, Alain – Gopinath, Ramesh A. 1999. Improved speaker segmentation and segments clustering using the Bayesian information criterion. In: *Proceedings of European Conference on Speech Communication and Technology 1999*. 679–682.

Trueba-Hornero, Beatriz 2008. *Handling overlapped speech in speaker diarization*. Master's thesis. Universitat Politecnica de Catalunya, Barcelona.

Tucker, Roger 1992. Voice activity detection using a periodicity measure. In: *IEEE Proceedings, Communications, Speech and Vision* 139. 1350–2425.

Vandecatseye, An – Martens, Jean-Pierre 2003. A fast, accurate and stream-based speaker segmentation and clustering algorithm. In: *Proceedings of European Conference on Speech Communication and Technology 2003*. Geneva, Switzerland, 941–944.

Vandecatseye, An – Martens, Jean-Pierre – Neto, Joao – Meinedo, Hugo – Garcia-Mateo, Carmen – Dieguez, Javier – Mihelic, France – Zibert, Janez – Nouza, Jan – David, Petr – Pleva, Matus – Cizmar, Anton – Papageorgiou, Harris – Alexandris, Christina 2004. *The COST278 pan-European Broadcast News Database*. LREC'04, Lisbon, Portugal.

Vescovi, Michele – Cettolo, Mauro – Rizzi, Romeo 2003. DP algorithm for speaker change detection. In: *Proceedings of European Conference on Speech Communication and Technology 2003*. 2997–3000.

Vipperla, Ravichander – Geiger, Juergen T. – Bozonnet, Simon – Wang, Dong – Evans, Nicholas – Schuller, Bjorn – Rigoll, Gerhard 2012. Speech overlap detection and attribution using convolutive non-negative sparse coding. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 4181–4184.

Wactlar, Howard D. – Hauptmann, Alexander G. – Witbrock, Michael J. 1996. News on-demand experiments in speech recognition. In: *ARPA STL Workshop*. https://www.ri.cmu.edu/pub_files/pub2/wactlar_howard_1996_2/wactlar_howard_1996_2.pdf (A letöltés ideje: 2013. szeptember 1.)

Wegmann, Steven – Scattone, Francesco – Carp, Ira – Gillick, Larry – Roth, Robert – Yamron, Jonathan P. 1998. Dragon system's 1997 broadcast news transcription system. In: *DARPA Broadcast News Transcription and Understanding Workshop*. Landsdowne, USA, 89–108.

Wei, Koh Chin Eugene 2008. *Speaker diarization of news broadcasts and meeting recordings*. Master Thesis. Nanyang Technological University, Singapore.

Wilcox, Lynn Donelle – Chen, Francine R. – Kimber, Don – Balasubramanian, Vijay 1994. Segmentation of speech using speaker identification. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Adelaide, Australia, 161–164.

Willsky, Alan S. – Jones, Harold L. 1976. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. In: *Proceedings of IEEE Transactions on Automatic Control* AC-21/1. 108–112.

Woo, Kyoung-Ho – Yang, Tae-Young – Park, Kun-Jung – Lee, Chungyong 2000. Robust voice activity detection algorithm for estimating noise spectrum. *Electronics Letters* 36/2. 180–181.

Woodland, Phil C. – Gales, Mark J. F. – Pye, David – Young, Stephen J. 1997. The development of the 1996 HTK broadcast news transcription system. In: *Speech Recorgnition Workshop*. 73–78.

Wooters, Chuck – Fung, James – Peskin, Barbara – Anguera, Xavier 2004. Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system. In: *Fall 2004 Rich Transcription Workshop*. Palisades, New York. ftp://ftp.icsi.berkeley.edu/pub/speech/papers/EARS-RT04f-spkr.pdf (A letöltés ideje: 2013. szeptember 1.)

Wooters, Chuck – Huijberts, Marijn 2007. The ICSI RT07s speaker diarization system. In: *Proceedings of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop*. Baltimore, Maryland, 509–519.

Wu, Tingyao – Lu, Lie – Chen, Ke – Zhang, Hong-Jiang 2003a. UBM-based incremental speaker adaptation. In: *Proceedings of IEEE International Conference on Multimedia & Expo*. 721–724.

Wu, Tingyao – Lu, Lie – Chen, Ke – Zhang, Hong-Jiang 2003b. UBM-based real-time speaker segmentation for broadcasting news. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 193–196.

Wu, Tingyao – Lu, Lie – Chen, Ke – Zhang, Hong-Jiang 2003c. Universal background models for real-time speaker change detection. In: *International Conference on Multimedia Modeling*. 135–149.

Xiao, Bo – Ghosh, Prasanta K. – Georgiou, Panayiotis – Narayanan, Shrikanth S. 2011. Overlapped speech detection using long-term spectro-temporal similarity in stereo recording. In: *Proceeding of International Conference on Acoustics, Speech and Signal Processing.* 5216–5219.

Yamaguchi, Masahide – Yamashita, Masaru – Matsunaga, Shoichi 2005. Spectral cross-correlation features for audio indexing of broadcast news and meetings. In: *Proceedings of International Conference on Speech and Language Processing*. Lisbon, Portugal, 613–616

Yamamoto, Kiyoshi – Asano, Futoshi – Yamada, Takeshi – Kitawaki, Nobuhiko 2006. Detection of overlapping speech in meetings using Support Vector Machines and Support Vector Regression. In: *Journal IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences Archive.* E89-A/8. 37–40.

Yella, Sree H. – Valente, Fabio 2012. Speaker diarization of overlapping speech based on silence distribution in meetings recordings. In: *Proceedings of International Conference on Speech and Language Processing 2012.* Portland, USA, 490–493.

Yella, Sree Harsha – Bourlard, Hervé 2013. Improved overlap speech diarization of meeting recordings using long-term conversational features. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 7746–7750.

Ying, Dongwen – Yan, Yonghong – Dang, Jianwu – Soong, Frank K. 2011. Voice activity detection based on an unsupervised learning framework. In: *IEEE Transactions on Audio, Speech and Language Processing* 19/8. 2624–2633.

Yngve, Victor H. 1970. On getting a word in edgewise. In: *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*. Chicago Linguistic Society, Chicago, 567–577.

Zelenak, Martin – Hernando, Javier 2011. The detection of overlapping speech with prosodic features for speaker diarization. In: *Proceedings of International Conference on Speech and Language Processing 2011.* 32–35.

Zelenak, Martin – Segura, Carlos – Hernando, Javier 2010. Overlap detection for speaker diarization by fusing spectral and spatial features. In: *Proceedings of International Conference on Speech and Language Processing 2010*. Makuhari, Japan, 2302–2305.

Zhou, Bowen – Hansen, John H. 2000. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In: *Proceedings of International Conference on Speech and Language Processing.* Beijing, China, 714–717.

Zhu, Xuan – Barras, Claude – Lamel, Lori – Gauvain, Jean-Luc 2006. Speaker diarization: From broadcast news to lectures. In Renals, Steve – Bengio, Samy – Fiscus, Johnatan G. (eds.): *Machine learning for multimodal interaction*. Lecture Notes in Computer Science. Spinger-Verlag, Berlin – Heidelberg – New York, 396–406.

Zhu, Xuan – Barras, Claude – Meignier, Sylvain – Gauvain, Jean-Luc 2005. Combining speaker identification and BIC for speaker diarization. In: *Proceedings of International Conference on Speech and Language Processing 2005.* Lisbon, Portugal, 2441–2444.

Zochova, Polina – Radova, Vlasta 2005. Modified DISTBIC algorithm for speaker change detection. In: *Proceedings of International Conference on Speech and Language Processing*. Lisbon, Portugal, 3073–3076.

Automatic speaker diarization in Hungarian spontaneous conversations

In human-machine communication, several processes have been modelled by applying speech technology, such as speech decoding (automatic speech recognition), speech production (speech synthesis) or speaker identification based on voice (speaker recognition). These processes are linked together in conversation where the operation of decoding and production are circularly interleaved. This circulation is further promoted by speaker changes. The automatic detection of the speaker change is therefore very important. Speaker diarization is the process of partitioning an input audio stream into homogeneous segments according to the speaker's identity. It can enhance the readability of an automatic speech transcription by structuring the audio stream into speaker turns and, when used together with speaker recognition systems, by providing the speaker's true identity. It is used to answer the question "who spoke when?".

In the literature, extensive research effort is concentrated on speaker diarization, but principally for English. However, for the Hungarian language, no work is known which addresses the field of speaker diarization. The aim of this research is to develop a speaker diarization system for the Hungarian language. The main focus is to create algorithms for speaker diarization (speaker segmentation, speaker clustering, overlapping speech detection) and to implement and enhance some already existing algorithms in speaker diarization (voice activity detection, speaker recognition), focusing on Hungarian conversation. The adopted approach is mainly based on unsupervised methods. The main motivation of this thesis is to develop a speaker diarization for spontaneous conversations, because the most of speaker diarization systems were created for broadcast shows or telephone call speech material. Broadcast shows typically contain read or prepared speech, characterized by minimally overlapping speech segments. Telephone calls mainly contain conversations between only two persons. Spontaneous conversation is the most challenging task for speaker diarization, as it presents many overlapping speech segments and very short speaker turns.

This work contains 10 chapters. The first chapter provides an introduction, where the theoretical and practical background of the introduced scientific areas is presented focusing on speaker diarization. *Chapter 2* reviews the state-of-art in speaker diarization (speaker segmentation and cluster methods). *Chapter 3* describes the goal of the research, research questions and hypotheses. The material, subject of the research and the evaluation of the speaker diarization approach (DER: Detection Error Rate) is presented in *Chapter 4.* Our proposed algorithm for a speaker diarization system presented in *Chapter 5.* In this Chapter, the development of our two-pass speaker diarization system is described. The pre-processing step relies on applying VAD (voice activity detection) and proposes several modifications in the

original diarization algorithm, which are presented in *Chapter 5. Chapter 5* describes the examination of speaker specific acoustic features for speaker recognition. The algorithm for overlapping speech detection is described in *Chapter 5. Chapter 6* concludes the results of this thesis with conclusions and by summarizing its main contributions *(Chapter 7)*. The summary of this work is presented in *Chapter 8. Chapter 9* provides an outlook for possible directions of future work in the field. Chapter 10 shows the references.

For this research, 100 spontaneous conversations (total duration is of 55 hours) were selected from the BEA database (Gósy 2012), recorded in a laboratory environment. In each case, three persons were involved in the conversations. Two of them were permanent (2 females, average age 32 years old). The third subject (interviewere) was one out of the 43 male and 67 female (average age 35 years old) speakers. To test our speaker diarization system, 12 conversations were selected randomly from the BEA database. The total duration of conversations was 2.8 hours which contained 490 speaker-changes. We implemented standard BIC-base speaker segmentation to compare it to the proposed system. The standard BIC-base speaker segmentation used standard MFCC and the λ parameter value was 0. In this standard system we did not use any speech detection or overlap detection algorithm. By using standard BIC-base speaker segmentation the best DER value was 39.43%. To improve this result we used MFCC for the spectral subband between 2.5 and 3.5 kHz and energy along with deltas in standard BIC-base speaker segmentation. The result showed that when the BIC-base segmentation included MFCC (2.5–3.5) feature, the proposed method achieved about 0.869% relative DER reduction (from 39.43.5% to 38.56.2%) which is statistically significant improvement. Performance of the BIC-base speaker segmentation is depending on the penalty factor λ. We tested our speaker segmentation system using various values for λ (from 0 to 4). The best DER is obtained if the penalty factor is $\lambda = 1$ (by DER = 35.73%). By adding the speech/nonspeech detector proposed for spontaneous speech, not only it does not improve the non-speech errors, but also reduces the speaker error, due to a reduction in clustering errors as noted above. The result showed that when the BIC-base segmentation included VAD, the proposed method achieved about 4.535% relative DER reduction (from 35.73% to 31.21%).

We present our initial work toward developing an overlap detection system based on a deep neural network for improved speaker diarization. We demonstrated a relative improvement of about 2.49% in DER over the baseline diarization system (from 31.21% to 28.713%).

This research addressed the topic of speaker diarization for spontaneous conversations. The presented BIC-base system uses as baseline the technology in speaker diarization for broadcast news and adapts it to the spontaneous speech by developing new algorithms and improving existent ones to use speaker specific features and to implement VAD and overlapping speech detection algorithm based on a deep neural network. In the field of discourse modelling, speaker diarization could benefit from research aiming at modelling the turn-taking between the speakers. Using information at a higher level than simple acoustics, the transition probabilities between speakers could be appropriately set to help the decoding.

A
BESZÉD • KUTATÁS • ALKALMAZÁS
című sorozat eddig megjelent kötetei:

Markó Alexandra:
Az irreguláris zönge funkciói a magyar beszédben
ISBN 978-963-312-195-5

Bóna Judit:
A spontán beszéd sajátosságai az időskorban
ISBN 978-963-312-199-3

Horváth Viktória:
Hezitációs jelenségek a magyar beszédben
ISBN 978-963-312-205-1

Neuberger Tilda:
A spontán beszéd sajátosságai gyermekkorban
ISBN 978-963-312-204-4

Beke András:
Gépi beszélődetektálás magyar nyelvű spontán társalgásokban
ISBN 978-963-312-234-1

Az emberiség régi vágya, hogy saját szóbeli kommunikációját reprodukálni tudja gépek által. A társalgás egyik legszembetűnőbb jelensége, hogy körkörös folyamattal megy végbe, ahogy a résztvevők egymást váltva mondják ki gondolataikat. Az ezt a váltakozást gépi úton modellező beszédtechnológiai eljárás a beszélődetektálás, amelynek alapvető feladata az, hogy automatikusan jelölje a folytonos beszédjelben, hogy *mikor ki beszél*.

A kutatás fő motivációja az volt, hogy spontán társalgásokra valósítsunk meg beszélődetektálót, mivel az eddigi beszélődetektálók híradós adásokra vagy telefonhívásokra készültek. A legnagyobb kihívást azonban a több résztvevős spontán társalgások beszélőkre való bontása jelenti. A beszélődetektálásnak elengedhetetlen szerepe lehet a napjainkban egyre növekvő adatmennyiség automatikus feldolgozásában, amelyeknek nagy része beszélők szerint strukturálható.

A kötet érdeklődésre tarthat számot a beszédtechnológusok, a fonetikai és más nyelvészeti tudományterületek kutatói körében.

ISBN 978-963-312-234-1