Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar

DOKTORI DISSZERTÁCIÓ

BEKE ANDRÁS

BESZÉLŐDETEKTÁLÁS MAGYAR NYELVŰ SPONTÁN TÁRSALGÁSOKBAN

Nyelvtudományi Doktori Iskola
vezető: Prof. Dr. Bárdosi Vilmos egyetemi tanár

Alkalmazott Nyelvészet Doktori Program
vezető: Prof. Dr. Gósy Mária egyetemi tanár

A bizottság tagjai

A bizottság elnöke:	Prof. Dr. Nyomárkay István akadémikus
Hivatalosan felkért bírálók:	Prof. Dr. Adamikné Jászó Anna DSc
	Prof. Dr. Olaszy Gábor DSc
A bizottság titkára:	Dr. Dér Csilla Ilona PhD
A bizottság további tagjai:	Prof. Dr. Vicsi Klára DSc
	Prof. Balázs Géza CSc
	Dr. Hámori Ágnes PhD

Témavezető
Prof. Dr. Gósy Mária DSc

Budapest, 2014

Tartalom

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 vizsgáló 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 alapvetően nem rendezetlen struktúra, hanem szabályok mentén rendeződik, dinamikusan alakul a beszédpartnerek mentén. 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 beszélődetektálás (speaker diarization). A beszélődetektálás alapvető feladata, hogy a társalgásokban automatikusan jelölje, hogy mikor ki beszél. Ennek során a folyamatos társalgások automatikusan lejegyzett szövegeit újrastrukturáljuk, 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 dolgozat célkitűzése egyrészt az, hogy a beszélődetektáláshoz kapcsolódó tudományterületeket bemutassa, illetve hogy maga a beszélődetektálás főbb módszertani ismereteit leírja. Mivel magyar nyelven nem születtek tanulmányok e témában, ezért ez egy jelentős lépés, hogy a hazai kutatások kiinduló munkája lehessen ez a dolgozat. A másik célja az volt, hogy a beszélődetektálóhoz szükséges algoritmusokat elkészítsük (egyszerrebeszélés-detektálás, beszélőszegmentáló, beszélőklaszterező), és a már létező algoritmusokat implementáljuk a beszélődetektálóba (beszéd/nembeszéd-detektáló, beszélőfelismerő algoritmus).

Az általunk javasolt rendszer célja, hogy magyar nyelvű spontán társalgásokban automatikusan detektálja a beszélőket pusztán akusztikai információk alapján, vagyis megoldást adjon arra a kérdésre, hogy „Mikor ki beszél?”. Az algoritmus kialakításához a BEA (Magyar spontánbeszéd-adatbázis; Gósy 2012) spontán társalgásait használtuk fel, amelyben három résztvevő társalog. Az általunk javasolt beszélődetektáló rendszer alapvetően nemellenőrzött tanulási eljárásokon alapul.

A disszertáció 11 fejezetből áll. Az első, legterjedelmesebb fejezetben bemutatjuk a témához kapcsolódó tudományágak elméleti és gyakorlati hátterét. A beszédprodukció és beszédpercepció ismertetésén keresztül bemutatjuk a spontán beszéd jellegzetességeit, majd annak a legtöbbet használt változatát, a társalgást. Ebben a fejezetben prezentáljuk a beszélőalkalmazkodás részletezésével a dinamikusan változó társalgást. A társalgás során a beszélő változtatja a beszédét, mimikáját, gesztusait a beszédpartner függvényében (Turner–West 2010). A beszélőalkalmazkodás kutatói azokat az okokat kívánják feltárni, amiért az egyének minimalizálják vagy hangsúlyozzák a maguk és a tárgyalópartnereik közti társadalmi különbségeket a szóbeli és a nonverbális kommunikáción keresztül. Szintén ebben a fejezetben kap helyet a társalgás építőköveinek bemutatása, amely a beszédforduló (angol terminussal: turn). A beszédforduló dinamikusan valósul meg a társalgás során, semmiképpen nem előre definiált helyen, vagyis rekurzív módon, lépésről lépésre különbözőképpen, mindig a lehetséges váltási ponton vagy váltásra alkalmas helyen (átmeneti relevanciahely). A beszédfordulók szerkezetét alapvetően meghatározza az a potenciális hely, ahol a társalgás résztvevői átvehetik a szót; 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 beszédlépés végét a megnyilatkozó az esetek többségében jelzi. A beszédlépés jelzésére számos akusztikai, szintaktikai, pragmatikai stb. eszköz áll a beszélő rendelkezésére. A *Bevezetés* fejezetben mutatjuk be az egyik lehetséges jelzési eszközt, amely a diskurzusjelölő. A diskurzusjelölő egy pragmatikai eszköz, amelynek számos funkciója lehet, többek között jelezhetik a beszédlépések elejét vagy végét. Az 1. fejezetben taglaljuk az egyszerre beszélések szerepét a társalgásokban, illetve, hogy hogyan dolgozhatók fel a beszéddetektálás során.

A disszertáció 2. fejezete módszertani áttekintést ad a beszélődetektálásban használt algoritmusokról. Itt kerül bemutatásra a beszéddetektálás folyamata, amelynek célja, hogy a folyamatos akusztikai jelben jelölje, hogy hol van beszédrész, illetve nem beszédrész. A 2. fejezet ismerteti a beszélőfelismerés alapvető módszertani kérdéseit, amelynek célja, hogy milyen módon lehet gépileg felismerni a beszélő személyt a hangja alapján. Ebben a fejezetben kap helyet az egyszerre beszélések automatikus osztályozása is, amelynek igen nagy szerepe van a beszélődetektálás téves riasztásaink csökkentésében. A *Módszertani fejezetben* (2. fejezet) ismertetjük magát a beszélődetektáló rendszerek elméletét és módszertanát.

A saját kutatásunk céljainak, kérdéseinek és hipotézisének ismertetése a 3. fejezetben történik.

A 4. fejezetben a kísérleti személyek, általános anyag és módszer ismertetése történik. Itt mutatjuk be a kísérletekhez használt adatbázis felépítését, tartalmát, illetve itt kerül bemutatásra a beszélődetektálás kiértékeléséhez használt NIST (National Institute of Standards and Technology, Nemzetközi, Szabványok és Technológiák Nemzetközi Intézete) által javasolt DER (Detection Error Rate) eljárás, és az osztályozás kiértékeléséhez használt DET (Detection Error Tradeoff) algoritmus.

Az 5. fejezetben a kísérletek és az eredmények ismertetésére kerül sor. Először a beszéddetektálás folyamat lépéseit és eredményeit írjuk le. Ezután a beszélőspecifikus akusztikai jellemzők vizsgálatát és az azzal elért eredményeket prezentáljuk. A következő alfejezet a beszélődetektálásban legtöbb hibát okozó egyszerre beszéléseket detektáló rendszert mutatjuk be. Az 5. fejezet utolsó fejezetében pedig az általunk fejlesztett beszélődetektáló rendszert és az azzal elért eredményeket prezentáljuk.

A 6. fejezet az általános következtetéseket tartalmazza, amelyet az általános összefoglalás követ (7. fejezet). Ezután ismertetjük a beszélődetektálás felhasználási és további fejlesztési lehetőségeit (8. fejezet). Végül a disszertáció téziseit fogalmazzuk meg (9. fejezet). Ezt követi az *Irodalom* (10), és a *Rövidítések jegyzéke* fejezet (11. 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). Napjaink egyik beszédtechnológiai

célkitűzése, hogy a beszédfelismerő (SST: speech-to-text) rendszereket egyesítsék metaadatokat kinyerő rendszerekkel (a metaadat itt minden olyan információt jelent, amely nem a hang-szöveg átalakításkor jön létre, például beszélőváltás, tartalomkinyerés stb). A kutatás és fejlesztés célja az, hogy a kapott információval segítsék, modellezzék az ember–ember és az ember–gép kommunikációt.

A disszertáció eredményei közelebb vihetik a kutatót az ember–ember kommunikáció megértéséhez, modellezéséhez, amely tovább mutat a mesterséges intelligencia, az ember–gép kommunikációja felé.

Köszönetnyilvánítás

Ezúton szeretném kifejezni hálás köszönetemet témavezetőmnek, dr. Gósy Máriának a sok év során nyújtott segítségéért, valamint opponenseimnek a hasznos megjegyzéseikért. Köszönettel tartozom dr. Szaszák Györgynek a szakmai és baráti támogatásáért.

Beke András

2014.01.20.

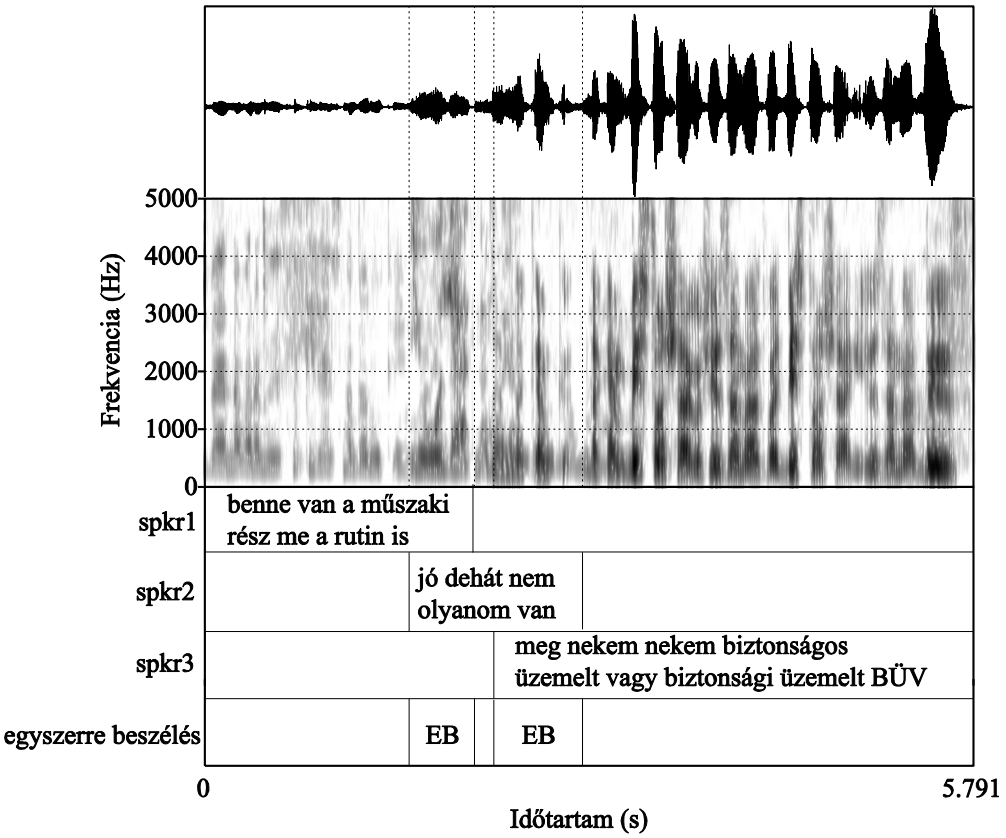
1. BEVEZETÉS

A kommunikáció alapvető feltétele a résztvevők, azaz a feladó (forrás) és a címzett (vevő). A feladó az, aki különböző nyelvi és nem nyelvi jelek segítségével üzenetet küld a címzettnek (kódolja), aki ezt az üzenetet felfogja, értelmezi (dekódolja) és válaszol rá (Denes–Pinson 1993). 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 nem nyelvi kódokat. A kommunikáció csak akkor lesz sikeres, ha a résztvevők közös nyelvet beszélnek, azaz egyformán ismerik a kódot. A megfogalmazott üzenet a csatornán keresztül jut el a feladatótól a címzettig, az továbbítja a közleményt. A csatorna lehet hallható (telefonbeszélgetés), látható (levél), érezhető (tapintás) vagy egyszerre többféle is (személyes beszélgetés). A tipikusnak mondható verbális kommunikációt mindig nonverbális elemek (Knapp–Hall 2001) kísérik (metakommunikációval), amelyek természetesen csak akkor érvényesülnek, ha a kapcsolat nem csak auditív, hanem vizuális formában is fennáll (vagyis nem csak hallják, hanem látják is egymást a felek). Ilyen a testtartás, hangsúly, mimika, gesztikulálás stb. (Goodwin 1979). A beszédkommunikációban zajnak nevezzük azokat a tényezőket, amelyek megzavarják, torzítják az üzenetet, gátolják annak eljutását a címzetthez (pl. ha recseg a telefon).

A társalgás az 1960-as években került a vizsgálatok középpontjába, elsősorban a szociológiai érdekeltségű társas nyelvészet (Streeck 1987; Kiss 1995), a szociálpszichológia, a pszicholingvisztika, a modern filozófia és a logika együttműködéseként (Pléh 2012). A társalgásokkal alapvetően a diskurzuselemzés, illetve a konverzációelemzés foglalkozik (pl. Austin 1962/1990; Sacks et al. 1974; Grice 1975; Franke 1990; Clark 1996; Teun−van Dijk 2006; Őványi 2001; Jakusné Harnos 2002; Hámori 2006; Boronkai 2006, 2008).

A konverzációanalízis (conversation analysis) néven megjelent új tudományág a hétköznapi társalgások verbális interakciók szerkezetét vizsgálja, amelynek bizonyos szerkezeti szabályosságokat feltételeznek a társalgások felépítésében (Garfinkel 1967; Goffman 1983; Schegloff 1992; Sacks et al. 1974; Sacks 1992; Őványi 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).

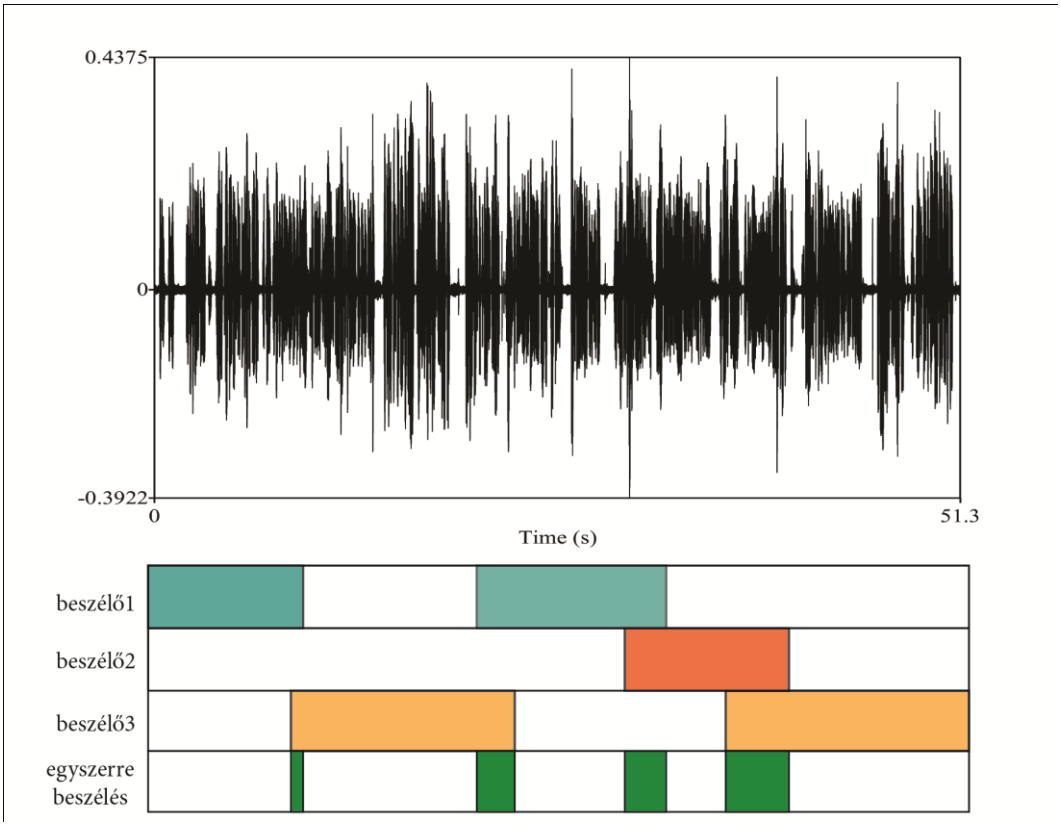


1.1. ábra

A társalgás felépítésének szemléltetése

A konverzációelemzés adta keretben tehát a társalgásban struktúrák jelennek meg, amely nem csak nyelvészeti szempontból fontos, hanem beszédtechnológiai szempontból is, hiszen ha a társalgás rendszerszerű, akkor feltételezhetően gépi úton modellezhető. A beszédtechnológia mesterséges intelligencián belül a beszéd alapú (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 beszédelőállítást (beszédszintézis), a beszélő személy

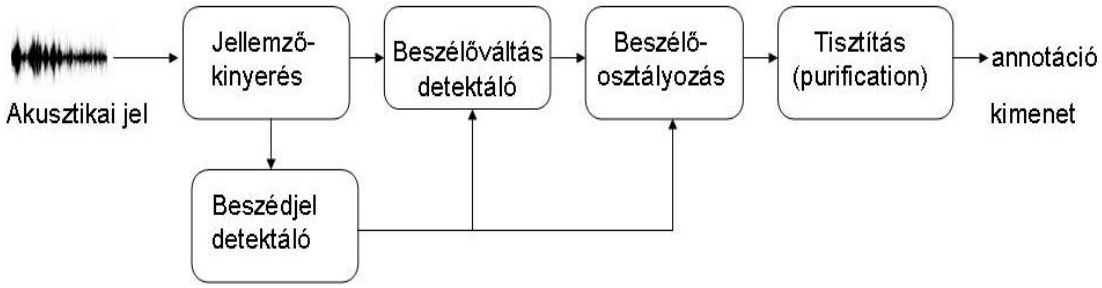
gépi azonosítását a hangja alapján (beszélőfelismerés). Ezek a részfolyamatok a társalgásban kapcsolódnak össze, ahol nem csak egyoldalú a folyamat, vagyis nem csak beszédfelismerésről vagy beszédelőállításról beszélhetünk, hanem ezek körkörös működéséről, ami a beszélők váltakozásából fakad, vagyis fontos lépés, hogy ezt a folyamatot gépileg tudjuk lekövetni, előjelezni. Ezt a dinamikus rendszert, amelyben a beszélők váltakozását igyekeznek leírni gépi eszközökkel a beszédtechnológiában beszélődetektálásnak (speaker diarization) neveznek. A beszélődetektálás során a folyamatos beszédben az akusztikai jelből gépi úton határozzuk meg, hogy mikor ki beszél (Jin et al. 2004) (1.2. ábra). A beszélődetektálás során a folyamatos társalgásokat automatikusan beszélőkre szegmentáljuk, ezzel a társalgások szövegeit beszélőkhöz rendelhetjük, így a szöveg sokkal könnyebben feldolgozható más, például tartalomkinyerő algoritmusok számára.



1.2. ábra

A beszélődetektálás sematikus annotációja

A beszélődetektálást alapvetően két alfeladatra lehet bontani (Jin et al. 2004; Kotti et al. 2008): a beszélő szerinti szegmentálásra (speaker segmentation) és a beszélőosztályozásra (speaker clustering). Az első feladat célja elkülöníteni az azonos beszélőktől származó beszédrészeket, a második részfolyamatban pedig ezeket a szegmentumokat kell osztályozni a beszélők szerint (1.3. ábra). A két részfeladaton kívül fontos feladat még a beszéddetektálás és az egyszerre beszélések detektálása.



1.3. ábra

A beszélődetektáló folyamatábrája

Az elmúlt évtizedekben tudományos közösségek jöttek létre, hogy a beszélődetektálást, mint beszédtechnológiai célt megvalósítsák (Nemzetközi, Szabványok és Technológiák Nemzetközi Intézete, NIST: National Institute of Standards and Technology, http://www.itl.nist.gov/iad/mig/tests/rt/). A kutatás fejlődését mindig valamilyen valós igény határozta meg. A 1990-as évek végén és a 2000-es évek elején a korai munkákban a telefonos beszélgetések és a híradások voltak a kutatások középpontjában, ahol a beszélődetektálást a műsorok automatikus lejegyzéséhez használták fel. Az automatikus lejegyzés többszintű volt, amely tartalmazta az elhangzott szöveget, illetve a beszélőváltásokat. 2002-től nőtt az érdeklődés az élő, spontán társalgások iránt (meating domain), amely 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 multimodális technológiáknak meg kell felelniük a valós igényeknek, mint a tartalmi indexelés, tartalmi kivonatolás, mind a verbális és a nem verbális 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. Emezekből olyan információkat nyerhetnek ki, amelyek segítségével a társalgások tartalma strukturálható, elemezhető (Ajmera–Wooters 2003; Barras et al. 2004; Wooters et al. 2004).

A beszélődetektálás megvalósítására jelentős mennyiségű kutatás történt idegen nyelvre (Tritschler–Gopinath 1999; Sivakumaran–Fortuna–Ariyaeeinia 2001; Lu–Zhang 2002a; Cettolo–Vescovi 2003; sian Cheng–min Wang 2003; Vescovi–Cettolo–Rizzi 2003; stb.). Magyar nyelvre azonban mind ez ideig 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.

1.1. Beszédprodukció és beszédpercepció

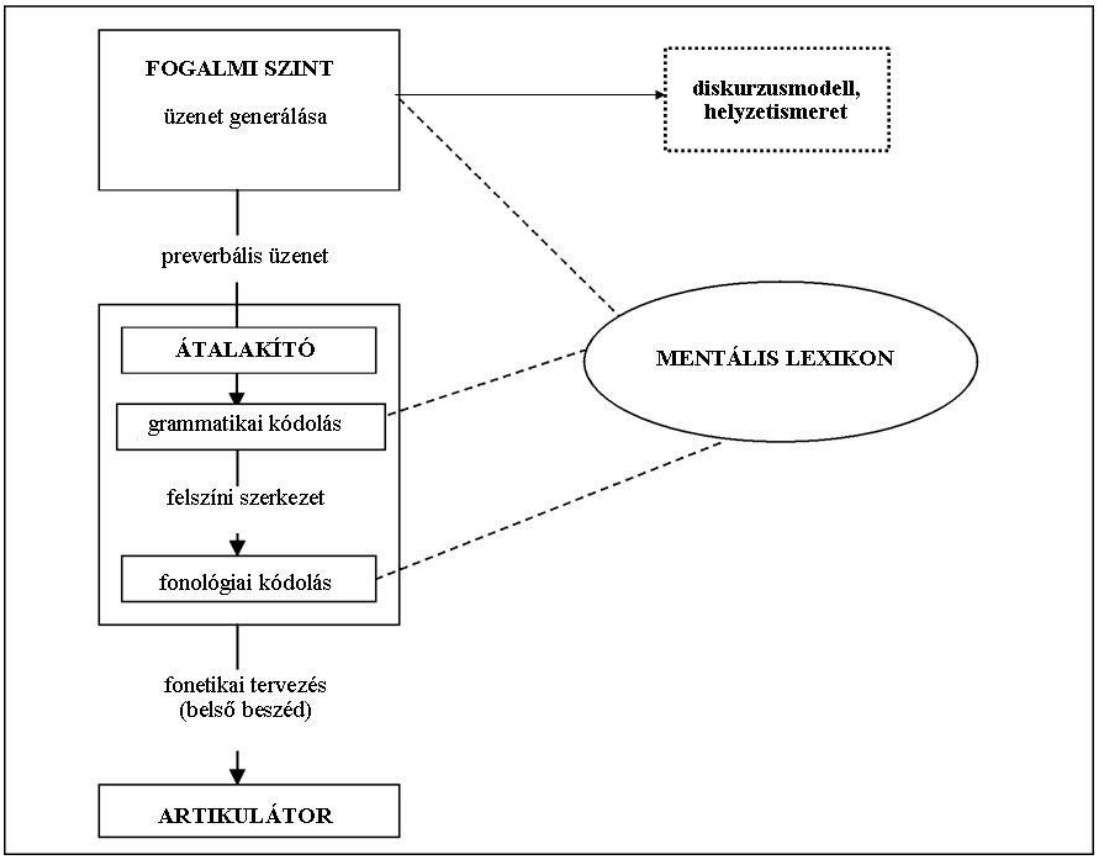
A beszéd önmagában alkalmas mindenféle információ továbbítására (amennyiben nem korlátozza azt a zaj, a távolság, a nem ép beszédszervek stb.). Ez a kommunikációs csatorna rendelkezik a legbonyolultabb kóddal valamennyi csatorna közül. A nyelv mint kódrendszer és maga a beszéd kulturális termék, amely az ember fejlődéstörténete során alakult ki (Buda 2009). A beszéd évekig tartó tanulással elsajátítható emberi képesség, amely az egyén fejlődéséhez és társadalmi szocializációjához kötődik (Levelt 1989). A beszéd pszicholingvisztikai szempontból alapvetően két nagy részre osztható: a beszédprodukcióra és a beszédpercepcióra. A beszéd legtermészetesebb formája a társalgás, amelyben – beszélőnként is – a beszédképzés és a beszédfeldolgozás váltogatja egymást (Gósy 2005).

1.1.1. Beszédprodukciós modellek

A beszédprodukció a megnyilatkozás szándékától a kiejtésig tart (Gósy 2005). A beszédképzés bonyolult biológiai mechanizmus, amely az agyi tervezés és irányítás eredménye. A beszédprodukció kutatása igen nehéz feladat, mivel nem lehet közvetlenül azokat a rejtett folyamatműködéseket vizsgálni, amelyek a megnyilatkozástól a kivitelezésig realizálódnak. Közvetett úton azonban lehet vizsgálni. A múlt század hatvanas éveitől kezdve megindultak a beszédprodukciós kutatások, amelyek a hezitációkat, illetve a nyelvbotlásokat elemezték, így kívántak meg közelebb jutni a tervezési folyamatok megértéséhez (Goldman−Eisler 1968; Fry 1973; Garman 1990; Meyer 1993; Fromkin 1999; Gósy 1998; Harley 2001; Horváth 2009; Gyarmathy 2011). A feltételezések szerint a hibás szerkezeteket ugyanazok a folyamatok hozhatják létre, amelyek a normának megfelelőket, így az eltérésekből következtetni lehet a beszédtervezési mechanizmus különböző szintjeinek működésére.

A Levelt-féle beszédprodukciós modell a kutatók körében a legtöbbet használt modell (1989). Levelt szerint a mondatokat nem szavak összekapcsolásával állítjuk elő, hanem az „intonációs frázison” belül tervezzük meg, amely egy vagy két fonológiai frázisból áll. Ezeken a fonológiai frázisokon belül a szavak szintaktikailag

összekapcsoltak. Levelt beszédprodukciós modellje különféle szintekből tevődik össze (1.4. ábra).



1.4. ábra

A beszédprodukció modellje (Levelt 1989 nyomán)

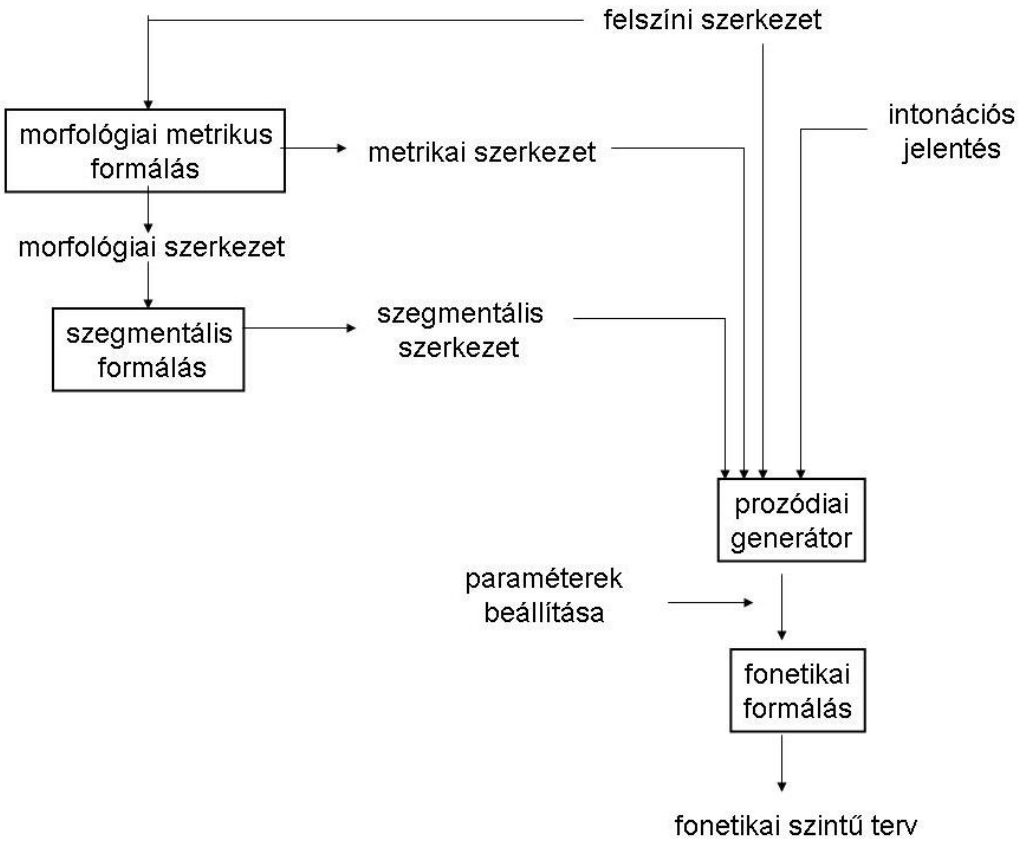
Az üzenet generálása a fogalmi tervezés (konceptualizálás) folyamata során történik, amikor a beszélő közölni kíván valamit. A konceptualizáció során jön létre kimenetként a preverbális üzenet. Levelt modelljében nagyon fontos szerepet kap a kommunikációs helyzetből fakadó előzetes ismeret, tudás, hiszen a kommunikációs helyzetben a beszélőnek az üzenet létrehozásához széleskörű háttérismeretekre van szüksége (ezek a legtöbb esetben a hosszú távú memóriában tárolódnak), amelyek a beszélő tapasztalatait tartalmazzák önmagáról és az őt körülvevő világról. Ez az előzetes tudás lehet procedurális, illetve deklaratív ismeret. A preverbális üzenetet két folyamat előzi meg: a makro- és a mikrotervezés. A makrotervezés folyamatában a beszélő elhatározza, hogy miről kíván beszélni és milyen céllal. Már ekkor is lehetséges bizonyos nyelvi

meghatározottság, de a vizualitás, az asszociációs kapcsolatok vannak előnyben. A mikrotervezés során a beszélő az egyes beszédaktusokhoz nyelvi formákat rendel, ennek során figyelembe veszi saját, a cél által meghatározott sorrendiségi stratégiáját és kognitív stílusát.

A fogalmi tervezés kimenete a preverbális üzenet, amely aztán továbbítódik az átalakítóhoz. Az átalakítóban a preverbális üzenet a grammatikai és fonológiai kódolón keresztül transzformálódik felszíni szerkezetté, vagyis a konceptuális szerkezet nyelvi szerkezetté alakul. Az átalakítóban lévő grammatikai és fonológiai átalakítón keresztül kapcsolatban van a lexikonnal. A mentális lexikonban tárolt lemmák előhívását követően a grammatikai átalakító a szintaktikai szerkezeteket hozza létre, amelyek szemantikailag és szintaktikailag meghatározott jelentését adják a szónak, de a fonológiailag meghatározottságát nem. A grammatikai kódoló után a fonológiai kódolás történik, ahol kialakul szó végső, adott szerkezetnek megfelelő alakja, a lexéma.

Az átalakítóban a felszíni szerkezet a kívánt lemmák aktiválódása és a grammatikai struktúra tervezése után áll elő. A fonológiai kódolás előkészíti a kiejtést az adott nyelvre jellemző fonológiai szabályoknak megfelelően, így kialakulnak a lemmák végleges fonológiai formái, amelyek megfelelnek a lexémáknak. Az átalakítóból az információ a fonetikai tervezés szintjére kerül, amely az artikulátorhoz továbbítódik. Az artikulátorban az artikuláció tervezése történik (vagyis a beszédszervek izom mozgatásának előkészítése) és a kivitelezés.

A szupraszegmentális jellemzők funkciói a beszédprodukció felől határozhatók meg (Levelt 1989; Markó 2005). Az 1.5. ábrán a Levelt-féle beszédprodukciós modell (Levelt 1989) egyszerűsített részletét, egész pontosan a prozódia generátort láthatjuk, amely a szupraszegmentális szerkezet megvalósulását szemlélteti.



1.5. ábra

A folyamatos beszéd képzésének Levelt-féle modellje (Levelt 1989)

A prozódia generálásának alapja és elsődleges bemenete a felszíni szerkezet. A felszíni szerkezet elméletileg a beszélő által már gondolatilag megfogalmazott megnyilatkozás váza, amelyeket a beszéd folyamán szeretne kifejezni. A felszíni szerkezetet a metrikai szerkezet módosítja, meghatározza a hangsúlyok helyét és a hangsúlyozási mintát. A felszíni szerkezet a metrikai szerkezetre döntő hatással van, ami a kötött hangsúlyú nyelvre fokozottabban igaz. A harmadik bemenet a szegmentális szerkezet, amelyet kvázi a kiejteni kívánt beszédhang-sorozat. A metrikai szerkezet a felszíni szerkezet közvetett függvénye. A negyedik bemenetként a beszélő a felszíni szerkezethez hozzáadja az intonációs jelentést, ami a beszélő szándékát és érzelmeit tükrözi. A megnyilatkozás szupraszegmentális szerkezetét ebből a négy bemenetből generálja a beszélő. A prozódia generátor kimenetén a paraméterek beállítása az artikulációs paraméterekre vonatkozik.

A magyar nyelvben a szupraszegmentális szerkezetet alapvetően a felszíni szerkezet és az intonációs jelentés határozza meg (Markó 2005), hiszen a prozódia elsődleges funkciója a magyar nyelvben a megnyilatkozás felszíni szerkezetének és a beszélő viszonyulásának, érzelmeinek, szándékainak realizálása. Ezek mellett természetesen megjelenik a metrikai szerkezet és a szegmentális szerkezet hatása is. A szókezdet detektálásában elsődlegesen a prozódiai szerkezet nyomon követése és az általa hordozott információ kinyerése a cél, azon belül pedig a szó eleji hangsúlyok kinyerése.

1.1.2. A beszédpercepciós mechanizmus

A beszédpercepció az a folyamat, amelyben a hallgató az artikuláció során realizálódott, folyamatos akusztikai jelből képes diszkrét nyelvi egységeket létrehozni (Gósy 2005). Ez a folyamat igen komplex mechanizmus, amelynek működéséről relatíve kevés biztos tudással rendelkezünk. A beszédpercepció két nagy részből tevődik össze: a beszédészlelésből és a beszédmegértésből. A beszédészlelés során azonosítani tudjuk a beszédhangokat, hangkapcsolatokat. Ebben a folyamatban a nyelv által használt különbségeket ismeri fel a beszédészlelési rendszer. A beszédmegértés során a beszédészlelés során felismert hangok, hangkapcsolatok kódolása történik diszkrét nyelvi egységekké. Itt alakul ki a szavak, mondatok, szövegek megértése, értelmezése (Cutler–Norris 1979; Gernsbacher–Faust 1991; Gernsbacher 1994; Pickering 1999; Gósy 1999, 2000, 2005). A beszédpercepcióban nem válik el élesen a beszéd értelmes és értelem nélküli egysége (pl. beszédhangok, hangkapcsolatok, jelentéses hangsorok) vagy a szegmentális és a szupraszegmentális része (Gósy 2005). „A beszédészlelés a kommunikációs lánc harmadik nagy egységét képzi a beszédprodukció és a nyelvi jel után” (Mády 2008: 1). A kommunikációs folyamatban a beszélő is ugyanúgy alkalmazkodik a percepcióhoz, mint a hallás a beszédprodukcióhoz; a beszélő igyekszik olyan ejtésben megvalósítani megnyilatkozásait, amelyre az emberi beszédészlelés érzékeny. Ebből adódik, hogy a beszélői-hallgatói szerepben mindkét fél előzetes ismeretekkel rendelkezik mind a beszédprodukcióról, mind a beszédpercepcióról, illetve a nyelvi jelről, ezáltal a kommunikáció is egyszerre zajló folyamatok komplex működése (Mády 2008).

Minden nyelvnek saját percepciós bázisa van, amely lehetővé teszi az elhangzó közléssorozat feldolgozását „mégpedig úgy, hogy a nyelvi sajátosságok meghatározók,

és hatnak a fiziológiai rendszer működésére" (Gósy 2004: 164). A beszédpercepció rugalmasságának bizonyítéka, hogy a megnyilatkozás különféle realizációinak azonosítására is képes, hiszen képesek vagyunk megérteni a zajos vagy a rosszul artikulált, gyors beszédet (Clark−Clark 1977; Gósy 2008), illetve a különböző beszélők különböző stílusú, artikulációs megformáltságú közléseit (Gósy 2000d). A megértési kulcsokat a beszédben található invariáns és redundáns jegyek őrzik. Az invariáns jegyek azokat a beszédparaméterek, amelyek az adott szegmentumot egyértelműen meghatározzák, és biztosítják az észlelést (Gósy 2004). A redundáns beszédparaméterek információtöbbletet adhatnak zajos vagy zavart helyzetben. Az akusztikum és a percepció viszonylatában az invariáns jegyek az elsődleges akusztikai kulcsoknak mondhatók, míg a redundáns jegyek tartalmazzák a másodlagos akusztikai kulcsokat.

A nyelv- és életkorspecifikus, tipikus beszédészlelés és megértés ép halláson alapszik. A hallás során még nem történik beszédelemzés, de előzetes döntések történnek a frekvencia, az intenzitás és az idő szempontjából. Felismerjük, hogy milyen típusú hangjelenséget hallottunk; beszéd vagy nem beszéd, gyors volt-e vagy lassú, halk vagy hangos, kellemes vagy kellemetlen (Pauka 1982; Gernsbacher 1994; Gósy 2004, 2005). A beszédészlelés (amelynek működését a szeriális észlelés, a beszédhang-differenciálás, a transzformációs észlelés, a ritmusészlelés és a vizuális észlelés biztosítja) három szinten megy végbe. Az akusztikai elemzés a beérkezett jel időtartamáról, zöngés vagy zöngétlen voltáról, frekvenciaszerkezetéről, intenzitásviszonyairól nyújt információt. A fonetikai osztályozás során az akusztikai jelekhez hozzárendelődnek az adott nyelvre jellemző beszédhangok. A fonológiai szinten a nyelvspecifikus szabályok alapján megtörténik a fonémadöntés. Az azonosított hangsorokhoz a beszédmegértés során kapcsolódik jelentés, és a szemantikai és szintaktikai elemzések is visszahatnak az észlelési folyamatokra. Ez a szint a szavak, szókapcsolatok, mondatok és szövegek tartalmának a megértését jelenti. A hallott és megértett közlések összekapcsolása az emlékezetben már korábban tárolt ismeretekkel, tapasztalatokkal az asszociációk vagy értelmezés szintjén történik (Gósy 2005; Honbolygó 2009).

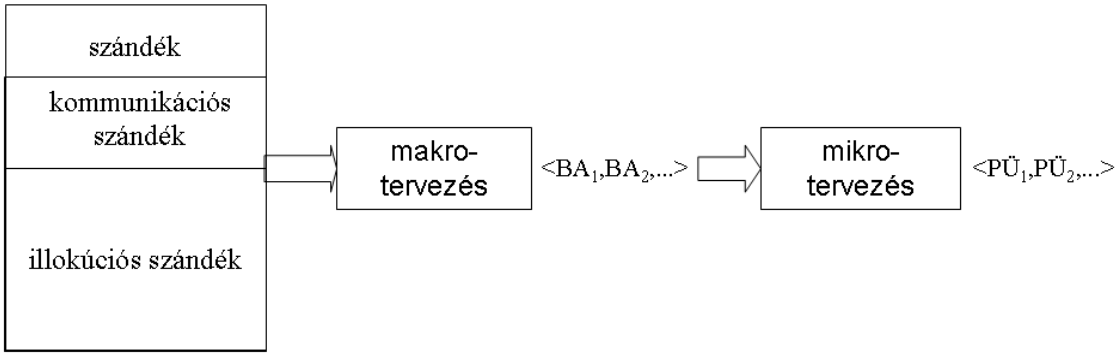
A beszédfeldolgozási mechanizmusba bármelyik szinten keletkezhet hiba, tehát az észlelés és a megértés, illetve az asszociációk és az értelmezés szintjén is. A hallási

feldolgozás is működhet hibásan ép hallás esetén is, ha például az átviteli körülmények nem megfelelőek (pl. zaj) (Gósy 2008; Gyarmathy 2008).

1.2. Spontán beszéd (tervezés és stílus)

Napjainkban egyre fontosabbá válik a verbális nyelvhasználat, mindamellett hogy az írott és a verbális nyelv mintegy köztes formái (cset, sms stb.) egyre nagyobb tért hódítanak (Thurlow 2006). Ezt jól mutatja, hogy a nemzetközi konferenciák és publikációk többségének a témája a spontán beszéd valamilyen aspektusának vizsgálata. A spontán beszéden azt folyamatot értjük, amikor a beszélő előzetesen nem készül fel a mondanivalójára, a közölni kívánt gondolatokat az adott pillanatban önti nyelvi formába, azaz a gondolatok kialakulása, kiválogatása és a kivitelezés az adott helyzetben történik (Gósy 2005a), vagyis a spontán beszéd tervezése és a kivitelezése gyakorlatilag egyszerre zajlik. Minthogy kutatásunkban a spontán társalgásokat vizsgáljuk, így az eredményleírások jelentős része erre a kommunikációs stílusra szorítkozik.

A beszédszándéktól a kivitelezésig nagyon sok változó hat a beszédre. Levelt (1989) beszédprodukciós modelljében a szándéktól a kivitelezésig számtalan részfolyamatot különböztet meg, mint arról korábban volt szó, amelyek a rájuk ható tényezők következtében változtathatják aktivitásukat és befolyásukat a végeredményre. Az elhangzó beszédet, vagyis a beszédszándékot, a beszélés gondolata előzi meg. A makrotervezés előtt is feltételez bizonyos részfolyamatokat, szándékokat (1.6. ábra). Ilyen szándék a kommunikációs szándék, amelynek csak egy része fejeződik ki a beszédaktusokban.



1.6. ábra

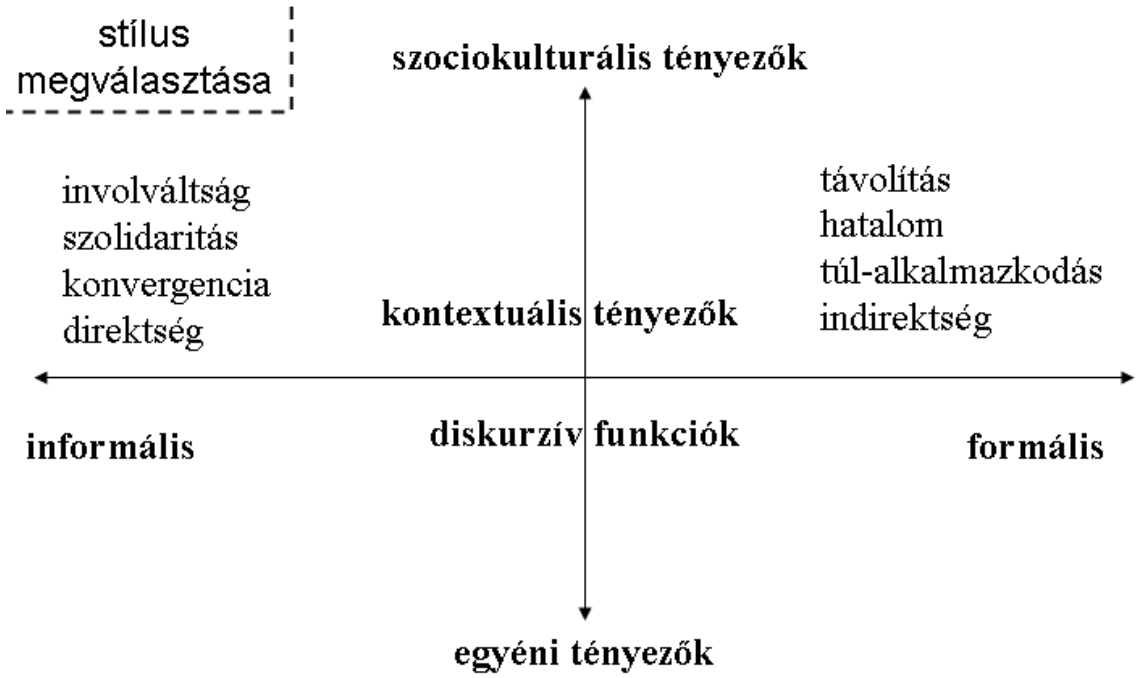
A szándéktól a kivitelezésig tartó folyamatábra és a stílust meghatározó tényezők kétdimenziós ábrázolása (BA=beszédaktus; PÜ=preverbális üzenet) (Levelt 1989 alapján)

Feltételezhetjük, hogy a kommunikációs szándékban kerül kiválasztásra, hogy milyen beszédstílusban nyilatkozzunk meg. A stílus megválasztása igen fontos, amelyet számos tényező befolyásol. A nyelvhasználat egyik legjellegzetesebb összetevője a stílus. Maga a stílus a kommunikáció folyamatában az üzenet közlésének módja, vagyis a „hogyan mondjuk”. A jelen disszertációnak nem célja bemutatni a különféle stíluselméleteket, csupán azokat, amelyek célkitűzéseinknek legjobban megfelelnek.

A társas konstruktivista stílusmegközelítések előzményeként megemlítendő Eskenazi (1993), aki a stílusokat három tengely mentén tartja elhelyezhetőnek a külső szituációs tényezőkből kiindulva: (i) a formalitás mentén, (ii) a familiaritás (közelség vs. ismeretlenség) és (iii) az érthetőség tengelyén. Az érthetőség dimenziója ebben a keretben már előre mutat a beszédalkalmazkodás felé, amely a beszédpartner befogadói folyamataihoz való alkalmazkodást jelenti, mind a kognitív pragmatika elméletei felé (Tomasello 1999; Verschueren 1999; Clark 1996; Tomasello et al. 2005; Smith 2007). Előzményként még ebben a keretben megemlítendő Bell „hallgatóságra tervezés” modellje („audience design”, Bell 1984, 2001; Bell–Johnson 1997), amelynek alapgondolata a beszélőnek a hallgatósághoz való dinamikus adaptációja. Elméletében a beszélő és a hallgató a diskurzus folyamatában dinamikusan vagy közelednek egymáshoz beszédmódjukban (konvergál), vagy távolodnak egymástól (divergál), például a szolidaritás, a távolság vagy a tekintély fenntartása vagy teremtése függvényében (Bell 1984: 162; Giles–Coupland–Coupland 1991: 18).

Az interakcionális stílusvizsgálatok a szociolingvisztikai stíluselméletek legújabb, „harmadik generációjaként” (Eckert 2010) vagy „interakcionális stilisztikaként” (Selting 2008) is jellemezhető. Vizsgálatuk tárgya a nyelv- és stílushasználat („styling”), amelyet a társas jelentések, az egyéni identitás és társas kötődések konstruálásának keretében vizsgáltak (Schiffrin 1996; Eckert 2000; Schilling-Estesc 2004; Buchholtz 2004; Tannen 2005; Coupland 2007; Selting 2008).

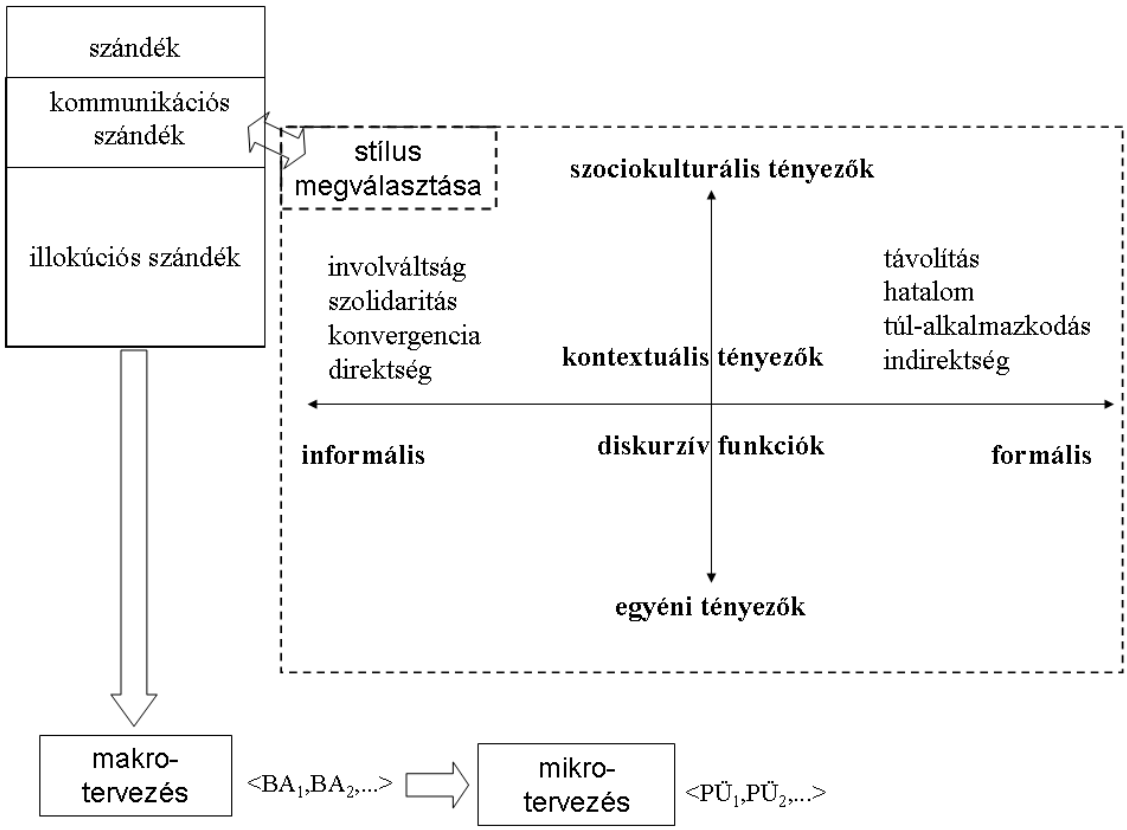
Az „interakcionális stílusvizsgálat" modell középpontjában a beszélő alkotó tevékenysége áll (vö. „speaker agency”, ill. „performativity”). A középpontban a társas interakció jelentéslétrehozó dinamizmusa és a beszélő aktív identitás- és világalakítása áll. A fő kérdés ebben a megközelítésben az, hogy a beszélők miért alkalmazzák egy adott szituációban az adott stilisztikai választásokat. A beszélők a stilisztikai választásokba mind külső, tágabb (mint beszédpartner, téma, helyszín), mind a közvetlen beszélői interakciós célokat is bevonják (Coupland 1980, 1985, 2001, 2007a, 2007b; Eckert 2000; Schilling-Estes 1998, 2004a, 2004b, 2004c) (Bartha−Hámori 2010). Bartha és Hámori (2010) munkájukban az egyes stíluselméletek egy ábrába implementálták, elhelyezve a stílus változatait két kontinuum mentén (1.7. ábra).



1.7. ábra

Stílusválasztási modell (Bartha–Hámori 2009 alapján)

A levelti beszédprodukciós modellbe ez a stílusválasztási modell jól illeszkedik, amely a kommunikációs szándékba implementálható (1.8. ábra), így a társalgásban létrejövő megnyilatkozások stílusváltozatai, vagyis a választott nyelvi forma érthetőbbé válik.



1.8. ábra

A szándéktól a kivitelezésig tartó modell a stílusválasztási modell függvényében (Levelt 1989 és Bartha–Hámori 2009 alapján)

A társalgásban a stílusválasztások dinamikusan változnak a 1.7. és 1.8. ábrán látható kontinuumok mentén. A stílusváltozatok a stíluselemzés szokásos szintjein elemezhetők (fonológiai, lexikai, morfoszintaktikai, pragmatikai). A társalgás elemzéséhez érdemes tekintetbe venni a konverzációra jellemző sajátosságokat, elsősorban a diskurzusjelenségeket (pl. szünetek, fordulóhossz, egyszerre beszélések, beszédmennyiség). Mindezek mellett fontosnak tartjuk az akusztikai, illetve fonetikai elemzések elvégzését is (alaphangmagasság, beszédhang-realizációk, hangszínezet stb.)

A spontán beszédnek is több stílusváltozata van: a narratíva során a beszélő egy vagy több témáról viszonylag hosszabban közöl valamit; a társalgásban pedig két vagy több személy felváltva közli gondolatait, folyamatosan egymásra reagálva. A beszédalapú játékok, az irányított spontán beszéd egyaránt a spontán beszédhez sorolhatók (Gósy 2005).

A dialógusok esetében a beszélők szinte egyik pillanatról a másikra tervezik meg a közlést annak függvényében, hogy miként kell reagálniuk az elhangzottakra. A beszélőváltások kivitelezése, egy beszélgetés kezdeményezése vagy befejezése különféle szabályok szerint zajlik. A tervezés és a kivitelezés egyidejű működéséből következően a spontán beszédnek tipikus megvalósulási formája van. A pszicholingvisztikai szemléletű spontánbeszéd-kutatások egyik fő témája a megakadások vizsgálata. Jóllehet igen sok kutató vizsgálja a megakadásokat mind nemzetközileg, mind a hazai szakirodalom tanúsága szerint, a terminológiája nem egységes, sőt nemegyszer a terminológiai különbségek nehézzé teszik a különböző publikációk eredményeinek értelemezését és összehasonlíthatóságát (Shirberg 2001). Mint említettük, mind a nemzetközi, mind a hazai szakirodalom igen gazdag ebben a témában. A hazai kutatásokat tekintve Gósy (1998, 2001, 2002, 2003, 2008), Horváth (2004, 2007a, 2007b), Gyarmathy (2009, 2010, 2011), Markó (2004, 2006), Bóna (2006) vizsgálatai irányadók. A hezitálások módszeres elemzése úttörőnek tekinthető (Horváth 2009). Megállapították, hogy a spontán beszédben a szünetek gyakran kitöltött szünetekként jelennek meg, ellentétben a felolvasással, és igazolták, hogy a hezitálások több funkcióval is bírnak.

Az akusztikai-fonetikai célú kutatások száma is igen jelentős. Kimutatták, hogy a magánhangzók spontán beszédben redukálódnak, azaz a neutralizált, semleges célkonfiguráció felé tolódik el a képzésük (Gósy 2006; Beke–Gráczi 2008; Beke–Szaszák 2009; Gráczi–Horváth 2010). A koartikulációban is jelentős különbséget adatolt Markó, Gráczi és Bóna (2009). Vizsgálataikkal kimutatták, hogy a zöngésségi hasonulás nem szabályos megvalósulásai gyakrabban jelentek meg a spontán beszédben, mint felolvasásban. A spontán és a felolvasás között temporális jellemzőkben különbségek adatolhatók. Váradi (2010) eredményei azt igazolják, hogy a spontán beszédben a beszédszakaszok rövidebbek, mint felolvasásban, illetve a szünet a tagolás és a levegővétel funkción kívül beszédtervezési időt is biztosít a beszélő számára. A spontán beszédre tovább jellemző a hangtörlődés, amely a szavak közel 12%-át érinti (Váradi 2012). A spontán beszéd egy igen jelentős kihívása a tagolás. Számos hazai és nemzetközi kutatás készült már ebben a témában, de még mindig számtalan kérdés megválaszolatlan és sok részterületet ez idáig nem vizsgáltak. Gósy (2003) kutatása alapján virtuális mondatokat feltételez a spontán beszédben, ez a

vizsgálat a humán percepció felől közelítette a tagolást. Kísérletében a virtuális mondatok percepciós határait a szünetek (néma vagy kitöltött) megjelenése, az alaphangmagasság csökkenése vagy lebegése és a szemantikai strukturáltság eredményezte. A tanulmány következetése, hogy létezik ’virtuális mondat’, amely a beszélő produkciójában és a hallgató percepciójában is jelen van, és amelynek akusztikai-fonetikai összefüggései vannak, s a hallgatók nagymértékben egyöntetűen ismerik fel ezeket a spontán beszédben. Más kísérletek azonban ezt nem vagy csak igen korlátozott mértékben tudták kimutatni (Markó 2010), ami felveti a jelentős egyéni különbségek kérdését, illetőleg a hangzó szöveg egyéb (kevéssé vizsgált) tényezőinek hatását (Váradi 2013).

A spontán beszéd tagolása a beszédtechnológia egyik fontos kihívása. Magyar spontán beszéd automatikus tagolására algoritmust elsőként Beke és Szaszák készített (2011). Kutatásukban alapvetően prozódiai jellemzőket, az osztályozáshoz pedig nem-ellenőrzött tanuláson alapuló k-közép eljárást használtak fel. Az eredmények alapján nem mondható ki, hogy a spontán beszéd ilyen fajta megközelítésével automatikusan egyértelműen tagolható lenne. Hasonló probléma a spontán beszédben a szintaxis kérdése is, amely szintén csak nagyon korlátozottan kutatható az írott nyelvre kialakított szabályrendszerekkel (Szaszák et al. 2011; Szaszák–Beke 2012; Beke–Szaszák 2012).

A spontán beszéd megismerése tehát napjainkban jelentős kihívás a nyelvészet, a beszédtechnológia és még számos más tudományterület számára is.

1.3. Társalgás

Ahogyan azt fent említettük, a prototipikus társalgás során a résztvevői szerepek folyamatosan váltakoznak a beszélő és a hallgató között. A társalgó felek ugyanabban a fizikai kontextusban helyezkednek el, és ebben a helyzetben a megnyilatkozásaikat interakciós viszonyban, spontán módon hozzák létre, és szóban közvetítik azt (ez mára megváltozóban van az internetes konferenciabeszélgetések következtében).

A spontán társalgásokat az antropológiai és szociológiai gyökerű irányzatok, mint az etnometodológia, sőt az etnográfia kezdték el új szempontok alapján kutatni. A kutatások eredményei azt mutatták, hogy a társalgás szabályszerűségeken alapszik, még ha nem is alapvetően nyelvi oldalról, inkább interakciós oldalról (Boronkai 2009; Pléh 2012). Az empirikus megközelítésükben a társalgás kiemelkedően fontos eleme a forduló, amely az egy beszélő által létrehozott megnyilatkozás (szövegrész). A társalgásban dinamikusan váltakoznak a részvevők, amelyet a beszélőváltás jelenségeként azonosítanak. A beszélőváltások jellegzetes szomszédsági párokban jelennek meg, más néven kérdés-válasz szekvenciákban, amelyek a társalgás felépítését adják. A beszédlépések határainak azonosításában a témakapcsolódás és a témaválasztás meghatározó.

A társalgások stílusa – mint ahogyan azt a spontán beszédről szóló alfejezetben kifejtettük – dinamikusan változhat. A formalitás dimenziójában a két szélsőérték a formális és az informális, amely egy kontinuum, vagyis számtalan változatban jelenhet meg. Az informális beszélgetések azok, amelyek a mindennapi kommunikációban ismerőseinkkel, családtagjainkkal, barátainkkal folytatunk. A társalgások másik típusát a formális adja, amely lehet például hivatali beszélgetés, tanórai kommunikáció, vizsga vagy terápiás beszélgetés. A formalitás dimenzióján kívül szociokulturális, egyéni, kontextuális tényezők, illetve diskurzív funkciók is befolyásolják (Iványi 2001; Teun−van Dijk 2006).

A társalgás olyan dialógusokból épül fel, amelyekben az információk kisebb részletekben oda-vissza áramlanak a felek között. A dialógusokban a kontextustól, a társalgó féltől stb. függően folyamatosan változik a mondanivaló tartalma, szerkezete, nyelvi formája, de ezek az egységek adják összekapcsolódva a szöveget magát, amely

nyelvi-grammatikai, viselkedési és pragmatikai szabályozók mentén rendeződnek (Teun−van Dijk 2006; Boronkai 2009).

A beszédaktus-elmélet (Austin 1962/1990) a párbeszédet olyan szövegformaként határozza meg, amelynek beszédaktusai két nagy csoportba sorolhatók (Boronkai 2009): (i) a beszélgetést szervező aktusok: az üdvözlés, a búcsúzás, a témaváltást jelző vagy a szövegegységet záró megnyilatkozások; (ii) a beszédaktust konstituálók: a kérdés-felelet vagy a parancs-teljesítés/elutasítás szekvencia-párjai (Huszár 1994).

Franke (1990) beszédaktus-elméleti keretében a dialógus beszédaktusai a kezdeményező és a reagáló aktusok két nagy kategóriájába tartoznak, s így a beszédaktusoknak a beszédcselekvés végrehajtásán kívül a párbeszéd szerkezeti kialakításában is nagy szerepük van (Boronkai 2009).

Az interakció-elmélet középpontjában szintén a közös cél áll, amely a társalgás eredményességét biztosító együttműködésben foglalható össze. Ennek fő eszközei az ún. konverzációs maximák (Grice 1975). Ezek az elméletek már nagy hangsúlyt fektetnek a szövegek implikált információira, az előfeltevésekre, a kontextus és a szituáció szerepére is.

Langleben párbeszédleírása (1983) Grice (1975) munkájából indul ki. Elméletében a dialógus alapegységének nem a beszédfordulót, vagyis a turn-t tartja, hanem a replikát, amelyek láncszerűen időben követik egymást. Az első replikát ingernek (replikastimulus, Rstim), a másodikat válasznak (replikareakció, Rresp) nevezi, amely a résztvevők értelmezési folyamataiban visszafelé hatva alakítják ki, gyakran módosítva az első replika jelentését. A válaszreplika sok esetben a következő egység ingerreplikája is egyben (Iványi 2001).

A kognitív nyelvészet a párbeszédes társalgást a nyelvi tevékenység egyik legjellemzőbb megnyilvánulásaként értelmezi, amely magában foglalja a megnyilatkozások létrehozásának és befogadásának egymást feltételező folyamatát. Ennek eredményeképpen jönnek létre az értelemmel bíró nyelvi közlések, a kommunikációs üzenetek. A társalgás két alapvető résztvevői szerepe a megnyilatkozó és a befogadó, akik az üzenet létrehozása és értelmezése érdekében nyelvi tevékenységet végeznek (Clark 1996).

A beszélgetések célja a valóság létrehozása és fenntartása, amelyben a cselekedeteink által alakul ki a bennünket körülvevő világ. A konverzációelemzés fő

célja, hogy elemezze a beszédbe elegyedést, a beszélőváltásokat, a beszédlépéseket, beszédlezárásokat, és leírja ezek szabályszerűségeit. Ezen társalgásrészek a társalgó felek interakciójában jönnek létre, szekvenciálisan, vagyis lépésről lépésre. A sikeres kommunikációhoz természetesen szükség van a társalgásban részt vevő felek együttműködési szándékára, a személybeli, valamint a tér- és időbeli vonatkozások közös ismeretére, és a beszélgetés időbeli sorrendjének megszervezésére. Az így létrejövő dialógusok egy nyelvi interakciós folyamat elvei és szabályai szerint épülnek fel (Iványi 2001).

A társalgáselemző vizsgálatok általában a beszélgetések globális és lokális struktúrájának elemzésére irányulnak. A globális struktúrában a belebonyolódási és kihátrálási stratégiákat, a lokális szerkezetben pedig főként a fenti szempontok között is szereplő beszélőváltás mechanizmusát, a szekvenciális rendezettséget és a hibajavítások diskurzusszervező szerepét vizsgálják.

Saját kutatásunkban a társalgás lokális szerkezetével foglalkozunk a tekintetben, hogy miként jönnek létre a beszélőváltások, és hogy ezeket hogyan lehet gépileg modellezni, detektálni.

1.3.1. Beszédforduló

A társalgás alapegysége a beszédforduló (a terminus szinonimái: beszélőváltás, beszédlépés). 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 (pl. Brown−Yule 1989; Iványi 2001; Markó–Dér 2011). A beszédforduló lehet egyetlen mondat, egy frázis, vagy lehetnek különböző lexikai konstrukciók (1.9. ábra).

A	(nem is lehet) hát annak
T2	(ugyanúgy a szülés) 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 mert én azt tudom ám hogy mentek ott a dolgok hozzá nem nyúlhatott a másik orvos az ügyeletes orvos nem nyúlhatott a másik orvos (betegéhez)

1.9. ábra

A társalgás szekvenciális szerkezetének reprezentálása

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; 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 beszédlépés végét a megnyilatkozó az esetek többségében jelzi. A beszédlépés jelzésére számos akusztikai, szintaktikai, gesztus, pragmatikai stb. eszköz áll a beszélő rendelkezésére.

A beszélőváltás rendje tehát szabályok által vezérelt, és amelyeket a következőképpen írja le Iványi (2001, alapul véve Sacks–Schegloff–Jefferson munkáját):

1. A beszéd jogának odaítélése a váltásra alkalmas helyen történik.

a) A következő beszélőt az addigi beszélő választja külválasztással. Csak a választottnak szabad, illetve kell beszélnie.

b) Ha nem történik külválasztás, a beszélgetőpartnerek egyike ön(ki)választás vagy belválasztás útján nyeri el a beszélő rangját. Aki elsőként kezd beszélni, az nyer jogot egy beszédlépés megtételére.

c) Ha sem kül-, sem belválasztás nem történik meg, az eredeti beszélő folytathatja a lépését.

2. 1.c) esetén a lehetséges beszélőváltás következő helyén újra az 1. a)–c) szabálysorozat lép érvénybe, és így tovább minden beszélőváltásra megfelelő helyen (vö. Sacks et al. 1978).

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.

Bár a beszédlépés szemantikája, szintaktikai felépítése, fonológiai és intonációs jellemzői megkönnyítik a beszélőváltást – minthogy lehetséges lezárásukat és a potenciális átadásra alkalmas helyet előre kiszámíthatóvá teszik –, ennek ellenére szerkezetük nyelvtanfüggetlen korlátozásoknak is alá van vetve. Interakcionális és nem nyelvi feltétel például, hogy a beszélő lépése lezárásával a beszédhez való jogát elveszítheti. Ennek elkerülésére és a beszédlépés lehetséges végpontja kitolásának az érdekében a lépés mellékmondatok beépítése, hozzáillesztése, melléknévhalmozás stb.

segítségével kiterjeszthető. Egy ily módon bővített beszédlépés létrehozásához szükség van mindkét interakciós partner beleegyezésére: a kezdeményezést a beszélő teszi, a befogadónak azonban le kell mondania a beszédhez való jogáról és lehetőségéről.

Az egyes megnyilatkozások a beszélgetőpartnerek interaktív együttműködése során jönnek létre a beszélőváltás modellben, ahol az egyes beszédlépések kontextusérzékenyek, vagyis a beszédlépések belső felépítése kizárólag az aktuális beszélgetés kontextusától függ. A kontextus háromrészes: „interaktív szerkezete rendelkezik egy visszautaló – azaz a megelőző beszédlépéssel való kapcsolatát demonstráló –, egy előremutató – a következő, lehetséges vagy kívánatos szerkezetű beszédlépés realizálását meghatározó – és egy, az aktuális helyzettel összefüggő elemmel, amelyek mindegyike a mindenkori beszédlépés feladataihoz igazodik (vö. Sacks–Schegloff–Jefferson 1978)" (idézi: Iványi 2001). A társalgás során az egyén úgy alakítja mondanivalóját, hogy az megfeleljen a társalgó félnek, vagyis nagyban beszédpartnerfüggő (lásd részletesebben a Beszélőalkalmazkodás fejezetben). A társalgásban nemcsak a beszélőnek kell figyelni a beszédpartnerére, hanem fordítottan is. Tehát a hallgatónak is aktív funkciója: biztosítja a beszélőt figyelméről különböző interakciós technikákkal, mint szemkontaktus, testtartás (vö. Streeck 1983; Bergmann 1988). Így egy megnyilatkozás hosszúságának, tartalmának és struktúrájának oka sok esetben a különböző „címzettek" váltakozása és a mindenkori befogadónak a beszélőre gyakorolt hatása (azaz interaktív közreműködése) lehet (Goodwin 1979; Iványi 2001).

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 jelez. Az elmúlt évtizedekben számos jellemzőt vizsgáltak, hogy megállapítsák, milyen szerepet játszanak 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), testmozgással (kézmozdulat befejezésével vagy egy megfeszített kézpozíció ellazulásával), konvencionális nyelvi jelekkel, szófordulatokkal – ez a diskurzusjelölőknek felelnek meg – (*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 beszédátadásban. 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 (1998) 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 és munkatársai (1982) Margaret Thatcherrel készített interjúkat elemeztek, amelynek eredménye az volt, hogy kimutatták több helyen is átvette a szót a riporter a beszélgetés során, még akkor is, ha ő maga nem is akarta átadni a szót. Ezeknél a pontoknál az alaphangmagasság csökkenése volt megfigyelhető éppúgy, mint a szándékolt beszédlépés végénél. Tehát az alapfrekvencia változása eredményezte a riporter közbevágásait, amellyel a szerzők bizonyították az F0 fontos szerepét a beszélőváltásokban. Stephens és Beattie (1986) egy olyan kísérletet terveztek, ahol a részvevőknek a társalgás átiratait kellett olvasni, illetve annak hanganyagát meghallgatni. Az átirat és a hanganyag társalgásokból kivágott beszédforduló-közepi és -végi megnyilatkozásokat tartalmaztak. Az eredmények azt mutatták, hogy a hanganyag alapján a részvevők el tudták dönteni, hogy beszédlépés-végi megnyilatkozásról volt szó. Cutler és Pearson (1986) vizsgálatai szerint csak néhány dallammenet létezik, amely szóátadást jelezne, ezek karakterisztikája azonban kevésbé lényeges, mint a fonológiai szerepük.

A szóátadást szintén jelezheti hosszabb néma vagy kitöltött szünet (Maclay–Oswood 1959). Beattie (1977) azt figyelte meg, hogy a társalgásban részvevők 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-közepi, illetve -végi megnyilatkozások elkülönítésére (Stephens–Beattie 1986).

Ford és Thompson (1996) eredményei azt mutatták, hogy a szünet segít befejezetté tenni az intonációs egységeket. A legrövidebb észlelhető szünetet 0,3 percben határozták meg. Jóllehet ugyanakkor a szünet nem minden esetben jelzi előre az intonáció végét. Local és Kelly (1986) két funkcióját feltételezték a szünetnek: az első, amikor a beszédjelben szünet keletezik, amely lezárásra utal; a másik, amikor a szünet mégis a beszélő folytatását jelzi előre. A 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 kap a testmozgás, a gesztusok. Számos kutatás kimutatta, hogy a testmozgás igen fontos és integrált része a spontán társalgások beszédfordulóinak szerveződésében (Beattie 1979; Lerner 2003). Kendon 1994, 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 kifejezésében: a kéz vagy kar mozdulat lezárása előre jelezheti a beszédlépés végét; ennek ellentétjeként a mozdulat folytatása a szóátadást gátolhatja meg.

Az utóbbi évtizedben egyre fontosabbnak tűnik a diskurzusjelölők szerepe a beszédfordulók előrejelzésében (Sacks et al. 1974; Schiffrin 1987; Wennerston−Siegel 2003) (lásd bővebben a Diskurzusjelölők c. fejezetet).

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 testmozgás, a gesztusok, a kitöltött szünetek, a szomszédsági párok struktúrája. Ford és Thompson (1996) a szintaktikai szerkezeteket, az intonációt és a pragmatikai lezártságot vizsgálták. Az eredményeik azt mutatták, hogy a teljes szintaktikai egységet az

intonáció (F0 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 átmeneti lehetőséget a hallgatónak, hogy átvegye a szót. Wennerston és Siegel (2003) szintén a beszédlépéseket mint komplex folyamatot vizsgálta, főként fonológiai és szintaktikai interakciók együttes működéseként. Tanulmányaikban az intonáció, a szünet és a szintaxis együttes működését 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. Az 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, 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 tovább azt is, hogy amikor hosszabb szünet jelent meg (0,5 percnél nagyobb), akkor a beszélő folytatta beszédét továbbra is. Ezt azzal magyarázták, hogy a hallgatónak 0,3 percnél lett volna lehetősége átvenni a szót (Ford–Thomphson 1996), de ezt elmulasztotta, így a beszélő folytatta a megnyilatkozását. Ugyanakkor azonban ezt nagyban egyénfüggőnek találták.

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 magyarra is történtek már kísérletek, főként a prozódia és a szintaxis együttes működésével kapcsolatosan a beszédfordulókban (Markó 2006; Németh 2007, 2008; Bata 2009; Lerch 2011).

1.3.2. Diskurzusjelölők

A verbális jelzések közül a legtöbbet kutatott téma a diskurzusjelölők szerepe a beszédlépések szerkezetében. 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 ismertető jegyei lehetnek a beszédfordulóknak, így nagyban hozzájárulhatnak a beszélőszegmentáláshoz, a diskurzus működésének megértéséhez (Fraser 1999: 931, Louwerse–Mitchell 2003: 199; Markó−Dér 2011).

A diskurzusjelölőket egyre kiterjedtebben vizsgálják, illetve a kutatások által leírt eredményeket egyre több helyen alkalmazzák. Ezen nyelvi jelenségekkel foglalkozó tanulmányok egy része igyekszik az összes lehetséges DJ-t vizsgálni (angol nyelvben: Schourup 1982; Schiffrin 1987; Watts 1989; olasz nyelven: Bazzanella 1990; francia nyelvben Holker 1991), a kutatások másik része azonban csak egy-egy DJ kiterjedt elemzését tűzte ki célul (Lakoff 1973; Svartvik 1980; Owen 1981; James 1983; Carlson 1984; Schiffrin 1985; Watts 1986; Blakemore 1988).

A humán-humán társalgásokban számtalan diskurzusjelölőt találhatunk. A DJ egy olyan nyelvi egység, amelyet a beszélők a közléseik elején használhatnak jelezvén, hogy a diskurzus mely állapotában vannak (Cohen 1984; Grosz−Sidner 1986). A DJ arra is alkalmas például, hogy egész diskurzusszerkezetet változtasson meg, ilyen például az angolban a *by the way* (’mellesleg’, ’egyébként’).

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. A kategorizálását azonban nagyban megnehezíti, hogy igen heterogén csoport az eredetüket tekintve, 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; Markó 2005, 2006). Diskurzusjelölők nagyobb számban a beszélt nyelvben fordulnak elő, de egyes írott műfajokban is megtalálhatók (Schiffrin 2001; Dér 2006).

A diskurzusjelelölők számos funkciója közül az egyik legjelentősebb, hogy sokszor utalhat a beszédlépések végére, illetve annak kezdetére. Mivel a jelen dolgozat alapvető célkitűzése a beszédfordulók automatikus előrejelzése, így a DJ-k ezen szerepéről szóló munkákat mutatjuk be itt.

A diskurzusjelölőkkel kapcsolatban alapvetően azt kell elemezni, hogy a beszédváltásokban milyen szerepet játszanak, hogyan működtetik a beszédlépések kezdetét, fenntartását, végét, hiszen ez közelebb vihet a DJ-k szemantikai és

pragmatikai megkülönböztetéséhez. A feladat elsőre igen könnyűnek tűnhet, azonban korántsem olyan egyszerű. Számtalan szabályt, előírást és jellemzőt írtak le a kutatók a DJ szemantikai, illetve pragmatikai funkcióinak elkülönítésére (Turner 1999). Az alapfeladat tehát az, hogy hogyan lehet elkülöníteni azt, hogy a DJ szemantikai vagy pragmatikai szerepben áll-e a beszédfordulóban (Fisher 2000).

A társalgásban minden jelenségnek funkcionális szerepe van. Például a szünet gondolkodási időt biztosít a beszélőnek a gondolatainak a megformálására, a kötőszavaknak a beszédlépések összeköttetését biztosítja, a háttércsatorna-jelzések a hallgatói figyelmet jelzik a beszélő felé. A töltelékszók előhívása és kivitelezése közben a közlőnek lehetősége van a megfelelő szó aktiválására (Gósy–Horváth 2009).

A kutatások többsége megegyezik abban, hogy a DJ-knek fontos szerepe van beszédlépések szerveződésében, de önmagukban nem elégségesek. Schiffrin (1987) szintén amellett érvel, hogy számos különböző tényező vesz részt a társalgás szerkezetváltásában (1987). Emellett a DJ-knek beszédlépésvéget jelölő szerepet feltételez (Schiffrin 1987), illetve, beszédlépést fenntartó szerepet: pl. a *you know* (Schiffrin 1987). Sacks és munkatársai (1974) a DJ-k beszédlépéskezdő szerepét hangsúlyozták *well*, *but*, *and*, *so* DJ-ket vizsgálva. Sok esetben az adott diskurzusjelölő akár multifunkcionális is lehet: lehet beszédlépést indítói, zárói és fenntartói szerepben is, mint ahogy az angolban az *uhm, yes* (Fisher 2000). Azonban éppen ezeket a hangsorokat az angolban tartalmas szóknak, DJ-knek és fillereknek egyaránt tartják, ez részben a szerzőtől, a szemlélettől, a funkcionális vizsgálattól és az adott közlésrészlet jellemzőitől is függ.

A diskurzusjelölő egyik gyakori funkciója a háttércsatorna-jelzés. Egyes elméletek szerint a háttércsatorna-jelzés lehet beszédforduló, de többségében elkülönítik attól (Yngve 1970; Duncan 1972). A jelen munkában a háttércsatorna-jelzést nem tekintjük azonosnak a beszédlépéssel. A háttércsatorna-jelzés (lásd részletesebben az Egyszerre beszélés fejezetben), mint az *aha*, *ühüm* stb., jelzik, hogy a kommunikációs csatorna még nyitott, és egyben jelzi a beszélőnek, hogy folytathatja a mondanivalóját, illetve a hallgató nem kívánja átvenni a szót (Markó 2005).

A diskurzusjelölők funkciói a következő lehetnek: (i) kapcsolatot teremt a megnyilatkozások között (Knott−Sanders 1998; Fraser 1999), (ii) összeköti a globális és a lokális diskurzusstruktúrákat (Schiffrin 1987; Redeker 1990), (iii) jelezheti a nem

kívánt második részt a szomszédsági párban (Pomerantz 1984; Schiffrin 1987); (iv) nyugtázó szerepet is betölthet (Jefferson 1984); (v) lehet háttércsatorna-jelzések (Yngve 1970); (vi) segítheti a hallgatót az információk feldolgozásában a spontán társalgásban (Fox Tree−Schrock 1999); (vii) a visszacsatolást adhat a beszélőnek a hallgatói jelenlétről (Fox Tree−Schrock 2002). Bangerter és munkatársai (2004) az *uh-huh, yeah, right, okay* DJ-ket vizsgálták, amiket ők „projekt marker”-nek hívtak. A DJ-ket mint átmeneteket vizsgálták a telefonbeszélgetések különböző részeiben. Az elemzéseik szerint a DJ-nek inkább a globális szintű diskurzusszerkezetekben van jelentősége, kevésbé a lokális beszédfordulók szerkezetében.

A diskurzusjelölők fontos szerepet töltenek be a diskurzus struktúrájában, így felhasználhatók a diskurzusban lévő egységek szegmentálásában, és az egységek közötti kapcsolatok feltérképezésében. A diskurzusjelölők jelentésének meghatározása gyakran hozzásegíthet a kommunikációs üzenet megértéséhez is. A diskurzusjelölőket és azok szerepét vizsgálták nagy szövegkorpuszokban (Mann−Thompson 1988; Sporleder−Lascarides 2008), érvelő dialógusokban (Cohen 1984), interjúkban (Schiffrin 1987; Hirschberg−Litman 1993), illetve olyan társalgásokban, amelyek igen interaktívak, így igen gyakoriak bennük a beszélőváltások, mint például a feladatvezérelt társalgások (Heeman−Allen 1999) és spontán társalgások (Popescu-Belis−Zufferey 2006). Ezekben a kutatásokban kimutatták, hogy a diskurzusjelölők fontos szerepet töltenek be társalgási egységek határainak meghatározásában, illetve azok kommunikációs szándékának kijelölésében (vö. Hirschberg−Litman 1993; Heeman−Allen 1999; Popescu-Belis−Zufferey 2006).

Petukhova és Bunt (2009) kutatásukban a diskurzusjelölőket mint többfunkciós jelenségeket vizsgálta. Már Shiffrin (1987) is kimutatta, hogy a diskurzusjelölőknek több funkciója lehet egy időben, például az *and* gondolatmenet-összekötői szerepe, és a beszédlépés folytatására tett cselekvés szerepe is lehet egyazon időben. A szemantikai keretében Dynamic Interpretation Theory (DIT, Bunt 2000) úgy vizsgálja a diskurzusjelölőket mint a dialógus egyszerre több tevékenységet végző elemeit, mint például a folytatásra ösztönző cselekvés, váltásra ösztönző cselekvés, kommunikatív visszajelzés stb.

A nemzetközi kutatásokban a diskurzusjelölőknek jelentős szerepet tulajdonítanak a társalgások egységeinek meghatározásában is, illetve az egységek kommunikációs

funkcióinak azonosításában (vö. Heeman–Allen 1999; Popescu-Belis–Zufferey 2006; Volha Petukhova–Bunt 2009). Angol nyelvű társalgásban a beszédlépések 44%-ának az elején szerepelt diskurzusjelölő (Heeman–Allen 1999). Dér (2012) szintén nagy arányú előfordulást adatolt a beszédlépések elején a magyar nyelvre. Úgy tűnik tehát, hogy ez az arány több nyelven is igazolódott.

Magyar nyelven a verbális eszközök spontán társalgásokban való vizsgálatával, azon belüli a diskurzusjelölők szerepével kevés tanulmány foglakozott (Dér 2010; Markó–Dér 2011, Shirm 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 tipikusan az általa bevezetett beszédlépés elején fordul elő, mint például a kötőszói eredetű jelölők (pl. *tehát*, *és*, *de*). Kivételt képeznek azok, amelyek kötőszóként sem vagy nem mindig tagmondatkezdő helyzetűek (pl. *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 nyelvészeti indíttatású kutatások kimutatták, hogy a diskurzusjelölők nagy százalékban fordulnak elő a beszédlépés elején, jelezvén a szóátvétel szándékát. Ez a jelenség fontos jellemzője lehet a beszédváltás-detektálásnak. Tehát az első feladat az, hogy a folyamatos beszédben valamiképpen automatikusan megtaláljuk a diskurzusjelölőket.

A társalgásstruktúrák kutatásának fontos célja, hogy feltárja, hogy milyen lehetséges jelentései és funkciói vannak a diskurzusjelölőknek a dialógusokban, és azok hogyan korrelálnak a megnyilatkozás megfigyelhető jellemzőivel (prozódia, szintaxis, lexikai infirációk), ahhoz hogy azokat minél pontosabban lehessen automatikusan felismerni és osztályozni (Petukhova−Bunt 2009).

A DIT a diskurzusjelölőkhöz a következő funkciókat társítja: feladat (Task), visszacsatolás az előző közlés feldolgozásáról a beszélő által (Auto-feedback), vagy

más partnerek által (Allo-feedback), beszélő megnyilatkozásának produkciós nehézségei (Own-Communication Management) vagy más partnereké (Partner–Communication Management), a beszélő számára idő, hogy folyatatni tudja a beszédet (Time Management), kapcsolatok létrehozása és fenntartása (Contact Management), a következő beszédlépés bevezetése (Turn Management), egy lehetséges mód, hogy a beszélő megváltoztassa a párbeszéd struktúráját (Dialogue Structuring), az interakció szociális aspektusának mutatója (Social Obligations Management). Petukhova és Bunt (2009) ezeket a funkciókat osztályozta automatikusan spontán társalgásokban. A funkciók elkülönítésére prozódiai, szógyakorisági és kollokációs jellemzőket használt. A prozódiai jellemzőkön belül az F0-t, az energiát, zöngeminőségre utaló jellemzőket, beszédtempót, beszédhang-időtartamot mért. A szógyakorisághoz szózsák modellt alkalmazott (lásd részletesebben Zhang et al. 2010). A lexikális felépítést bigrammokkal és trigramokkal modellezte. Az eredményei szerint a diskurzusjelölők valóban fontos szerepet töltenek be a DIT által meghatározott funkciók betöltésében. Ebből kiemelnénk a beszédváltások működtetését. Továbbá azt is bizonyították, hogy tanuló algoritmusokkal a diskurzusjelölők azonosíthatók a szövegben, és funkcióik elkülöníthetők.

Kawahara és munkatársai (2004) a diskurzusjelölőket mint a topik kezdetének markerét használták fel a szóbeli előadások szegmentáláshoz. Megfigyeléseik szerint a beszédlépés első egységében tipikusan diskurzusjelölők fordulnak elő, például a *Now, I would like to…* esetében. A rendszerük egyik kiindulópontja a szünet. Ugyanis a szóbeli előadásokban a beszélő szünetet tart a diák között, illetve az egyes topikok előtt. Ezt kihasználva jelölhetők ki a beszédlépések. A diskurzusjelölőket szógyakoriságuk alapján nem ellenőrzött tanulással, küszöbértéket használva indexelik. Az eredményeik bizonyítják a diskurzusjelölők szerepét a topikok szegmentálásában a szóbeli előadásokban. Sajnos, a rendszer működése igen nagyban függ a beszédstílustól.

A diskurzusjelölők lényeges közös vonásaként említi több tanulmány, hogy elkülönülnek a mondat többi részétől, prozódiailag függetlenek, amit szóban az egység előtti és utáni szünet, írásban központozás jelez (például: Zwicky 1985: 303–304, idézi Fraser 1999: 933; Jucker–Ziv 1998: 3, idézi González 2004: 43–44). A nemzetközi szakirodalomban megjelenő, a diskurzusjelölők előtti és utáni szünettartásra irányuló kísérleti fonetikai kutatások eredményei ugyanakkor ellentmondani látszanak ennek.

Dér és Markó kutatásukban azt találták (2012), hogy sem a diskurzusjelölőket megelőzően, sem őket követően nem jelentős mértékű a szünetek előfordulása. A beszédlépéskezdő helyzet is csak bizonyos diskurzusjelölőket jellemez, általánosságban nem elegendő támpont a diskurzusjelölői és a szintaktikai szerep elkülönítéséhez. Megállapították továbbá, hogy a vizsgált homofónok többsége gyakrabban fordult elő tagmondatzáró helyzetben diskurzusjelölőként, mint szintaktikai szerepében. A szegmentális realizációban semmilyen különbség nem volt felfedezhető: minden lenizált alak megjelent mindkét funkcióban.

1.3.3. Egyszerre beszélések (átfedő beszéd)

A társalgás során a monologikus beszédre jellemző akusztikai és nyelvtani szabályok nagyszámú varianciája mellett más nehezítő jelenségek is megjelennek. 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). A jelfeldolgozás szempontjából a társalgás során több esetben van olyan időkeret, amikor a beszélők egyszerre nyilatkoznak meg, vagyis a tőlük származó jel párhuzamosan zajlik. Ez azért kiemelkedően fontos, a jelfeldolgozás szempontjából, mert ezek korlátozottan feldolgozható részei a beszédnek.

Az egyszerre beszéléseken belül a korai kutatások kezdetben csupán a társalgások egy igen érdekes jelenségeként vizsgálták a háttércsatorna-jelzéseket, amelyeknek tipikusan szociális interakciós szerepet tulajdonítottak (Yngve 1970; Sacks et al. 1974; Duncan and Fiske 1985; Ward 1997). Számos megnevezése létezik ennek a fajta hallgatói viselkedésnek: a folyamatos figyelem jele (‘signals of continued attention’) (Fries 1952), elismerés (‘recognition’) (Rosenfeld 1966, 1967), párhuzamos visszacsatolás (‘concurrent feedback’) (Krauss−Weinheimer 1966), kísérő jelzések (‘accompaniment signals’) (Kendon 1967), hallgatói válasz (‘listener responses’) (Dittmann−Llewellyn 1967, 1968; Bavelas et al. 2002), jóváhagyás (‘assent terms’) (Schegloff 1968; Leet−Pellegrini 1980), háttércsatorna-jelzés (‘backchannels’) (Yngve 1970; Duncan 1972, 1973; Duncan−Niederehe 1974 Duncan−Fiske 1977, 1985), ösztönzés (‘encourager‘) (Edelsky 1981), korlátozott visszacsatolás (‘encourager‘) (Kraut et al. 1982), hallgatói fogékonyság (‘responsive listener cues’) (Miller et al.

1985), minimális válaszok (‘minimal responses‘) (Fishman 1978; DeFrancisco 1991; Bennett−Jarvis 1991), reaktív tokenek (‘reactive tokens’) (Clancy et al. 1996), elismerő tokenek (‘acknowledgment tokens‘) (Jefferson 1984, 1983/1993; Drummond−Hopper 1993a, 1993c), nyugtázó tokenek (‘receipt tokens‘) (Heritage 1984).

A háttércsatorna-jelzések többsége olyan jelenség, amely igen rövid időtartamú, a hallgató a beszélő megnyilatkozása alatt produkálja őket, ezek funkcionálisan nem a szóátvételre irányulnak, sokkal inkább a beszélőt motiválják beszédének folytatására (1.10. ábra).

A	és akko[r] ez mérhető ezt tudják fél év van rá tudják hogy mi a helyzet a másik pedig ez a bizonyos zenetörténet dolgozat nálam is puskáznak össze-vissza puskáznak ugyanakko[r] pedig azt mondom hogy meghallgatta egyszer tőlem jegyzetelte
T2	*ühüm*
T1	*aha*
	A megírta a puskát
T2	*így van*
A	leírta
T1	*igen*
A	kijavítom legalább négyszer hallotta és hogyha azt mondom neki legközelebb hogy súbert [Schubert] akkor nem azt mondja rá hogy artista hanem azt mondja rá hogy zeneszerző
T2	*igen* azt hiszem hogy ezt így
(**T2**	<u>így lehet ??</u>)
(**A**	<u>*nem nem*</u>)
A	*szóval*
T2	nem nem ilyen terrorral
T1	*aha*
	A nem úgy csinálom mint a magyarórán
T2	*aha*
	A magyarórán nyilván nem engedem hogy puskázzon
T2	*aha igen*
A	mer[t] mer[t] muszáj
T1	*persze*
A	énekórán puskázik puskázzon ne feltűnően kész nem érdekel **de de** hallott ezekről
(**A**	<u>a dolgokról és az</u>)
T2	*igen* és akkor már nem
A	az bőven elég bőven elég nekem

1.10. ábra

Háttércsatorna-jelzések (*dőlt betűvel*) és az egyszerre beszélések (aláhúzással jelölve) a társalgás során

Az általános definíció szerint ez a jelenség alapvetően arra szolgál, hogy a beszélőt informálják arról, hogy a hallgató az üzenetet megkapta, megértette, elfogadta vagy

valamilyen hiba miatt a beszélőt kiegészítésre kéri. A későbbi kutatások alapján azonban kiderült, hogy ennél jóval tágabb a háttércsatorna-jelzések funkciója. A háttércsatorna-jelzések besorolását nehezíti, hogy számos változatban jelenhetnek meg. Sok esetben lehetnek önmagukban állók, mint például *mmm*, *hm*, *aha*, de nagyon sokszor kapcsolódnak egymáshoz, mint például *ja ühüm*. De megjelenhetnek más tartalmas kifejezésekkel is, amelyek legtöbbször igenlő vagy tagadó szavak, például *aha igen*, *aha tudom*. Mivel nagy a háttércsatorna-jelzések megvalósulásainak száma, ezért a gyakorlatban a leggyakoribb előfordulásokat szokás elemezni, illetve kategorizálni. A szakirodalomban alapvetően két megközelítés létezik ezen hallgatói reakció kategorizálására. Az egyik az egyesítő megközelítés (Fries 1952), amely úgy kezeli a háttércsatorna-jelzéseket, mint egy kategóriát vagy osztályt, amely egy csoportja a különböző hallgatói válaszok megvalósulásainak. A másik a felosztó megközelítés, amely főként a felosztást veszi alapul, és amely etnometodológiai megközelítésű. Ez a megközelítés egy vagy több diszkrét szekvenciális összefüggésben elemzi a háttércsatorna-jelzéseket, és igyekszik bizonyítani, hogy minden ilyen jelenség megkülönböztető interakciós funkcióval bír.

Az egyesítő megközelítést számos területen alkalmazzák, mint a nyelvészetben, a nyelv és a társadalom viszonylatában, interkulturális megközelítésekben, a kísérleti- és szociálpszichológiában. A korai kutatások az 1960-as évektől a kísérleti és szociálpszichológia felől érkeztek (vö. Rosenfeld 1966, 1967; Dittmann−Llewellyn 1967, 1968; Kendon 1967), majd folytatódtak napjainkig (vö. Bavelas et al. 2002). A korai munkákban a háttércsatorna-jelzések két jellegzetességére koncentráltak. Egyrészt a háttércsatorna-jelzések strukturális jellegzetességére, másrészt a társalgásban betöltött általános, illetve még inkább specifikus szerepéről a beszélgetés tervezésében és megértésében. A strukturális jellegzetességgel foglalkozó tanulmányok többsége a háttércsatorna-jelenségeket más nem verbális jelenségek, mint kézmozgás, gesztus, nevetés összekapcsolódásában vizsgálta a társalgásokban (vö. Birdwhistell 1962; Kendon 1967; Dittmann–Llewellyn 1967, 1968; Brunner 1979; Bavelas et al. 2002). A társalgásban betöltött szereppel foglakozó kutatások a háttércsatorna-jelzéseket mint a nem-beszédlépés meglétét elemezték a társalgásokban (Yngve 1970; Duncan 1972, 1973; Duncan–Niederehe 1974; Duncan–Fiske 1977, 1985). Az újabb kutatások szerint ezek a jelenségek nem beszédlépések, és nem hordoznak új információt, hanem

elősegítik a társalgás folyamatosságát, dinamikus szerkezetét. Továbbá az is jellemző rájuk, hogy többségükben átfedésben jelentkeznek a beszélő megnyilatkozásának utolsó szakaszával, ugyanakkor függenek az aktuális beszélő következő beszédlépésétől. Azt is megfigyelték, hogy számos esetben 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 (Pipek 2007).

A felosztó megközelítésben a kutatók a társalgásokban egy-egy háttércsatorna-jelenséget elemeztek szekvenciális környezetben. Ellentétben az egyesítő megközelítéssel, ebben a felfogásban nem vizsgálják a háttércsatorna-jelzések és a külső körülmények viszonyát. Vizsgálják azonban a háttércsatorna-jelzések beszédlépések szerveződésében betöltött szerepét. 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* és *mm hm* (Schegloff 1982; Jefferson 1983/1993, 1984; Drummond–Hopper 1993a, 1993b, 1993c); *oh* (Heritage 1984), *wow* és *good* (Goodwin 1986), *okay* (Beach 1993, 1995; Pillet–Shore 2003), és *mm* (Gardner 2001). Schegloff (1982) az *uh huh* háttércsatorna-jelzést vizsgálta angol nyelvű társalgásokban. Az eredményei azt mutatták, hogy a háttércsatorna-jelzéseket lehet úgy vizsgálni, mint interakciós teljesítményt, amely részben önmaga is szervezi a beszédlépések felépítését. Megfigyelte, hogy az *uh huh* leggyakrabban folytatásra való ösztönzésre használják; ennek szerepe, hogy ösztönözze az előző felszólalót, hogy folytassa a beszélgetést. Schegloff azt is megjegyzi, hogy az *uh*, *huh* jelzést hosszabb megnyilatkozásokkor is használják. Jefferson (1983/1993, 1984) a *mm*, *hm* és a *yeah* háttércsatorna-jelzéseket vizsgálta, mint elismerő jelenségeket. Az eredményei azt mutatták, hogy a két háttércsatorna-jelzés funkcionálisan és szekvenciálisan különbözik egymástól. Az *mm hm* inkább passzív részvételt jelezhet, míg a *yeah* a beszélői együttműködést. A passzív részvétel Jefferson (1984) szerint itt azt jelenti, hogy a hallgató jelzi, hogy a beszélő egy hosszabb megnyilatkozásának közepén jár, és folytatni kívánja.

A háttércsatorna-jelzések osztályozása nehéz feladat. Mivel a tanulmányok többsége a háttércsatorna-jelzést analitikus szempontból elemzi a társalgásokban, mint az egyéni jelek egyedülállóságát, ezért az osztályozási rendszer kialakítására az összegző megközelítések törekedtek. Az osztályozási rendszerek többsége Duncan és munkatársai munkásságán alapszik (Duncan 1972, 1973; Duncan–Fiske 1977, 1985; Duncan–Niederehe 1974). A 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. Számos szerző ezzel szemben a háttércsatorna-jelzéseket beszédlépésként értelmezi, ezt a felfogás azért problematikus, mivel maga a beszédlépés sem egyértelműen definiált. Ez vezetett Schegloff (1982) felvetéséhez, hogy a háttércsatorna-jelzések turn-státuszát eseti elbírálás alapján kell értékelni a lokális szekvenciális környezetet figyelembevételével, utalva a szekvenciális és az interakciós célokra, amelyek megteremtik ezt a környezetet.

A háttércsatorna-jelzések beszédlépésként és nem-beszédlépésként történő definiálásának problémája más kritériumok kereséséhez vezetett. Új kritériumként jelent meg a *floor*, vagyis a beszédjog (vö. Hayashi–Hayashi 1991) és a formája és/vagy a hallgatói válaszok szekvenciális felépítése (Tottie 1991; Clancy et al. 1996). Ennek a kritériumnak az alkalmazása azonban ugyanolyan problémás kérdéseket vetett fel, mint a beszédlépés.

A háttércsatorna-jelzések kategorizálását nehezíti tovább az, hogy használatuk erősen kultúrafüggő. Az ötvenes évektől kezdve megnőtt azon tanulmányok száma, amelyek a háttércsatorna-jelzéseket különböző kultúrákban vizsgálták (Hayashi–Hayashi 1991; Clancy et al. 1996; Ward–Tsukahara 2000; Beach–Lindstrom's 1992). Tao és Thompson az amerikai angol és a japán nyelvben vizsgálta a háttércsatorna-jelzések megvalósulásait. Kutatásukban kimutatták, hogy az amerikai angolban sokkal gyakoribb ez a jelenség, mint a japánban. Továbbá azt is leírták, hogy az amerikai angolban gyakran jelenik meg átfedő beszédként, és a beszédlépések végén, míg a japánban sokkal gyakrabban a beszédlépés végén, ritkábban átfedő beszédként. A két nyelvben a funkciót tekintve is találtak különbséget. Míg a japánban többségében megerősítő, támogató funkcióban használják a háttércsatorna-jelzéseket, addig az amerikai angolban sok esetben folytatásra biztató funkcióban is megjelenik.

Az egyszerre beszélések vizsgálata is igen nagy számú (pl.: Çetin–Shriberg 2006, Markó 2005). 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 beszédek 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 csak korlátozottan feldolgozható szakaszai a beszédnek. Shriberg és munkatársai kutatásukban kimutatták, hogy az egyszerre beszélések igen fontos részei a társalgásnak, illetve hogy közel azonosan nagy számban fordulnak elő mind spontán társalgásokban, mind telefonbeszélgetésekben (Shriberg et al. 2001). Az átfedő beszéd aránya a spontán társalgásokban meglehetősen nagynak mondható (Cetin–Shriberg 2006; Gráczi–Bata 2010). Beattie a beszélőváltásokat elemezve (1983) 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. Jurafsky és munkatársai (1997) angol társalgásokban kimutatta, hogy a megnyilatkozások 19%-át a háttércsatorna-jelzések teszik ki. Ennél is nagyobb arányban fordulnak elő az egyszerre beszélések a japán társalgásokban (Majnard 1989). Cetin és Shriberg (2006) 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 (2009) 1,7–3%-ot adatolt kutatásában, amelyben spontán társalgásokat elemzett.

Az egyszerre beszélések tehát láthatóan nagy arányban fordulnak elő, és ez az átfedő beszéd funkciójából adódik. A társalgás során ugyanis az egyszerre beszélés alapvetően két funkciót tölt be. Egyrészt megerősítő szerepe van (pl. *igen*, *aha*, *ühüm* stb., vö. backchannel), másrészt versengő funkciójú, amikor a társalgás egyik szereplője át kívánja venni a szót, és már az alatt elkezdi a beszédét, amikor az aktuálisan beszélő még nem fejezte be mondanivalóját (Iványi 2001; Hámori 2006; Bata 2009). Az egyszerre beszélések megvalósulásainak funkcionális kategorizálására számos munka született, lásd például Bata (2009). A szerző az első csoportba azokat az egyszerre beszéléseket sorolta, amikor a társalgó felek egyszerre szólaltak meg, vagyis egyszerre

két önkiválasztás történt. A másodikba azok kerültek, amikor a hallgató a beszélő megnyilatkozására azonnal reagálni próbált; vagyis ez csupán egy pillanatnyi reagálási szándék volt. Ekkor a hallgatói szándék csak egy rövid kommentár, valójában nem kívánja átvenni a szót. A harmadik csoportba azon egyszerre beszélések kerültek, amikor a szóátvétel sikertelenül jött létre. Ez esetben a beszélő magához kívánja ragadni a szót, de az eredeti beszélő folytatja a saját megnyilatkozását. A negyedik kategóriába a beszédforduló vége előtti megszólalások kerültek. Ekkor aa következő beszélő már elég információhoz jutott ahhoz, hogy átvegye a szót, és az eredeti beszélő is hajlandó megválni beszédjogától.

Dér (2012) elemzéseiben az egyszerre beszélések tipikus megnyilvánulási formáit írta le. Az első kategóriába a figyelem verbális visszajelzései kerültek. Ekkor a hallgatót különféle igenlő kifejezéseket közöl (*ühüm*, *igen*, *ja*, *aha*), miközben az egyik beszédpartner beszél. Kiemeli, hogy ebben az esetben ezek a közlések nem kérdésre vagy feltevésre reagálnak, sokkal inkább a hallgató figyelmét, érdeklődését fejezi ki. A második kategóriába a kommentárokat sorolta be. Ezek saját vélekedések, mint például az egyetértés, a csodálkozás rövid kifejezése). Eközben a jelenlegi beszélő folytatja a megnyilatkozását, vagy rövid szünetet tart. Külön csoportba kerültek azok az egyszerre beszélések, amikor a beszédpartner megismétli a beszélő mondandójának egy részét, kifejezve figyelmét és/vagy egyetértését. Sok esetben a hallgató képes befejezni a beszélő megnyilatkozásának záró részét, amely szintén okozhat egyszerre beszélést, így ez is külön kategóriába került. Fontos mozzanat a társalgásban, amely szintén egyszerre beszélést okozhat, amikor a hallgató megkísérli a szóátvételt, de ez sikertelen. Dér külön csoportként veszi fel az előbbiek kombinációját is, nyitva hagyva ezzel az egyszerre beszélések megjelenési formáinak újabb kategóriák kialakítását is.

1.3.4. Beszélőalkalmazkodás

A társalgásban nemcsak a gondolataink vagy információk átadása történik csupán, hanem fontos interakciós jelenségek valósulnak meg, mint a másokkal való kapcsolattartás, önmagunk értelmezése stb. A kommunikációs interakció során felépítjük, létrehozzuk, reprodukáljuk, bemutatjuk identitásunkat, a szerepeinket és a kapcsolatainkat. Mindezek megjelennek mind a verbális, mind pedig nem verbális kommunikációnkban. A hétköznapi kommunikációban megfigyelhető például, hogy

másképpen formáljuk meg gondolatainkat, ha egy idős vagy egy gyermek a beszédpartner.

Ezen jelenség leírására született meg a beszéd akkomodációs elmélet (Speech Accommodation Theory: SAT) (vö. Giles 1973; Giles et al. 1991), amely a kommunikációs akkomodáció elmélet előzményeként tekinthető. A SAT azért jött létre, hogy demonstrálják a szociálpszichológiai tényezők fontosságát a beszéd dinamikájának megértésében.

A kommunikációs akkomodáció elméletet (communication accommodation theory: CAT) Howard Giles és munkatársai dolgozták ki (Giles 1973; Giles et al. 1991). Elméletük központi elgondolása, hogy az emberi interakció során a beszélő változtatja a beszédét, mimikáját, gesztusait a beszédpartner függvényében (Turner–West 2010). Azokat az okokat kívánják feltárni, amiért az egyének minimalizálják a maguk és a tárgyalópartnereik közti társadalmi különbségeket vagy hangsúlyozzák őket a szóbeli és a nonverbális kommunikáción keresztül. Ez az elmélet „a nyelv, a kontextus és az identitás" közötti kapcsolattal foglakozik (Gallois et al. 2005). A CAT-ban a nyelvi keret olyan eszköz, amely megmutatja, és összhangba hozza a társas kapcsolatokat a társadalmi konvenciókon keresztül (Giles−Coupland 1991). Ez összhangban van a társadalmi identitás elmélettel (Tajfel−Turner 1979), amelyben az egyének identitásukat a csoportban másokkal és önmagukkal folyamatosan összehasonlítják, és pozitívan vagy negatívan értékelik. Általánosságban elmondható, hogy az egyének a saját csoportjukban lévő egyéneket kedvezőbben értékelik, mint az azokon kívülieket. A CAT azt vizsgálja, hogy hogyan járulhat hozzá a nyelvi viselkedés az identitások fenntartására a verbális és nem verbális kommunikációs eszközökön keresztül a társalgási partner függvényében. A beszélő széles változatban módosíthatja a saját beszédprodukcióját (i) társadalmi tényezők mentén, amelyek jellemzik az egyént vagy a csoportot, valamint (ii) szituációs és szociális kontextus mentén, amelyben a megnyilatkozás megvalósul. Az akkomodáció tehát a kommunikációnk módjának módosítását jelenti a beszédpartner függvényében (Giles−Coupland 1991). Alapvetően kétfajta akkomodációs jelenséget írnak le: (i) közeledés: konvergencia és (ii) távolodás: divergencia. A konvergencia az a stratégia, amelyben az egyének adaptálódnak egymáshoz a kommunikációs viselkedésükben azért, hogy csökkentsék a szociális különbséget. A divergencia során a verbális és nonverbális eszközökkel kívánja az

egyéni növelni a társadalmi különbségeket. A CAT továbbá megkülönböztet három további funkciót, amellyel a beszélő adaptálódni képes a beszélgetés folyamán: fenntartás, kiegészítés, távolodás (Giles−Coupland 1991). A távolodás akkor történik, amikor a beszélő nyelvileg igyekszik eltérni a hallgatótól, hogy a társadalmi elkülönülést kifejezze. A fenntartó funkció nem jár módosítással, vagy nem mozdul el a produkcióban. A kiegészítő (complementarity) funkció olyan helyzetet ír le, amelyben sem a konvergencia, sem a divergencia nemkívánatos kommunikációs cselekvés. Giles és munkatársai (1987) kimutatták, hogy a beszélők gyakran alkalmazzák ezt a stratégiát olyan helyzetekben, amelyek hatalomkülönbséggel járnak. Ekkor a meglévő társadalmi különbségekből adódó verbális jelenségek már az elején kimutathatók, amelyek a beszélgetés folyamán nem változnak, sem nem konvergálnak, sem nem divergálnak.

A SAT és a CAT mellett igen fontos modell a Vocal Channel Social Status Model (VOCSTAT), amelyet Gregory és munkatársai dolgoztak ki (2001). Munkájukban amellett érvelnek, hogy a szóbeli kommunikációban az alaphangmagasság egy igen fontos paraméter, amely közvetíti a társas viszonyokat az egyénről, hogy milyen adaptációs tevékenységet kíván folytatni: konvergál vagy divergál.

A nemzetközi szakirodalomban a beszédpartnerhez való alkalmazkodás számos fonetikai aspektusát vizsgálták. Bell 1984-es összefoglaló munkájában ad átfogó képet a hallgatóhoz való alkalmazkodás kutatásának addigi eredményeiről. Modellje hallgatóközpontú elgondolásnak nevezhető, alapgondolata, hogy a stílus változása a beszédhez köthető személyek tényezőiből vezethető le. Tehát a beszélő igazodni igyekszik a címzetthez, sőt beszédére még azok a személyek is hatással vannak, akik nem címzettjei az üzenetnek, de valamilyen módon meghallják azt: a vendéghallgatók, akikről tudja, hogy hallják, és akik fontosak a beszélő számára. A beszédalkalmazkodás egyik legtöbbet kutatott területe a gyermekekhez szóló beszéd (vö. Sundberg 1999; Liu et al. 2003; magyarra Kátainé Koós 1998). A szóbeli szövegtípusok közül a társalgásban fokozottan jelenik meg a beszédpartnerhez való alkalmazkodás szükségessége, hiszen a társalgás interaktív nyelvi aktivitás, amelynek megvalósulásához két vagy több személy szükséges, és mindegyiküknek lehetőségük kell, hogy legyen az aktív részvételre (vö. Iványi 2001). Szerkezete sajátságos szabályokból, szabályrendszerekből és konvenciókból épül fel, együttműködést feltételező szóbeli tevékenység (vö. Grice 1975), a beszélgetés menetét nemcsak az

adott pillanatban beszélő fél határozza meg, hanem a hallgató is aktív formáló szerepet tölt be visszajelzéseivel, közbeszólásaival, a témától való eltérítésekkel, közbekérdezéssel, így a társalgás egyik legfőbb jellemzője az újraszabályozás állandó jelenléte (Boronkai 2008). A hazai akusztikai-fonetikai kutatásokban Bata és Gráczi (Bata 2009a, 2009b, Bata–Gráczi 2009) munkássága a meghatározó. Eredményeik azt mutatták, hogy nem csak az életkor volt meghatározó a beszéd temporális jegyeire és a beszédlépések alakulására, hanem az ismertség foka is.

2. A BESZÉLŐDETEKTÁLÁS MEGOLDÁSI MÓDOZATAI

2.1. Beszéddetektálás (VAD)

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, a zajos részeket stb.

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–Kondoz 2001; Gazor–Zhang 2003; Armani et al. 2003). A VAD 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 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 1993; ETSI 1999; Tanyer–Özer 2000).

A tipikus VAD 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 szabály alapúra és a statisztikai alapúra. A jellemzőkinyerés során olyan akusztikai paramétereket keresünk az akusztikai jelben, amelyek alapján elkülöníthetők a beszéd- és a nem-beszédrészek. A Shone és munkatársai (1999) által kifejlesztett VAD rendszerben egy statisztikai alapon működő algoritmust alkalmaztak, amely a jel energiáját használja akusztikai jellemzőként, az adatok eloszlásának modellezéséhez pedig Gauss-eloszlást használtak. A Ying és

munkatársai (2011) által javasolt VAD nemellenőrzött tanuláson alapuló eljárással (szekvenciális keverék Gauss-modell) automatikusan osztályozta a beszéd- és a nem-beszédrészeket, akusztikai jellemzőként pedig Mel-frekvenciás spektrális szűrőt használt. Mind a beszéd, mind a nem-beszédrészek modellezése két keverék Gauss-szal történt. Az ITU által kidolgozott G.729B VAD-szabvány egyszerre négy akusztikai paramétert is alkalmaz: a teljes és az alacsony frekvencia energiáját, a lineáris spektrális elemzést (LPC-vel) és a nullátmenet számát (zero-crossing rate).

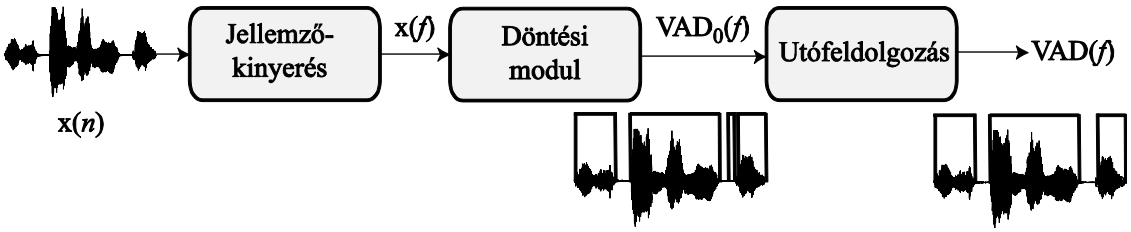
2.1.1. A VAD általános leírása

A VAD kihívása, hogy detektáljuk a jelen levő beszédjelet a zajos jelben. A VAD-ba alkalmazott döntés a bemenő jellemzővektorokra történik, ami legyen *x*. Feltételezve, hogy a beszédjel és a zaj additív, a VAD 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 (2.1. ábra), amely három fő modult tartalmaz: (i) jellemzőkinyerés, (ii) döntési modul, (iii) utófeldolgozás.



2.1. ábra

A VAD tipikus megvalósításának blokkdiagramja

2.1.2. Jellemzőkinyerés a VAD megvalósításához

A VAD létrehozásakor olyan akusztikai jellemző(ke)t szokás használni, amely diszkriminatív tulajdonsággal bír a beszéd és a nem-beszéd szegmensek automatikus osztályozásához. Számos megvalósítás alkalmazza (i) fullband (energia a teljes spektrumra) és részsáv (subband: a spektrum felbontása kisebb frekvenciatartományokra) energiát (Woo et al. 2000), (ii) spektrális eltérést a beszéd és

a háttérzaj között (Marzinzik–Kollmeier 2002), (iii) alaphangmagasságot (Tucker 1992), (iv) nullátmenetek számát (zero crossing rate) (Rabiner et al. 1975), és (v) magasabb rendű statisztikákat (Nemer et al. 2001; Ramírez et al. 2006; Górriz et al. 2006; Ramírez et al. 2007).

A legtöbb VAD 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).

2.1.3. A VAD döntési modulja

A VAD döntési modulja tartalmazza azt a szabályt vagy eljárást, amely meghatározza a jellemzővektorra (*x*-re), hogy beszéd vagy nem-beszéd. Sohn és munkatársai (1999) olyan VAD algoritmust hoztak létre, amely statisztikai eljárást, valószínűségi arány tesztet (likelihood ratio test, LRT) alkalmazott egyetlen jellemzőre. Az eljárás két-hipotézis tesztet használ, ahol az optimális döntési szabályt a valószínűségi hiba minimalizálásával éri el Bayesian 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 válasszuk $P(H_i|x)$:

$$P(H_1|x) \gtrless_{H_0}^{H_1} P(H_0|x)$$

Felhasználva a Bayesian szabályt a LRT kiszámolásához:

$$\frac{P(H_1|x)}{P(H_0|x)} >^{H_1} <^{H_0} \frac{P(H_0)}{P(H_1)}$$

2.1.4. A VAD utófeldolgozása (simítás)

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

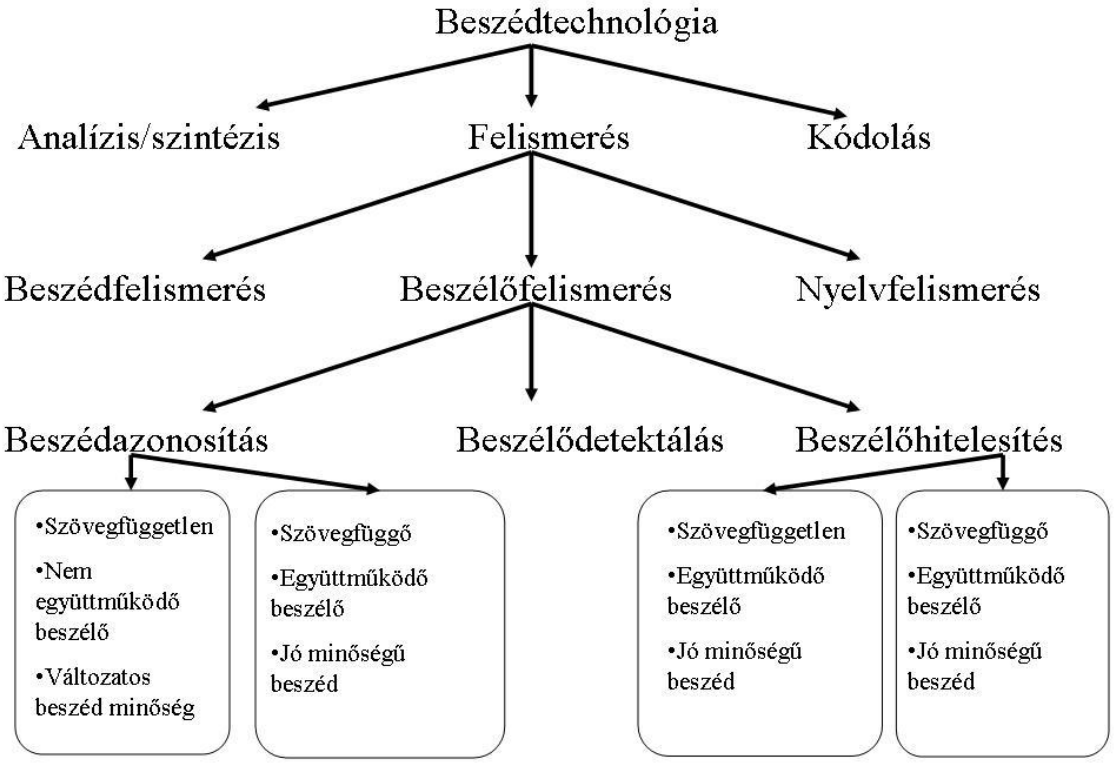
2.2. Gépi beszélőfelismerés

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 (l. INTERSPEECH 2011). 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 beszéd során (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 (Less et al. 1976), az életkorra (Ptacek–Sander, 1966; Gocsál 1998), testalkatra (Dommenlen–Moxness 1995; Gósy 2001) vagy hangulatra (Scherer et al. 2001) stb. vonatkozóan.

A gépi beszélőfelismerés alapvetően három területre osztható (Campbell 1997) (vö. 2.2. 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 alapvetően egy 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 kell, hogy 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 törekszenek, mivel ekkor tetszőleges tartalmú beszédminta alapján történhet a beszélő azonosítása vagy hitelesítése (Campbell 1997).

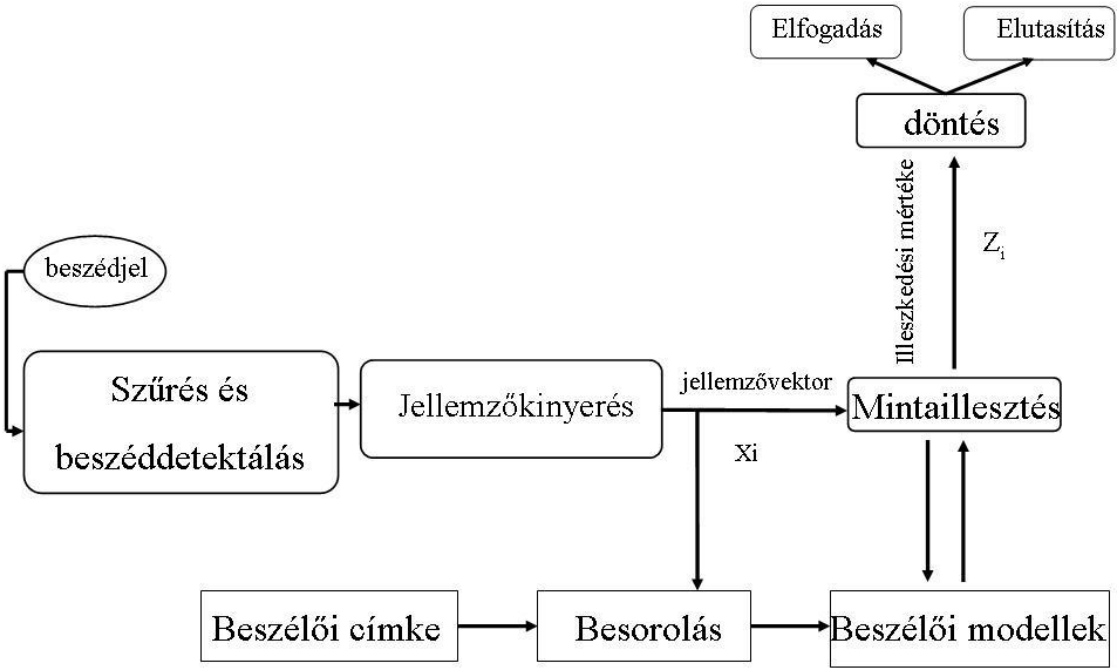


2.2. ábra

A gépi beszédfeldolgozás területei

A beszélőhitelesítés napjainkban egyre inkább megoldottnak tűnik, mivel közel 98-99%-os eredménnyel működik (Avi–Yuval 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, ami még alkalmas az azonosításhoz, 16 másodperc (Nikléczy–Gósy 2008).

A beszélőfelismerés öt lépésből áll: a beszédjel tisztítása, jellemzőkinyerés, beszélőmodellek létrehozása, mintaillesztés, döntés (2.3. ábra).



2.3. ábra

A beszélőfelismerés 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 MFC (Mel-Frequency Cepstral) 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 2500–3500 Hz közötti sáv az, amely beszélőspecifikus jegyeket hordoz (Furui 1986). 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 (GMM: Gaussian Mixture Model, HMM: Hidden Markov model, ANN: Artificial Neural Network, SVM: Support Vector Machine, 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.a és b táblázat). 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, mintaillesztéshez pedig vektorkvantálót és különböző neurális hálózatokat (MLP: Multilinear Perceptron, RBF: Radial Basis Function). Az eredményei azt mutatták, hogy a beszélő identifikációs feladatában az MLP, míg a beszélő verifikációs feladatban a vektorkvantáló teljesített jobban.

2.1.a 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%/0,5s v: 2%/1s
Markel–Davis 1979	Labor	LP	Long term statistics (egyfajta távolságfüggvény)	független	17	i: 2%/39s
Furui 1981	telefonos	Normalizált kepsztrum	Mintaillesztés	függő	10	v: 0,2%/3s
Schwartz et al. 1982	telefonos	LAR	Nemparametri-kus (pdf) valószínűség eloszlás függvény	független	21	i: 2,5%/2s
Li–Wrench 1983	labor	LP, kepsztrum	Mintaillesztés	független	11	i: 21%/3s i: 4%/10s
Doddington 1985	labor	Filter-bank	DTW: dinamikus idővetemítés	függő	200	v: 0,8%/6s
Soong et al. 1985	telefon	LP	VQ (64) Likelihood Ration Distortion	izolált szavas	100	i: 5%/1,5s i:1,5%/3s
Higgins–Wohlford 1986	Labor	kepsztrum	DTW Likelihood scoring	független	11	v:10%/2,5s v:4,5%/10s
Attili et al. 1988	Labor	Kepsztrum LP Autocorr	Projected Long term statistics	függő	90	v: 1%/3s
Higgins et al. 1991	irodai	LAR, LP-kepsztrum	DTW, Likelihood Scoring	függő	186	v:1,7%/10s
Tishby 1991	telefonos	LP	HMM (AR mix)	10 isolated digits	100	v: 2,8%/1,5s v: 0,8%/3,5s
Reynolds 1995; Reynolds–Carlson 1995	irodai	Mel-kepsztrum	HMM (GMM)	függő	138	i: 0,8%/10s v:0,12%/10s

2.1.b 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**
Che–Lin 1995	irodai	kepsztrum	HMM	függő	138	i:0,56%/2,5s i:0,14%/10s v:0,62%/2,5s
Colombi et al. 1996	irodai	Kepsztrum, Energia Első két derivált	HMM (GMM)	független	416	v: 11%/3s v: 6%/10s v: 3%/30s
Lindblom 2003	telefonos	MFCC Első két derivált	GMM-SVM	független	30	v:16,5%
Campbell et al 2006	telefonos	LPCC MFCC	SVM (GLDS kernel) GMM GMM-SVM	független	356	v: 6,1%/30s v: 4,8%/30s v: 3,2%/30s
Joshi et al. 2008	telefonos	LPCC	AANN-HMM	független	500	v: 6,69%
Avi–Yuval 2011	labor	LSF and MFCC Első két derivált	skew-GMM	független	100	i: 1,5

2.3. 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 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 beszédfelismerőt 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).

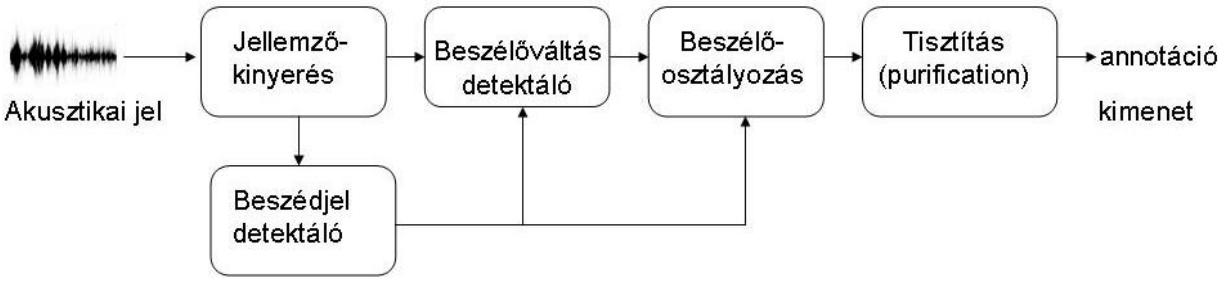
Az egyszerre beszélések aránya a spontán társalgásokban meglehetősen nagynak mondható (Gráczi–Bata 2010). Beattie a beszélőváltásokat elemezve (1983, idézi Levelt 1989) kimutatta, hogy a két résztvevős angol társalgásban 11%-ban fordul elő egyszerre beszélés (azaz a beszédpartner közbevág), több beszélőnél ez az arány már 31%. Az újabb kutatások ezeket az arányokat igazolták. Cetin és Shriberg (2006) 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 (2009) 1,7-3%-ot adatolt kutatásában, ahol 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 (pl. *igen*, *aha*, *ühüm*, vö. backchannel; lásd 1.3.3. fejezet), másrészt versengő funkciójú, amikor a társalgás

egyik szereplője át kívánja venni a szót, és már az alatt elkezdi a beszédét, mialatt az aktuálisan beszélő még nem fejezte be a mondanivalóját (Iványi 2001; Hámori 2006; Bata 2009).

Az egyszerre beszélések vizsgálata sokrétű (Çetin–Shriberg 2006). 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 beszédek 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.

2.4. Beszélődetektálás

A beszélődetektálás során a folyamatos társalgás beszédfordulóit automatikusan detektáljuk, majd az így kialakított beszédrészeket hozzárendeljük a beszélgetésben részvevő 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 a beszélők szerint, azaz az egyes beszélőkhöz rendelni. Egy általános felépítésű beszéddetektáló rendszer felépítése az 2.4. ábrán látható.



2.4. ábra

A beszélőosztályozó leegyszerűsített blokkdiagramja

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

A beszélődetektálás megvalósítására számos megoldás készült számos nyelvben. 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, amelyben a beszédet különböző beszélők törik meg (Raynold–Torres-Carrasquillo 2004). A beszélődetektálás fő kérdése, hogy „Ki mikor beszél?”, amely sok esetben referál a beszélő szerinti szegmentálásra és a klaszterezésre.

Raynold and Torres-Carrasquillo (2004) szerint a beszélődetektálásnak három fő alkalmazási területe van, amelyre az utóbbi években kiemelt figyelem került:

(i) Híradások (broadcast news): rádió és tv csatornák 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 felvett.

(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 a feladat, hogy az egyes (diszkrét) kategóriák legyenek megfeleltetve a folyamatos beszédjellel (időben illesztettek legyenek a beszédjelhez), és ezáltal a köztük lévő határok definiáltak 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, 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, adatfeldolgozás, mintafelismerés/-osztályozás stb.

Ahhoz, hogy beszédben meghatározzuk, hogy ki beszél a hangfelvételben (osztályozási technika alkalmazása), meg kell határoznunk először azokat a szegmenseket a hanganyagban, amelyeket klaszterezni fogunk, és amelyek különböző hosszúak és különböző akusztikai karakterrel rendelkezhetnek (beszéd, nem beszéd, 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 (Raynold–Torres-Carrasquillo 2004). A megfelelő algoritmus megválasztása azonban nem olyan egyértelmű. Gyakori, hogy a különféle osztályozó algoritmusok szignifikánsan különböző osztályozási eredményt adnak.

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, amelyeket a beszélőszegmentálásban, illetve a beszélőklaszterezésben használatosak. A legtöbb beszéddetektálóban a beszélőszegmentálás az első lépés, ezért először ezt, majd másodikként a beszélőklaszterezést mutatjuk be.

2.4.1. 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; Hermansky 1990); ezek kimenete a legtöbb esetben 10-20 együtthatóból álló paramétervektor. Ezen akusztikai jellemzőket más beszédtechnológiai rendszerekben, például beszédfelismerésben is szokás használni. 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. 2006).

Az alaphangmagasság modellezését a beszélőfelismeréshez Soenmez és munkatársai (1998) mutatták be tanulmányukban. A beszélő alaphangmozgását úgy modellezték, hogy az aktuális szakaszban történő alaphangmozgást lineáris modellel közelítették. Az így stilizált modell paramétereit használták fel a beszélő azonosításához. Janin és munkatársai (2003) Soenmez és munkatársai modellezési technikáját folytatva a lineárisan közelített stilizált alaphang- és energiamozgásból határozták meg, hogy az eső vagy emelkedő, így egy bigram modellt építettek. Számos tanulmány használta a szupraszegmentális eszközök közül az időtartamot, az alaphangmagasságot és az energiát, amelyet két szünet közötti egységen mértek ki (Kajarekar et al. 2004). Számos kutatásban ugyanezen akusztikai jellemzőket vizsgálták hosszabb időszakaszokon (a statisztikákat egy teljes beszélgetésre számolták, ahol a célszemély és az imposztor közötti távolságot minden egyes beszélgetésre mért jellemzővektor között log-likelihood módszerrel számolták (Peskin et al. 2003; Reynolds et al. 2003). Szótagokra kinyert akusztikai jellemzőket is szokás alkalmazni, amelynek előnye az, hogy nagymennyiségű mintát kapunk. A szótagokat ebben az esetben automatikusan, a beszédfelismerő kimenetéként kapják. Az akusztikai jellemzőket (alaphangmagasság, energia, időtartam) minden egyes szótagra kiszámítják, majd GMM-el (Gaussian Mixture Model, Gauss-keverék modell) vagy SVM-mel (Suppor Vector Machine, Szupport vektor gép) modellezik azokat (Shriberg et al 2005; Ferrer et al. 2007).

Friedland és munkatársai (2009) a rövid idejű spektrális jellemzők mellett az alaphangmagasságot és más hosszú idejű (hosszabb időkeretre számolt) spektrális jellemzőket használtak.

Számos kutatásban számoltak be olyan alternatív akusztikai paraméterről, amely kifejezetten a beszélő karakterisztikáját igyekezett reprezentálni, és amely kifejezetten a beszélő modellezésére alkalmazható önmagában, vagy a standard akusztikai jellemzőkkel együtt. Ezek használata még nem elterjedt, de feltehetően ezen akusztikai jellemzők lesznek a legalkalmazottabbak az elkövetkezendő időkben (Anguera 2006).

Yamaguchi, Yamashita és Matsunaga (2005) például beszélőszegmentáló rendszerükben az energiát, az alaphangmagasságot, a frekvenciacsúcs középpontját, sávszélességét, és még három új jellemzőt használtak: az energiaspektrum temporális stabilitása, a spektrális burkoló görbe alakja, és ezen jellemzők keresztkorrelációja az energiaspektrummal.

Nguyen (2003) egy új nem lineáris jellemző normalizációs eljárást (SWAMP: Sweeping Metric Parameterization) javasol a háttérzaj, illetve 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, Benetos és Kotropoulos (2006) az MPEG-7 alapú akusztikai jellemzők, mint például az AudioWaveformEnvelop és az AudioSpectrumCentroid, mellet érvelnek. Ez a két akusztikai jellemző az MPEG-7 Audio standard csomag részei (Manjunath et al. 2002). Az AudioWaveformEnvelop jellemző néhány értékkel reprezentálja az extrém adatokat (minimum és maximum) beszéd hullámformájából. Az AudioSpectrumCentroid jellemző pedig a spektrum log-frekvenciás energiaspektrum súlyközéppontját (CoG: center of gravity) határozza meg.

2.4.2. Beszélőszegmentálás

A beszélőszegmentálás sok esetben a beszélőváltás-detektálással hasonlatos, és igen közel áll a beszéd/nembeszéd-detektálásához. A beszéd/audiojel, beszélőszegmentálást/váltást detektáló rendszer az folyamatos akusztikai jelben megtalálja, hogy hol van beszélőváltás. Általánosabban az akusztikai változás detektálás alapvető célja megtalálni azt az időpillanatot, ahol az akusztikai jelben változás történik

a felvétel során, amely lehet beszéd/nem-beszéd, zene/beszéd és egyéb más kategóriák. A jelen esetben az akusztikaiváltozás-detektáló feladata a beszédlépések megtalálása.

Sok esetben tévesen a beszélőszegmentálás kifejezés egyszerre jelenti a beszélőváltások megtalálását, illetve ezen részek homogén csoportokba való klaszterezését. A beszélőszegmentálást és klaszterezést fontos megkülönböztetni, mivel alapvetően teljesen két különböző feladatról van szó, ugyanis a beszélőszegmentálás alapvető feladata, hogy megtalálja azon időpillanatokat, amikor beszédlépés történik, míg a beszélőklaszterezéskor ezeket a beszédlépéseket csoportosítjuk, vagyis hozzárendeljük az egyes beszélőkhöz.

A beszélőszegmentálás megvalósítására alapvetően két megoldási technika létezik a szakirodalom alapján. Az első megoldásban az akusztikai adatok alapján egyetlen lépésben határozza meg a váltási pontokat (vö. Kim et al. 2005). A második megoldásban ezt több lépésben teszi, úgy hogy iteratívan pontosítja a kimenetet (vö. sian Cheng–min Wang 2004). Az első lépésben több váltási pontot feltételez a rendszer; többet, mint amennyi valójában létezik, amely magas téves elfogadási hibát (false alarm rate) eredményez. A második lépésben ezeket iteratívan felülvizsgálja az algoritmus, és törli azokat, amelyek nem szükségesek.

Egy másik megközelítésben a beszélőszegmentálás három főbb kategóriába lehet sorolni (Shaobing Chen–Gopalakrishnan 1998; Kemp–Schmidt–Westphal–Waibel 2000; Chen–Gales–Gopinath–Kanvesky–Olsen 2002; Ajmera 2004; Perez-Freire–Garcia-Mateo 2004): metrikus alapú, szünetalapú, modellalapú algoritmusok.

2.4.2.1. Metrikus alapú szegmentáló algoritmusok

A metrikus alapú szegmentáló algoritmusok a legtöbbet használt eljárások. Az algoritmus alapja, hogy valamilyen távolságot mér az akusztikai szegmensek között, és megállapítja, hogy vajon az előző beszélőhöz tartozik-e, vagyis hogy beszédlépés 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, amit az akusztikaiváltozás-detektálásra használnak, alkalmazható a beszélőklaszterezésre is annak megállapítására, hogy a két beszélői csoport azonos beszélőhöz tartozik-e.

Legyen két audioszegmens *(i, j)*, amelyeket az akusztikai jellemzővektorukkal reprezentálunk X_i és X_j, és amelyek hossza N_i és N_j. Ezek átlaga és varianciája μ_i, σ_i és μ_j, σ_j. Mindegyik szegmenst modellezzük Gauss-eloszlással: $M_i(\mu_i, \sigma_i)$ és $M_j(\mu_j, \sigma_j)$, amely lehet egy Gaussos vagy több Gaussos. Másrészről a két szegmenst összevonva X, az átlag és a variancia μ, σ, amelyet a Gauss eloszlással közelítve $M(\mu, \sigma)$.

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 egy Gaussos átlaggal és a varianciával.

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

2.4.2.1.1. Bayesian Information Criterion (BIC: Bayes-féle Információs Kritérium)

A BIC az egyik legtöbbet használt algoritmus a szegmentálásban, illetve a klaszterezésben, mivel számítása igen egyszerű és hatékony. 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 modell komplexitást (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, és amely a következőképpen definiálható:

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

Mivel a $logL(X_i,M_i)$ az adatok log-likelihood értéke (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 az $\#(M_i)$ 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-t 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:

$$\Delta \mathrm{BIC}(\mathrm{i},\mathrm{j}) = -\mathrm{R}(\mathrm{i},\mathrm{j}) + \lambda \mathrm{P}$$

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

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

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

$$\mathrm{P} = \frac{1}{2}(\mathrm{p} + \frac{1}{2}\mathrm{p}(\mathrm{p}+1))\log(\mathrm{N})$$

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 $\Delta\#(\mathrm{i},\mathrm{j})$ a különbség értéke a szabad paraméterekben a kombinált modell és a két különálló modell között.

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 BI- kritériumként hivatkozni. A BIC-algoritmust elsőként Chen és Gopalakrishnan (1998) 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–Kryze–Wellekens 1999a; Lopez–Ellis 2000; Delacourt–Wellekens 2000; Mori–Nakagawa 2001; Vandecatseye et al. 2004).

Ajmera, McCowan és Bourlard (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 használatát, 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, amely azt feltételezi, hogy a 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 Shaobing Chen és Gopalakrishnan (1998) több váltási pontot feltételező kétutas algoritmust alkalmaztak, később számos tanulmány (Tritschler–Gopinath 1999; Sivakumaran–Fortuna–Ariyaeeinia 2001; Lu–Zhang 2002a; Cettolo–Vescovi 2003; sian Cheng–min Wang 2003; Vescovi–Cettolo–Rizzi 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 detektáláshoz.

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 átlagának és varianciájának kiszámítására. Roch és Cheng (2004) a MAP (Maximum A Posteriori) adaptációs algoritmust használtak a paraméterbecsléshez. Emellett számos tanulmány az ML (Maximum Likelihood) algoritmust használja a paraméterek becsléséhez (Miró 2006).

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. 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 használják a BIC-et, 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 T2 távolság használatát javasolják, míg Lu és Zhang (2002) KL2 (Kullback-Leibler) távolságot. Vandecatseye és munkatársai (2004) normalizált GLR-t (NLLR) használtak az előszegmentáláshoz, míg normalizált BIC-et az utószegmentáláshoz.

2.4.2.1.2. Generalized Likelihood Ratio (GLR: általánosított valószínűségarány)

A GLR-t mint változás detektáló algoritmust először Willsky és Jones (1976), illetve Appel és Brandt (1982) mutatta be. A GLR szintén valószerűségi alapú metrikai eljárás, amely két hipotézis közötti arányt fejez ki: a H_0 mindkét szegmens azonos személytől 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. Az hasonlóság aránya tulajdonképpen a valószerűség arányaként számolható a két hipotézis között:

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

és ebből meghatározva a két szegmens közötti távolságot $D(i, j) = -log(GLR(i, j))$, így egy megfelelő küszöbértéket megválasztva el lehet dönteni, hogy a két szegmens azonos beszélőtől származik vagy nem. 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 megbecsülni, míg a LLR-ben a modelleknek a priori ismertnek kell lenniük.

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

Bonastre és munkatársai (2000) a GLR egyutas megoldásként alkalmazta a szegmensekre, hogy a beszélőváltásokat előre jelezze. A küszöbértéket úgy állította be, hogy a tévesztési arányt minimalizálja (gyakoribb téves riasztások árán). A beszélődetektálójában minden egyes szegmenst önálló potenciális beszélőhöz tartozónak tekintett.

Gangadharaiah és munkatársai (2004) két beszélőre alkalmas szegmentálót fejlesztettek, amely kétutas szegmentáló volt. Az első lépésben GLR-t, míg a másodikban Viterbi-algoritmust használt 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 választja ki, amelyek az elsőben egyik beszélőhöz sem tartoztak, és összehasonlítja a két beszélői modell GLR értékeivel, és ahhoz a beszélőhöz rendeli a szegmenst, amelyiknél magasabb a hasonlósági mérték.

A váltásdetektálásában és indexelésben Liu és Kubala (1999) büntető GLR-eljárást alkalmazott, mint a szegmentáló második lépése. A váltási pont elfogadására/elutasítására egy előre tanított beszédhangalapú dekódert használtak (ahol az ASR beszédhangkészlet beszédhangcsoportokra volt klaszterezve). A büntetés, amelyet 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 a θ empirikusan meghatározott. Hasonló megfogalmazásban használta Metze és munkatársai (2004) a GLR-t szegmentációs lépésként a társalgást átiratozó rendszerében.

Az egyik legtöbbet használt beszélőszegmentáló algoritmus, amelyben a GLR-t használják a DISTBIC-algoritmus (Delacourt–Wellekens 1999; Delacourt et al. 1999a; Delacourt et al. 1999b, Delacourt–Wellekens 2000), ahol a GLR-t a szegmentálás első lépéseként alkalmazzák, majd BIC-et ennek a szegmentálásnak a finomítására.

2.4.2.1.3. Gish-distance (Gish-távolság)

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

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

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

2.4.2.1.4. 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 eltérését) a következőképpen tudjuk számolni:

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

ahol E_X várható érték tekintettel a X valószínűségi eloszlásfüggvényére. Ha 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 nemnegatív, de nem valódi metrika, mivel nem szimmetrikus, azaz megkülönböztetheti a modellt és modellezett eloszlást. A KL2-távolságot szimmetrikussá lehet tenni a következő lépéssel:

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

Delacourt és Wellekens (2000) a KL2-távolságot használta első két lépésként a beszélőváltási pont meghatározására. Zochova és Radova (2005) a KL2 egy továbbfejlesztett változatást használták. Elsőként a szüneteket és a lélegzetvételeket szűrték ki a beszédből rövid idejű spektrális energiával, illetve ZCR-t (zero-crossing rate, nullátmenetek aránya) használva.

Hung és munkatársai (2000) MFCC akusztikai jellemzőt használtak az egyes szegmensekre, és a hasonlóságot a két szegmens között KL2-vel és Mahalanobis és Bhattacharyya távolsággal mérték a beszélőváltás megállapításához.

2.4.2.1.5. Divergence Shape Distance (DSD)

A DSD-algoritmus nagyon közel áll a Gish-távolsághoz. A DSD a két osztály KL-távolságából származik, és *n*-számú normális valószínűségi eloszlásfüggvénnyel kívánja kiküszöbölni az átlag által meghatározott részt, amelyet könnyen befolyásolnak környezeti tényezők:

$$D(i,j) = \frac{1}{2}tr\left[\left(C_i - C_j\right)\left(C_j^{-1} - C_i^{-1}\right)\right]$$

Ezt a távolságmértéket használták egyutas (Kim et al. 2005), illetve kétutas beszélőszegmentáláshoz is (Lu–Zhang 2002a,b).

2.4.2.1.6. Cross-BIC (XBIC)

A XBIC a hasonlósági mértéknek az alapja a két szomszédos szegmens közötti kereszt-valószerűség (cross-likelihood) értékelése (Anguera–Hernando 2004; Anguera 2005). BIC-től inspirálva Juang és Rabiner (1985) a HMM-ek (hidden Markov-model, rejtett Markov-modell) közötti távolság mérésére használta:

$$XBIC(X_1, X_2) = L(X_1, M_2(\mu_2, \sigma_2))L(X_2, M_1(\mu_1, \sigma_1))$$

2.4.2.1.7. Más távolságmérési eljárások

A fent említett eljárásokon kívül még számos más technika van két szegmens összehasonlítására. A CuSum-távolság (Basseville–Nikiforov 1993); a Kolmogorov-Smirnov teszt (Deshayes–Picard 1986); Mahalanobis–Bhattacharyya-távolság (Campbell 1997); VQ (Vector Quantization, vektorkvantáló) algoritmus (Mori–Nakagawa 2001); Hotelling's T^2-távolság (Zhou–Hansen 2000; Tranter–Reynlods 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, b) 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).

2.4.2.2. Nem metrikán alapuló szegmentálók

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

2.4.2.2.1. Szünetalapú beszélőszegmentáló

Az ezen technikán alapuló beszélőváltást előrejelző eljárások azt feltételezik, hogy a beszédváltás előtt, vagyis a két beszélő mintái között szünet helyezkedik el. Ezek többsége a beszédfelismerő rendszereknek adják át a beszélőktől származó beszédszeleteket, így nagyon fontos, hogy egy-egy beszédrész ne tartalmazzon félbevágott szót, vagyis a beszédváltás átfedő beszéd nélkül jöjjön létre. Ebbe a technikai megoldásba tartoznak az energiaalapú (energy-based), illetve a dekóderalapú (decoder-based) eljárások.

Az energiaalapú eljárások általában valamilyen energia-szintkövetést (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 használta, amely az energia változatozását 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 előre definiált a szünetek minimális hossza, hogy csökkentsék a téves elutasítások számát. Nyilvánvalóan belátható, hogy a szünetek jelenléte és a beszélőváltás csak csekély mértékben korrelálnak egymással, így ezen rendszerekben általában csak hipotetikusan megjelölt szünetváltási helyeket feltételeznek, és egyéb algoritmusokkal egyértelműsítik azt.

2.4.2.2.2. Modellalapú szegmentáló

A modell alapú 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 (Maximum Liklehood) algoritmussal 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 megkülönbözetni. Ez a szegmentálási technika igen közel áll a beszélőklaszterező megoldásokhoz, amelyekben a különböző beszélők identitása (vagyis az 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 hozzanak létre, amelyben nem szükséges előzetes információ például a társalgásban részt vevő beszélők számáról. 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 (2004a) 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) SVMs (Support Vector Machines, Szupport vektor gépek) használt az akusztikai osztályok modellezéséhez, és ML-t dekódoláshoz.

2.4.2.3. 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 nemellenőrzötten működő rendszer. Emellett a bonyolultabb megoldások nem, vagy csak alig javítottak a beszélődetektálás eredményein. Ezért a jelen disszertációban mi is e mellett az algoritmus mellett döntöttünk.

2.4.3. Beszélőklaszterezés

A beszélőklaszterezés azon technikák és algoritmusok összességét jelenti, amelyekkel az egyes beszédszegmenseket homogén csoportokba, egy-egy beszélőhöz rendeli. 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. 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.

A szakirodalomban általában kétféle főbb alkalmazását találjuk a beszélődetektálásnak. Az egyik a beszédfelismerő rendszer, ahol homogén beszélői modellekhez adaptálják az akusztikai modelleket azért, hogy a rendszer beszélőfüggő legyen, növelve ezzel a felismerés eredményességét. A beszélőre adaptált modelleket a beszédfelismerő rendszer a beszélődetektáló kimenete függvényében aktiválja. A másik alkalmazás a beszélőindexelés (speaker indexing) és az információgazdag átirat (RT: rich transcription), amelyek a beszélődetektáló kimenetét használják fel, hogy számtalan információt nyerjenek ki a hangfelvételből, amely aztán később automatikusan indexálható, vagy más műveletekre alkalmazható.

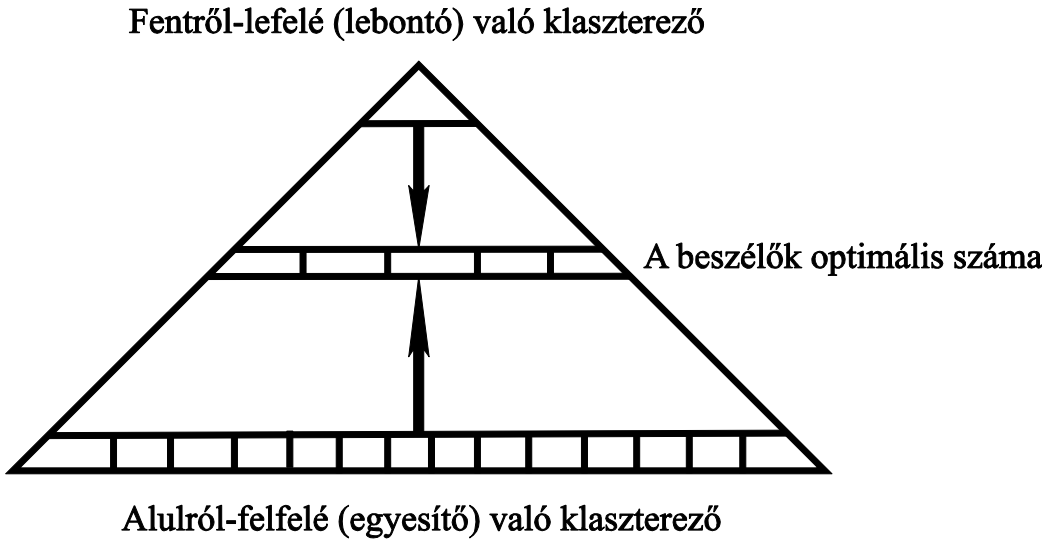
A beszélődetektálásnál is fontos megkülönböztetni az online és az offline rendszereket. Az offline rendszerben az adatok a rendszer működése előtt rendelkezésre állnak. A legtöbb kutatás ilyen rendszerekről számol be. Az online rendszerek esetében az adatok a felvétel során dolgozódnak fel. A legtöbb 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, mindezt egy dinamikus küszöbérték alkalmazásával. Ebben az esetben a hangsúly a beszélői szegmensek gyors beszélők szerinti osztályozásán van, amelyre döntési fát használtak a beszélői modellek kialakításához.

A legtöbb beszélődetektáló azonban offline rendszer, amelyben az algoritmusokat úgy mutatják be, mint egy online is működtethető rendszer. Ezen rendszereket alapvetően két csoportba lehet sorolni. Az egyikbe tartoznak a hierarchikus klaszterezési technikák (hierarchical clustering technique), amelyek az optimális beszélői számot keresik iteratív módon a különböző lehetséges klasztereknek a

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.

2.4.3.1. 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 (2.5. ábra).



2.5. 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 legjobban megoldani.

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

Ez a leggyakrabban használatos módszer a beszélőklaszterezésre, mert a beszélőszegmentálás technikája felhasználható a klaszter 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 alkalmazták, amelyben Gish-távolságot használtak mint távolságmátrixot, súlyokkal alkalmazva a közeli klaszterek összevonásához. Megállási feltételként a büntetősúlyt minimalizáló függvényt alkalmaztak: 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 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 kritériumot 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 mérésre.

Általában elmondható, hogy a statisztikai alapú távolsági metrikák (nem igényelnek tanítást) korlátozottan működnek a beszélőklaszterezésben, mivel impliciten határozzák meg a távolságot az átlag- 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 és M_2 és K_2 Gauss-

keverék mindegyik modellre, és a Gauss-súlyok $W_1(i)$, $i = 1 \ldots K_1$ 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 a $N(i)$ modellből származó egyik Gauss-komponens.

Begin és munkatársai (1998) az egyes Gauss-keverékek között mérte a távolságot. A távolságmátrixot $d(i,j), \forall i,j$ a két modell minden lehetséges Gauss párja között meghatározza (a távolságot euklideszi, Mahalanobis- és KL-függvénnyel méri), 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 az átlagok adaptáltak (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 keverékek száma és GMM-modell dimenziója.

Ben és munkatársai (2004) küszöbértéket használtak a távolságra, mint egyfajta megállási kritérium.

A statisztikai alapú rendszereken kívül Gauvain és munkatársai (1988), valamint Barras és munkatársai (2004) a GLR-metrikát mutattak be (hangoló paraméterekkel), ahol büntették a nagyszámú szegmenseket és 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 addig, amíg maximalizálva nincs az ún. becsült klaszter tisztasága (purity).

A legtöbbet alkalmazott távolsági és megállási kritérium a beszélőklaszterezés esetében is a BIC-algoritmus (Shaobing et al. 1998; Chen–Gopalakrishnan 1998). A páronkénti távolságmátrix minden egyes iteráció során kiszámításra kerül, és az a pár, amelyik a legnagyobb BIC-értékkel rendelkezik, összevonásra kerül. Ez a folyamat akkor kerül leállításra, 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 entrópia távolságát használták a beszélőklaszterezéshez, amelyet az ASR-rendszerben alkalmaztak a beszélőadaptáció megvalósításához. A távolság hasonló Anguera (2005) és Malegaonkar és munkatársai által használtéhoz, amelyeket a beszélőszegmentációhoz alkalmaztak. A következő egyenlettel kifejezve:

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

ahol D_{λ_i,λ_j} következőt jelenti:

$$D_{\lambda_i,\lambda_j} = \log \mathrm{p}(\mathrm{X_i}|\mathrm{M_i}) - \log \mathrm{p}\left(\mathrm{X_i}\middle|\mathrm{M_j}\right)$$

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 beszélőfelismerés területéről érkeztek (Barras et al 2004; Sinha 2005; Zhu et al. 2005; Zhu et al. 2006). A rendszerben standard összevonáson alapuló (agglomeratív) klaszterezést használtak BIC-val. A büntetőparamétert, 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 háttér modell, UBM: universal background model) és MAP-adaptálással alakították ki beszélői modelleket (Wu et al. 2003a, 2003b, 2003c). A legtöbb esetben lokális jellemzővetemítő normalizációs (local feature warping normalization) algoritmust használtak, hogy a jellemzőkből eltávolítsák a nem stacionárius részeket. A beszélői modelleket metrikusan hasonlítják össze

kereszt-likelihood algoritmussal (Reynolds 1998), amelyet a következőképpen lehet megfogalmazni:

$$D(X_1, X_2) = \frac{1}{N_1}\log\frac{\mathrm{p}(\mathrm{X}_1|\mathrm{M}_2 - \mathrm{UBM})}{\mathrm{p}(\mathrm{X}_1|\mathrm{UBM})} + \frac{1}{N_2}\log\frac{\mathrm{p}(\mathrm{X}_2|\mathrm{M}_1 - \mathrm{UBM})}{\mathrm{p}(\mathrm{X}_2|\mathrm{UBM})}$$

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) bűntető GLR-algoritmust hoztak létre, ezen belül tradicionális 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 koordinátarendszerben helyezik el, és ebben a koordinátarendszerben 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 keresztkorreláció 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.

2.4.3.1.2. Fentről lefelé (lebontó) 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–Raynolds 2004). Az algoritmus addig fut iteratívan, ameddig négy alcsoportot nem képez, majd ezután összevonja azokat, amelyek nagyon hasonlók egymáshoz. Johnson és Woodland (1998) két különböző algoritmus implementációját javasolták. Egyrészről az MLLR (Maximum Linear Logistic Regression) adaptációs lépést, másrészről az Arithmetic Harmonic Speericity eljárást ahhoz, hogy az egyes beszélői szegmenseket hozzárendeljék a megfelelő alcsoportokhoz. Az Arithmetic Harmonic Speericity-t egy gaussos modellként lehet definiálni:

$$d(X_1, X_2 = \log\left[tr\left(\sum_{X_2}\sum_{X_1}^{-1}\right).tr\left(\sum_{X_1}\sum_{X_2}^{-1}\right)\right] - 2\log(D)$$

ahol a *D* az adatok dimenziója.

3. AZ ÉRTEKEZÉS CÉLJA, KUTATÁSI KÉRDÉSEK ÉS HIPOTÉZISEK

3.1. Az értekezés céljai

A disszertáció fő célja, hogy nagy mennyiségű magyar nyelvű spontán társalgás felhasználásával elsőként hozzon létre egy nemellenőrzött tanuláson alapuló beszélődetektáló algoritmust. A dolgozat további 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). A dolgozat célja továbbá az volt, (ii) hogy vizsgáljuk, milyen sikerrel lehet implementálni a beszélődetektálóba a VAD és az egyszerre beszélést detektáló algoritmusokat. További célunk volt az is, (iii) hogy a beszélődetektálóban milyen akusztikai paraméterekkel lehet a legjobb eredményt elérni. Mindezen algoritmusokat a MATLAB 2011a verziójú szoftverben írtuk, és futtattuk.

3.2. 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 a képi feldolgozásban használt mély neuronhálókat alkalmazni az egyszerrebeszélés-detektáló jellemzőkinyeréseként. Elemeztük, hogy a beszélőszegmentálásban milyen beállítások mellett kapjuk a legjobb eredmény.

3.3. A kutatás hipotézisei

A kutatáshoz a következő **hipotézisek**et fogalmaztunk meg:

1. A magyar nyelvű spontán társalgásban jó minőséggel lehet detektálni a beszélőváltásokat pusztán nemellenőrzött módszerekkel.
2. A beszédfelismerésben a spektrumban célzott részsávjára történő akusztikai jellemzőkinyerés jobb eredményeket adhat, mint a teljes spektrumot feldolgozó eljárások.
3. A beszélődetektálásban kikísérletezett akusztikai jellemzők jól alkalmazhatók a beszélőszegmentálásban, illetve a beszélőklaszterezésben.
4. Az egyszerre beszélések detektálásában jól lehet alkalmazni a mély neurális hálózatokat (DBN) mint az akusztikai jellemezők reprezentációját.
5. Az egyszerre beszélések és a beszéddetektálás implementációjával a beszélődetektálás eredményei növelhetők.

4. A KUTATÁS MÓDSZERTANA

4.1. Beszédanyag, kísérleti személyek

Ebben a fejezetben a disszertáció alapjául szolgáló adatbázist írjuk le. Mivel azonban az egyes részfejezetek különböző részeit használják a korpusznak, ezért a specifikusan ahhoz tartozó adatokat az aktuális fejezet *A vizsgálat anyaga* című részében közöljük.

A disszertációban a BEA-adatbázisból (Gósy 2012) 100 társalgást választottunk ki, amely 55 órányi hanganyagot jelent. A társalgásokban minden esetben három személy vett rész. Ebből két társalgó állandó volt (2 nő, életkoruk 32 év). A harmadik személy 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 KHz-en, amelyet újramintavételeztünk 16 kHz-en. A BEA-korpusz 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 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 az artikulációból adódó néma fázisokat nem jelöltük még akkor sem, ha azok ezen küszöböt átlépték is.

(ii) Beszélőváltások: folyamatos jelben bejelöltük, hogy mely időpillanatban van beszélőváltás, illetve hogy az egyes beszédszegmensek mely beszélőhöz tartoznak. A háttércsatorna-jelzéseket nem vettük beszélőváltásnak, csak abban az esetben, ha tényleges szóátvételről volt szó.

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

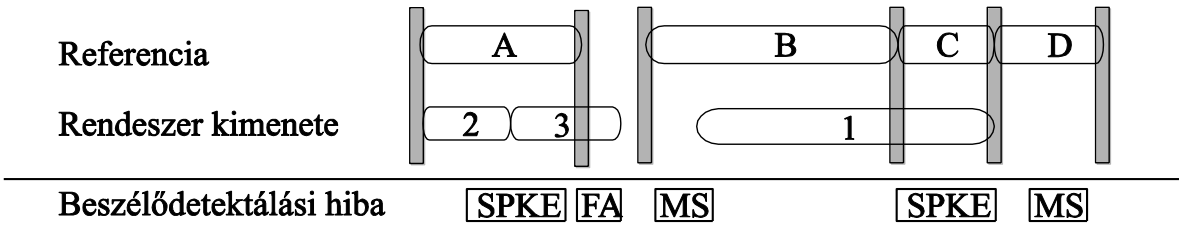
4.2. Kiértékelési módszer

A jelen dolgozatban kétféle kiértékelő rendszert alkalmaztunk. A beszéddetektálás, beszélőszegmentálás és a beszélőklaszterezés kiértékeléséhez a NIST által javasolt DER (beszélődetektálási hibaarány, Diarization Error Rate) módszert használtunk. Az egyszerre beszélés kiértékeléséhez pedig a kétosztályos kiértékelési metrikák számoltuk ki, 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, amelyet a NIST az RT kiértékelésekor alkalmazott (NIST Fall Rich Transcription 2006). A DER-t tulajdonképpen úgy értelmezzük, mint azt a törési időt, amely nem tulajdonítható helyesen sem a beszélőnek vagy a nem beszédnek. Ennek mérésére az MD-eval-v12.pl-t (NIST MD-eval-v12 DER kiértékelő scriptje 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 bementi (kézi) címkében és a kimeneti (automatikus) címkében. Ez a feladat tehát nem olyan, mint a beszéd/nem beszéd automatikus címkézése, amely során a szegmenst azonosító címkének egyezni kell a bementi és a kimeneti címkében (4.1. ábra).



4.1. ábra

A DER kiértékelési metódus sematikus ábrázolása

A kiértékelő script 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, ahol mind a hipotetikus, mind a referenciacímke tartalmazza ugyanazt a beszélői párt. Ezt úgy kapjuk meg, hogy összehasonlítjuk a hipotetikus, illetve a referencia-beszédfordulókat. A N_{ref} és N_{sys} 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-hiba tulajdonképpen különböző módon létrejött hibák összege:

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

2) Téves riasztások száma: teljes időtartamra vetítve a referenciacímkében a nem beszéd 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) Téves elutasítások száma: a teljes időtartamra vetítve a referenciacímkében a beszélő szerepel, de az automatikus címkesorban nem beszédnek 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: 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 függ attól, 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 ezt az időbeli határsávot ± 250 ms-ban határozta meg. A NIST DER script kiértékelő megadja minden egyes referencia-hipotetikus szegmentációra a DER értékét, illetve az összes kiértékeléshez használt fájlra ad egy súlyozott átlagot.

4.2.2. További kiértékelési technikák (DET: detection error tradeoff)

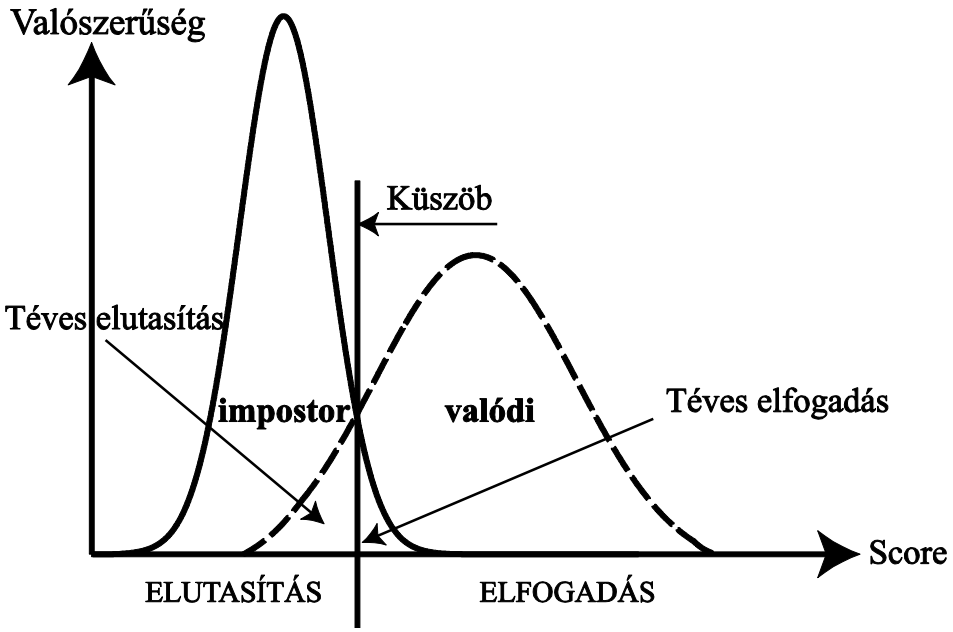
Az osztályozásra alkalmazott algoritmusok működésének kiértékelésére és összehasonlítására a DET (Detection Error Tradeoff) kiértékelő algoritmust használtuk. A DET kiértékeléséhez először bemutatjuk a bináris osztályozás esetén a tévesztési mátrixot (4.1. táblázat).

4.1. táblázat

A bináris osztályozás esetén a tévesztési mátrix

		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)

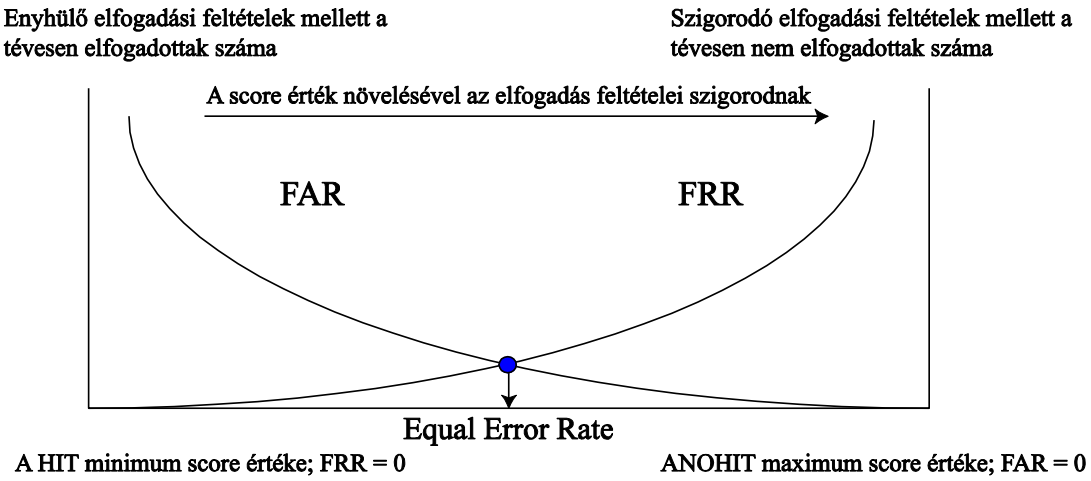
A bináris osztályozáskor megkülönböztetünk első- és másodfajú hibát. Az elsőfajú hiba a téves elfogadás (False Acceptance Rate: FAR; False Positives). A jelen munka során a téves elfogadásról akkor beszélnénk, ha a beérkező szegmens nem átfedő beszéd, de annak fogadja el a gép. A másodfajú hiba a téves elutasítás (False Rejection Rate: FRR; False Negatives rate: FNR) (4.2. ábra). A jelen munka során a téves elutasításról akkor beszélnénk, ha a beérkező szegmens átfedő beszéd, de nem fogadja el annak a gép.



4.2. 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 Caracteristic) 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 (4.3. ábra).



4.3. ábra

Az EER sematikus ábrázolása

5. EREDMÉNYEK

Ebben a fejezetben mutatom be azokat az eredményeket, amelyeket a jelen kutatás során létrehozott nemellenőrzött tanuláson alapuló beszélődetektáló algoritmussal értem el. A rendszert működését nagy mennyiségű spontán társalgásokban teszteltük. Az általunk megalkotott beszélődetektáló alapvetően a BIC-alapú algoritmuson alapszik, amelyet a Beszélődetektálás megoldási módozatai című fejezetben ismertettünk. A kutatássorozat alapvető újdonsága, hogy a beszélődetektálót spontán társalgásokra specifikáltuk nemellenőrzött módszerekkel. Az így kialakított beszélődetektáló további újdonsága, hogy beszélőspecifikus akusztikai jellemzőt kísérleteztünk ki, amelyet mind a beszélőszegmentálásban, mind a beszélőklaszterezésben kívánunk felhasználni. Emellett egy új, ún. mély tanuláson alapuló algoritmust teszteltünk az egyszerre beszélések automatikus osztályozásában. A fejezet további részeiben ezeket az eredményeket mutatjuk be részletesen, és értelmezzük az összefüggéseiket.

5.1. Beszéddetektálás

5.1.1. Bevezetés

Mivel a jelen disszertációnak 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 nemellenőrzött 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 nemellenőrzött tanulási módszer javít-e az eredményeken.

5.1.2. A vizsgálat anyaga

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.

5.1.3. Jellemzőkinyerés

A jellemzőkinyerés előtt a folytonos jelet rövid szegmensekre bontottuk, vagyis ablakoltuk (framekre: 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 centriod 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 koszinusz transzformáció koefficiense, és ahol az N a keret hossza. Ez a jellemző frekvenciákat mutatja meg 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 VAD 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.

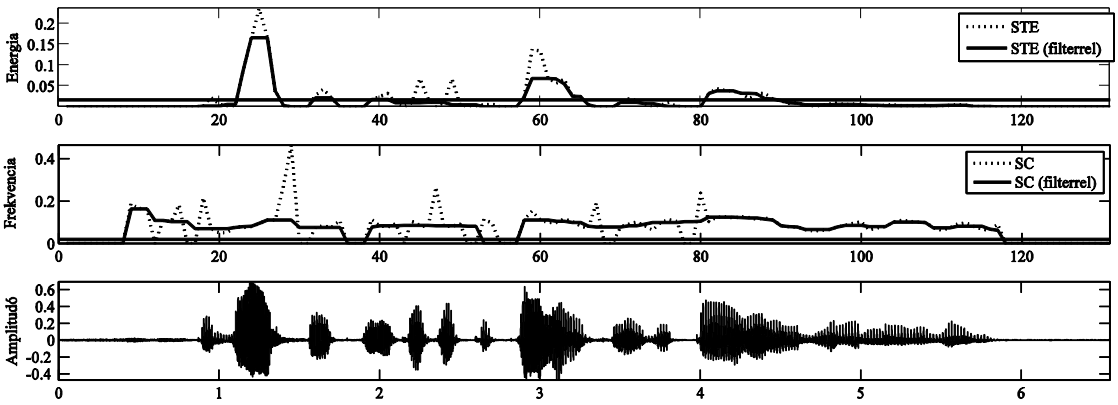
5.1.4. A VAD döntési metódusa

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 (5.1. ábra):

1) Az egyes jellemzők eloszlásának modellezése.
2) A hisztogram simítása.
3) A hisztogram lokális maximumainak detektálása.
4) A küszöb értékének kiszámítás: 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.



5.1 ábra

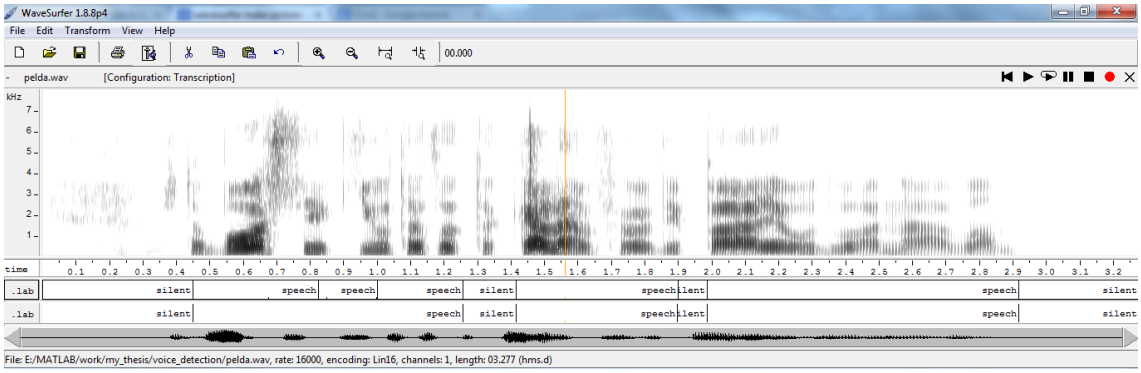
Az egyes jellemzőkre alkalmazott küszöbérték

Amikor a küszöbértékek rendelkezésre állnak mindkét akusztikai jellemzőre: T1 és T2, akkor az egyes ablakolt szegmensekre végrehajtódik a döntési feladat. Ebben az esetben a döntés a következő:

1. a rövid idejű energiafüggvényre: $M1>T1$, akkor beszédrész
2. a spektrális centriodra: $M2>T2$, akkor beszédrész
3. végső döntés: ha 1) és 2) beszédrész.

5.1.5. A VAD utófeldolgozása

A VAD 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 (5.2. ábra).



5.2. ábra

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

5.1.6. Az általunk javasolt eljárás a küszöb meghatározására

A jelen kutatásban a hisztogram számolása és a csúcsok megtalálása helyett nemellenőrzött módszert alkalmaztunk. A korábbi munkákban szintén használtak nemellenőrzött tanuló algoritmusokat a VAD megvalósításában. Ying és munkatársai (2011) szekvenciális Gauss-keverék modell alapú VAD-ot javasoltak, amelynek a bemenete az energia eloszlása Mel-szűrő frekvenciasávjaiban. A kezdeti lépésként a beérkezett keretekre nemellenőrzött módon két Guass-t illeszt, ahol az alacsonyabb középértékkel rendelkező klaszter a nem beszédnek felel meg, míg a magasabb

középértékkel rendelkező a beszédrésznek. Ezt a metódust a küszöbérték meghatározásában is alkalmazzák.

A jelen munkában a középpontok (beszéd és nem beszéd) megtalálásához klaszteranalízist használtunk. A klaszteranalízis lényege, hogy egy adattömböt több homogén részcsoportokra bontsuk 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 mérős eljárás az euklidészi 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:

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

ahol μi az i-edik klaszter középpontja. rij = 1, ha az i-edik klaszterbe soroljuk a j-edik mintát, amúgy 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 deriválja 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 M1 és M2 lesz.

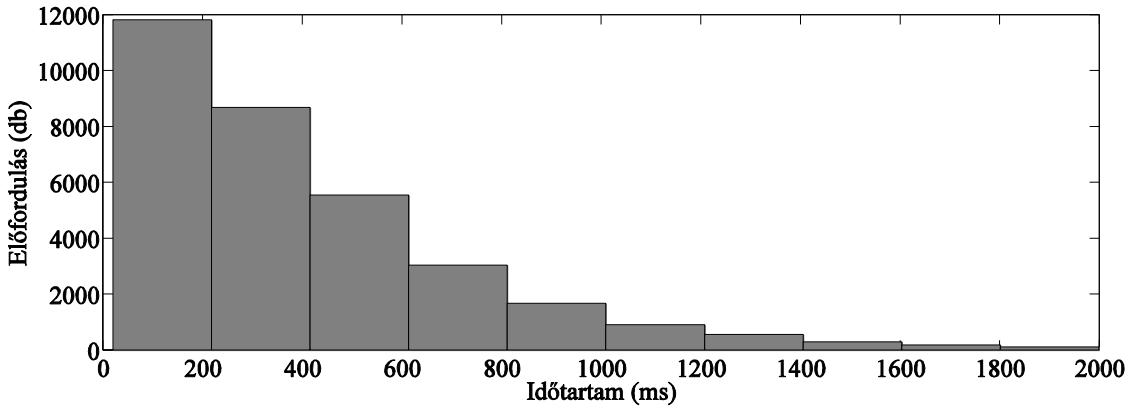
5.1.7. A VAD kiértékelése

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

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

5.1.8. Eredmények

A BEA-adatbázisban (Gósy 2012) a nembeszéd-részek átlagos időtartama 413 ms volt, a szórása pedig 438 ms (5.3. ábra).



5.3 ábra

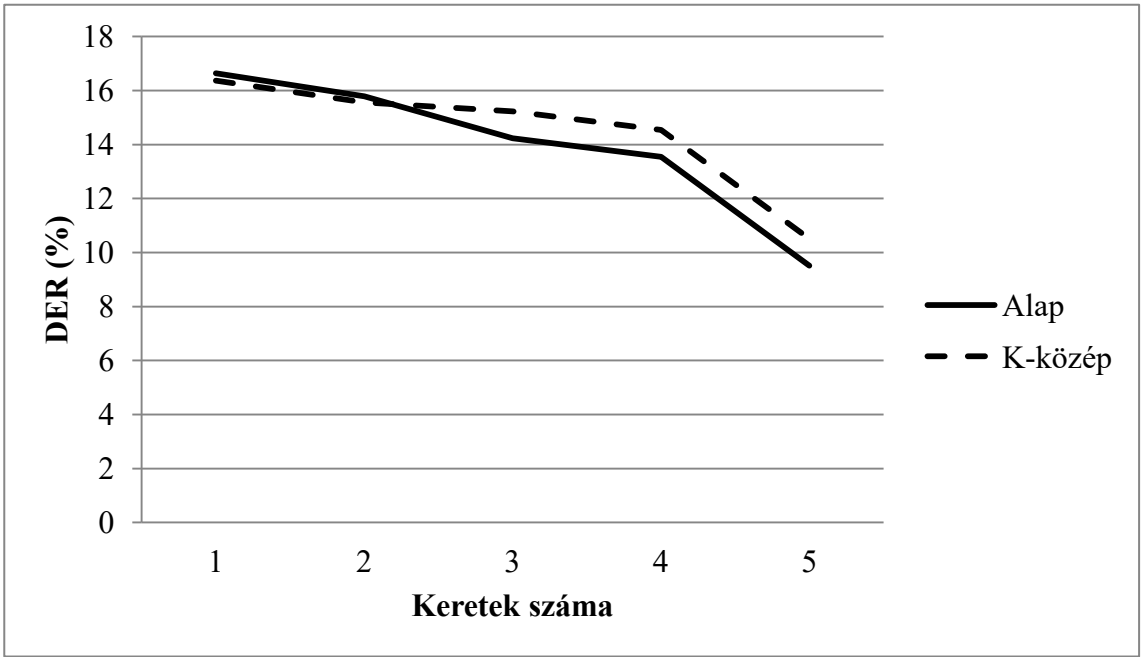
A nembeszéd-részek időtartamának eloszlása

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

A VAD 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 (5.4. ábra). Ezzel egy időben azt is teszteltük, hogy melyik módszerrel (az alap vagy az általunk javasolt) tudunk elérni jobb detektálási eredményt.



5.4. ábra

A DER értéke a keretek számának és a küszöböt meghatározó módszer 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.

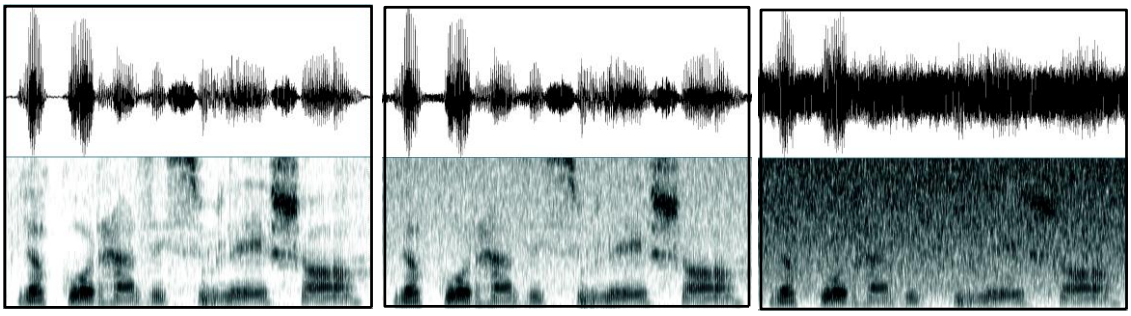
A Diarization Error-t felbontva láthatjuk (5.1 táblázat), hogy a legtöbb hiba abból adódik, hogy a gépi annotálásban sok helyen helytelen beszéd vagy szünet címkéje szerepel, vagy a szegmens helye megfelelő a beszédben, csak azonosítója téves.

5.1. táblázat

Az általunk javasolt algoritmus teljesítménye 250 ms-os ablakhosszúságú ablakozással

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

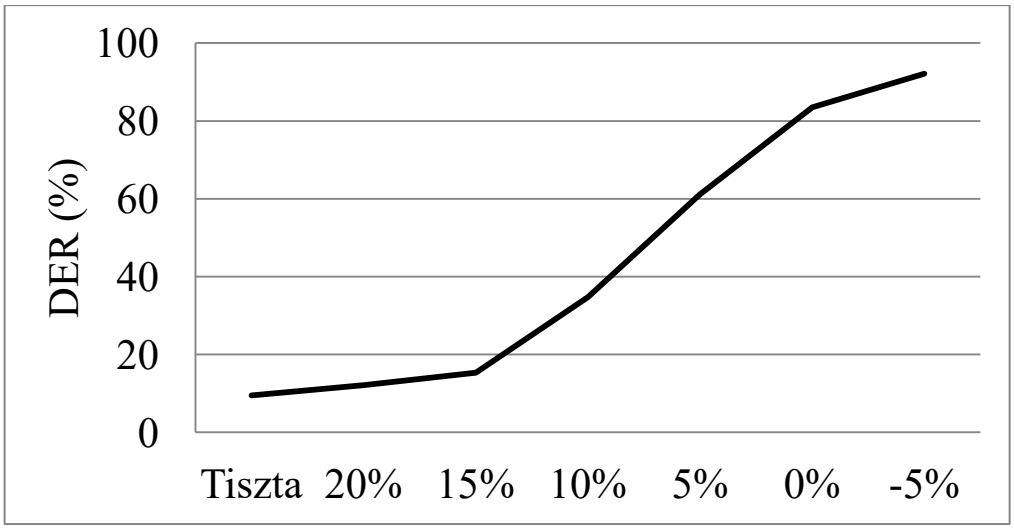
A tesztelés során megvizsgáltuk azt is, hogy az általunk javasolt VAD milyen eredménnyel működik különböző jel/zaj viszonyú (SNR: Signal to Noise Rate) spontán beszédben. Rendszerünk zajtűrését tehát úgy teszteltük, hogy a felvételhez fehér zajt kevertünk a jel/zaj arányt folyamatosan csökkentve ezzel. A felvétel eredeti SNR értéke átlagosan 25%. Az SNR értékét 5 dB-enként csökkentettük (5.5. ábra).



5.5. ábra

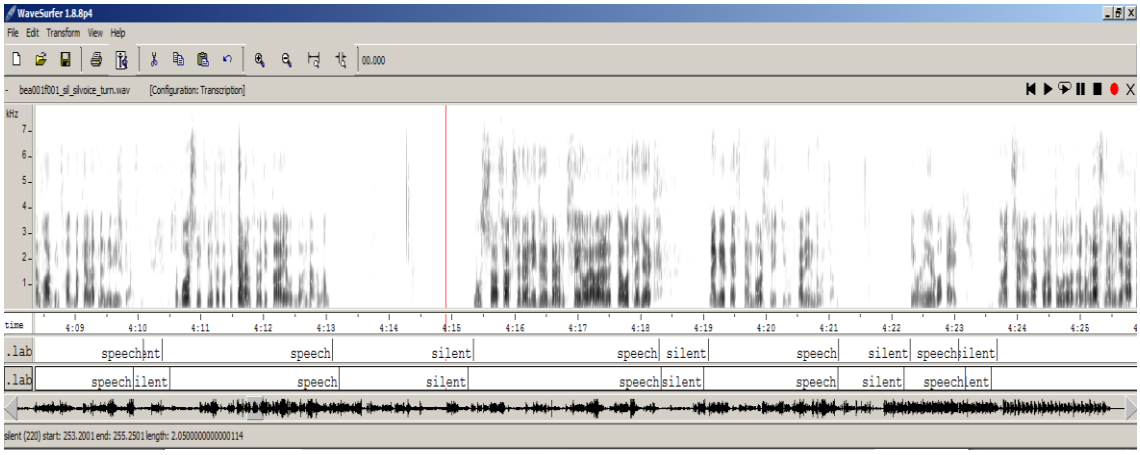
Egy megnyilatkozásnyi audioszegmens különböző SNR érték mellett (25, 15, 0) (spektruma lent és a hangnyomás-időfüggvénye fent).

A zaj hatására az eredményeink csökkennek, azonban még 10% SNR mellet is 34,72%-os volt a hiba értéke (5.6., 5.7. ábra).



5.6. ábra

A szegmentálási hiba értéke a jel/zaj arányának függvényében



5.7. ábra

Példa a manuális és az automatikus címkesor egyezésére (a felső címkesor az eredeti annotáció, míg az alsó az automatikus kiment)

5.1.9. Következtetések

A jelen kutatásban egy meglévő algoritmust módosítottunk, amely képes az akusztikai jelből kinyert jellemzők alapján automatikusan osztályozni a spontán

beszédben a beszéd- és a nembeszéd-részeket. Az általunk módosított VAD statisztikai módszereken alapul, azon belül nemellenőrzött tanuláson alapszik.

Az eredmények azt mutatják, hogy az általunk javasolt módszerrel a beszéddetektálási hiba csökkenthető, statisztikailag azonban a javulás nem volt igazolható.

Az általunk javasolt rendszer jó minőségű felvételen 90,49%-os eredménnyel működik. 10%-os jel/zaj arányig még közel 65,28%-os eredménnyel, 5%-os jel/zaj aránytól viszont már csak 38,8%-os helyes találati aránnyal működik a rendszer. Ez azzal magyarázható, hogy a VAD- algoritmusban nem használtunk zajszűrőt. Ezért tervezzük, hogy zajszűrőkkel is kísérletezni fogunk. Az elkészített VAD egy általunk fejlesztett beszélődetektálóba fogjuk integrálni, amely feltehetőleg javítani fogja annak működését.

5.2. Beszélőfelismerés

5.2.1. Bevezetés

A jelen kutatás célja kettős. Az első célja, hogy megvizsgálja, hogy a magyar nyelvű beszédben mely spektrális régiók beszélőspecifikusak. A második célja az, hogy a beszélőket MFC-vel előfeldolgozva GMM-kkel, illetve GMM-UBM-kkel modellezzük és osztályozzuk 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 kívánjuk integrálni.

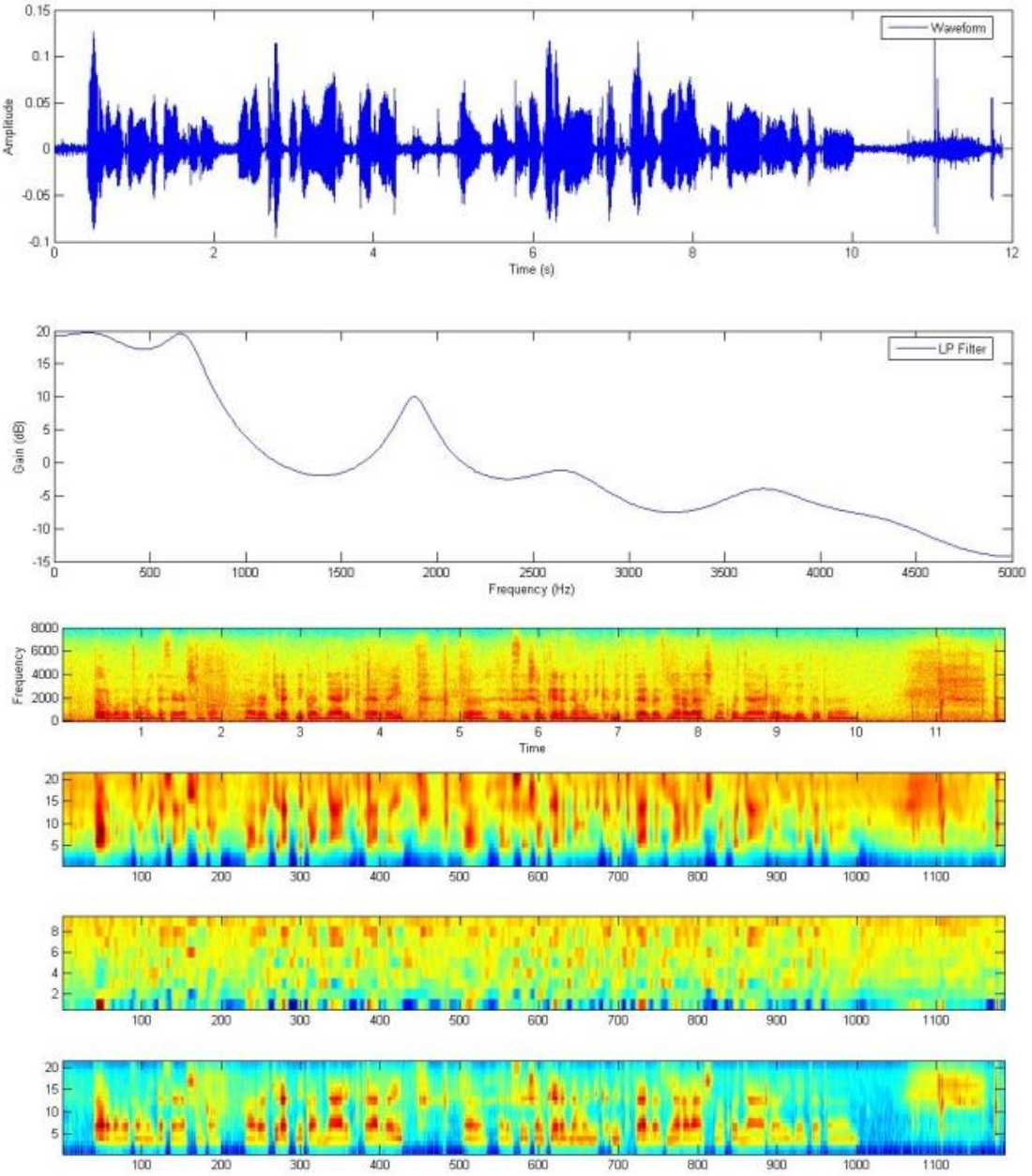
5.2.2. A vizsgálat anyaga

A jelen kutatásban 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 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.

5.2.3. Jellemzőkinyerés

Jóllehet még nem ismeretes olyan akusztikai jellemző, amely kifejezetten beszélőspecifikus lenne, mégis sok jól alkalmazható jellemző létezik: a beszéd spektrumában ugyanis fellelhetők olyan lenyomatok, amelyek jól megkülönböztetik az egyes beszélőket. Ez azért lehetséges, mert a spektrum reprezentálja a beszélő artikulációs csatornájának felépítését, amely a legdominánsabb fizikai jellemző a beszélőfelismerésben (Atal 1976). Az LPC (lineáris predikciós együttható, vö. 5.8. ábra) eljárás kiválóan leképezi a beszélő artikulációs csatornáját, ami miatt ez az egyik legtöbbször alkalmazott jellemző a beszélőfelismerésben. Az egyik hátránya azonban

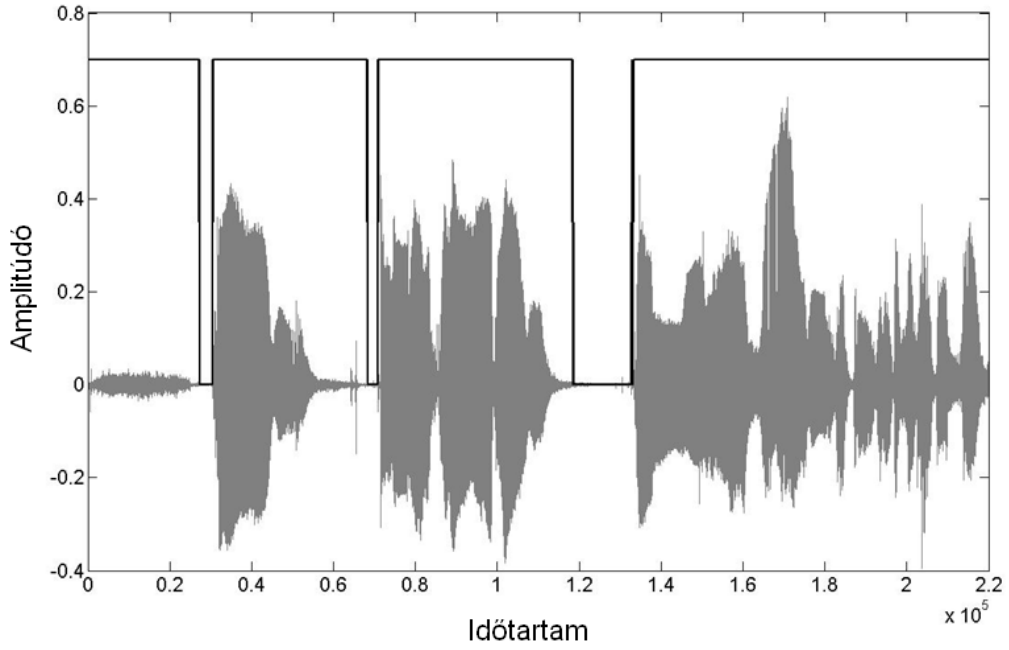
az, hogy nagyban függ a zaj hatásától (Tierney 1980), ezért a kutatások többsége a rövid idejű spektrális burkológörbe érzeti transzformációján alapuló eljárásokat alkalmazza (MFCC: Mel Frequency Cepstral Coefficients; PLP: Perceptual Linear Prediction, vö. 5.8. ábra) (Mermelstein 1976; Hermansky 1990).



5.8. ábra

Az LP együtthatók (fent) és az MFC együtthatók (lent) azonos beszélő esetében

Az MFCC kiszámítása előtt a beszédet előfeldolgozzuk, vagyis szűrjük. Először előkiemelést (pre-emphasis) hajtunk végre, amely során javítjuk a jel/zaj viszonyt. Az előkiemelés után a hangfelvételből kivágjuk azokat a részeket, amely nem tartalmaz beszédet. Ezt az eljárást beszéddetektálásnak nevezzük (Voice Activity Detection) (5.9. ábra).



5.9. ábra

A hanganyag beszédrészeinek kinyerése

A beszéddetektálás után számítjuk ki a beszélőfelismerőben használt jellemzővektorokat. Az MFC-együtthatók kiszámítási módja a következő: először a beszédjelen gyors Fourier-transzformáció történik (Fast Fourier Transform, FFT), ennek során a beszédből egy rövid részt (például 25 ms hosszú darabot) kivágunk, súlyozzuk egy ún. ablakfüggvénnyel, és Fourier-transzformációval meghatározzuk a spektrumát. Ezután 10 ms-ot továbblépünk, és ugyanezt ismételjük mindaddig, amíg el nem érünk a feldolgozandó szakasz végéig, ily módon 10 ms-onként megkaptuk a beszédjel spektrumát. Második lépésként ezeken a spektrumokon ugyanolyan elemzést végzünk, amilyet az emberi fül is, azaz a spektrumot az ún. kritikus frekvenciasávok szerint bontjuk fel. Ezt a műveletet szűrősoros elemzéssel hajtjuk végre, amely a kritikus sávoknak megfelelő frekvenciatartományokba bontja a jelet. A szűrősor általában 20

sávátereszt szűrőből áll, amelyek kimenetén egy, a sávba eső intenzitással arányos számszerű érték jelenik meg, azaz tulajdonképpen 10 ms-onként egy 20 dimenziós vektort kapunk. Mivel ezek az értékek egymással korrelálnak, ezért a dimenziószám csökkenthető, mégpedig egészen 12-re az ún. diszkrét koszinusz transzformáció segítségével (Discrete Cosine Transform, DCT). Ezután általában hozzáveszik a vektorhoz a teljes beszédjelszegmens átlagos energiáját, majd az így összesen 13 érték első és második deriváltjait, így összesen egy 39 dimenziós vektort, ún. jellemzővektort kapunk. Ez az eljárás tehát a teljes spektrumot kódolja le (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 jelen tanulmányban sem az energiát, sem pedig a deriváltakat nem vettük bele az akusztikai jellemzők közé. Az akusztikai jellemzőket a PLP-RASTA csomagban található MATLAB szoftverkörnyezetre írt algoritmust használtuk (vö. Daniel 2005).

5.2.4. Gauss-keverék beszélőmodell

5.2.4.1. Gauss-keverék modell

Ebben a részben bemutatjuk a Gauss-keverék 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 tanítóminták 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, amely mindegyike a beszédből kinyert d-dimenziós jellemzővektor. Mivel ezen jellemzővektorok eloszlása nem ismert, ezért Gauss-keverékkel szokás modellezni őket, amely súlyozott összege az m komponensű

eloszlásnak. A *d*-dimenziós jellemzővektor matematikai leírása *m* komponensű Gauss-keverékkel:

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

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

A kevert Gauss-modell paramétereinek, a_i, μ_i és Σ_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, kovarianciamátrixszával és a keveréksúlyukkal reprezentáljuk. Ezeket a paramétereket együttesen a következőképpen foglalhatjuk egybe:

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

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:

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

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

$$p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$$

.

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

Minden egyes iterációval újrabecsüljük (reestimation) a GMM paramétereit. A keverék súlyok újraszámolása a következőképpen történik:

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

.

Az átlag, μ újraszámolása:

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

.

A kovariancia mátrix, Σ újraszámolása:

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

.

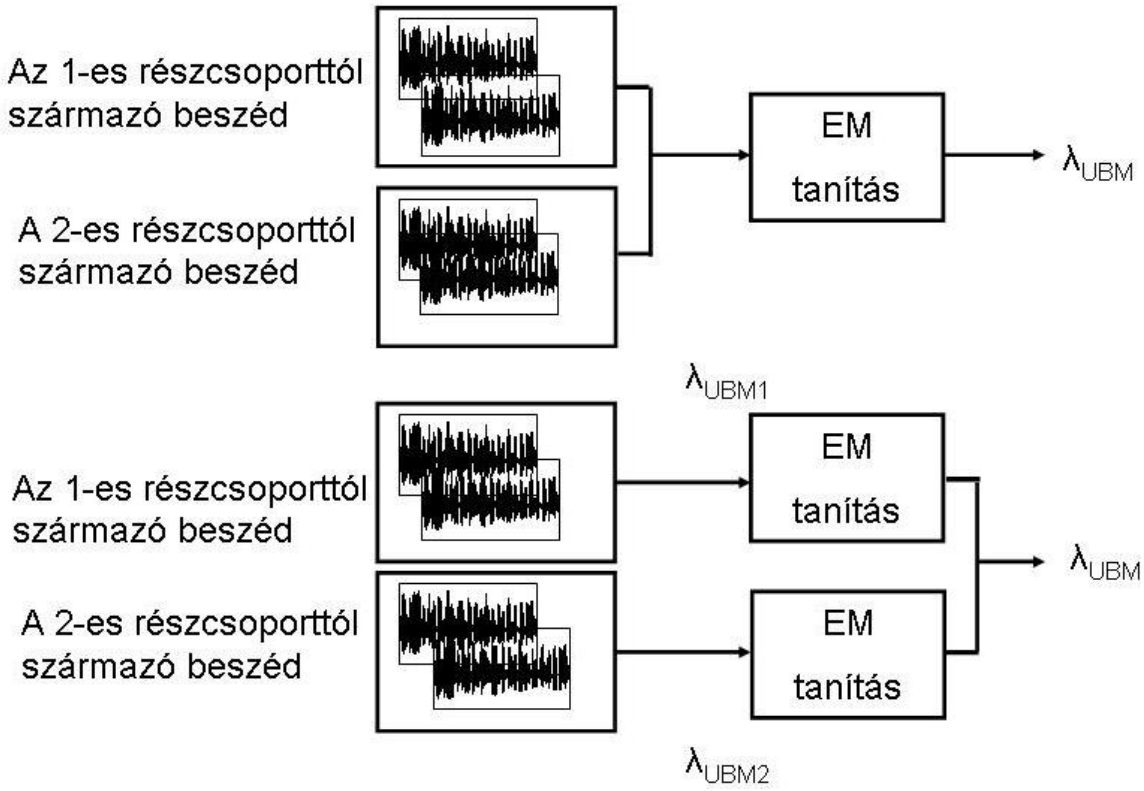
Az a posteriori valószínűségét az *i*-edik beszélőszemély-modellnek a következő egyenlettel számolhatjuk:

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

.

5.2.4.2. 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). A GMM-UBM alapú beszélőfelismerő rendszer egy beszélőfüggetlen háttérmodellt alkalmaz, amelyet a következőképpen reprezentálunk: $p(X \mid \lambda_{\overline{hyp}})$. 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 kialakításra, 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. 5.10. ábra).



5.10. ábra

A UBM kétféle tanítási módszere

5.2.5. A beszélőegyezés mérése

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

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

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

ahol m_{Ci}: i-edik keverék súly, μ_{Ci} : átlag vektor, Σ_{Ci}: kovarianciamátrixa az azonosítandó beszélői modellnek λ_C.

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:

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

,

ahol m_{bi}: i-edik keverék súly, μ_{bi}: átlag vektor, Σ_{bi}: kovarianciamátrixa a háttérmodellnek λ_b.

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 minimumhiba-arány osztályozóban a diszkriminanciafüggvény az *a posteriori* valószínűség:

$$g_k(X) = p(\lambda_k \mid X)$$

.

Felhasználva a Bayes-tételét:

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

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:

$$g_k(X) = p(X \mid \lambda_k)$$

.

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

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

,

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

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

.

5.2.6. Kiértékelés

A beszélőfelismerő rendszer kiértékelésére a felismerés pontosságát (Accuracy), vagyis a helyesen felismert adatok arányát adtuk meg. A tévesztési mátrixot többosztályos problémára kell számolnunk. A diagonális elemek tartalmazzák a helyesen felismert adatok számát osztályonként. A nem diagonális elemek jelentése: hány esetben lett a tesztadathoz a j-edik osztály rendelve, amikor valójában az i-edik osztályhoz tartozik. Ebből adódóan a pontosságot a következőképpenn számoljuk:

$$Pontosság = \frac{Diagonális\ elemek\ összege}{Összes\ tesztadat\ száma}$$

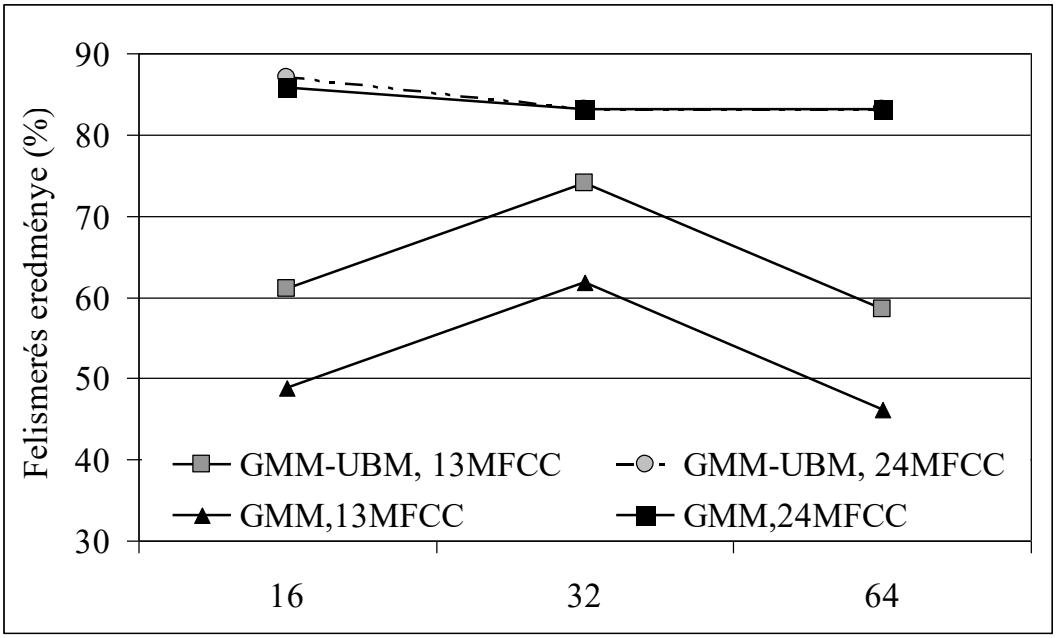
A statisztikai összehasonlításokat nem parametrikus, összetartozó (Wilcoxon-próba) és nem összetartozó (Mann–Whitney-próba) teszttel végeztük el, kiegészítve őket Monte Carlo szimulációval.

5.2.7. Eredmények

A jelen kutatás során elkészítettünk egy férfi/nő osztályozó algoritmust, vagyis egy nem szerinti osztályozót. Milan Sigmund (2008) kimutatta, hogy a magasabb számú koefficiensek őrzik a személy nemére vonatkozó ismertetőjegyeket, ezért az osztályozáshoz akusztikai jellemzőként 13 koefficienst tartalmazó MFCC mellett, 24 koefficienst tartalmazó MFCC-t is használtunk. A tanítás során egy női és egy férfi modellt hoztunk létre. Mindkét esetben 40-40 beszélő 25 s-os beszédmintán tanítottunk. A teszteléskor 40-40 13 s-os beszédmintát használtunk. A UBM kialakításához 20

beszélő 25 s-os részét használtuk fel. Mind a női, mind a férfi modell kialakításakor 16, 32 és 64 Gauss-komponenst használtunk.

A legjobb eredményt akkor kaptuk, ha 24 MFCC-t, GMM-UBM-et és 16 Gauss-komponenst tartalmazó osztályozót használtunk. Ekkor a felismerés eredménye 87,01%-os volt. A legrosszabb eredményt a 13 MFCC-t, GMM-et és 64 Guass-komponenst tartalmazó osztályozó adta (5.11. ábra).



5.11. ábra

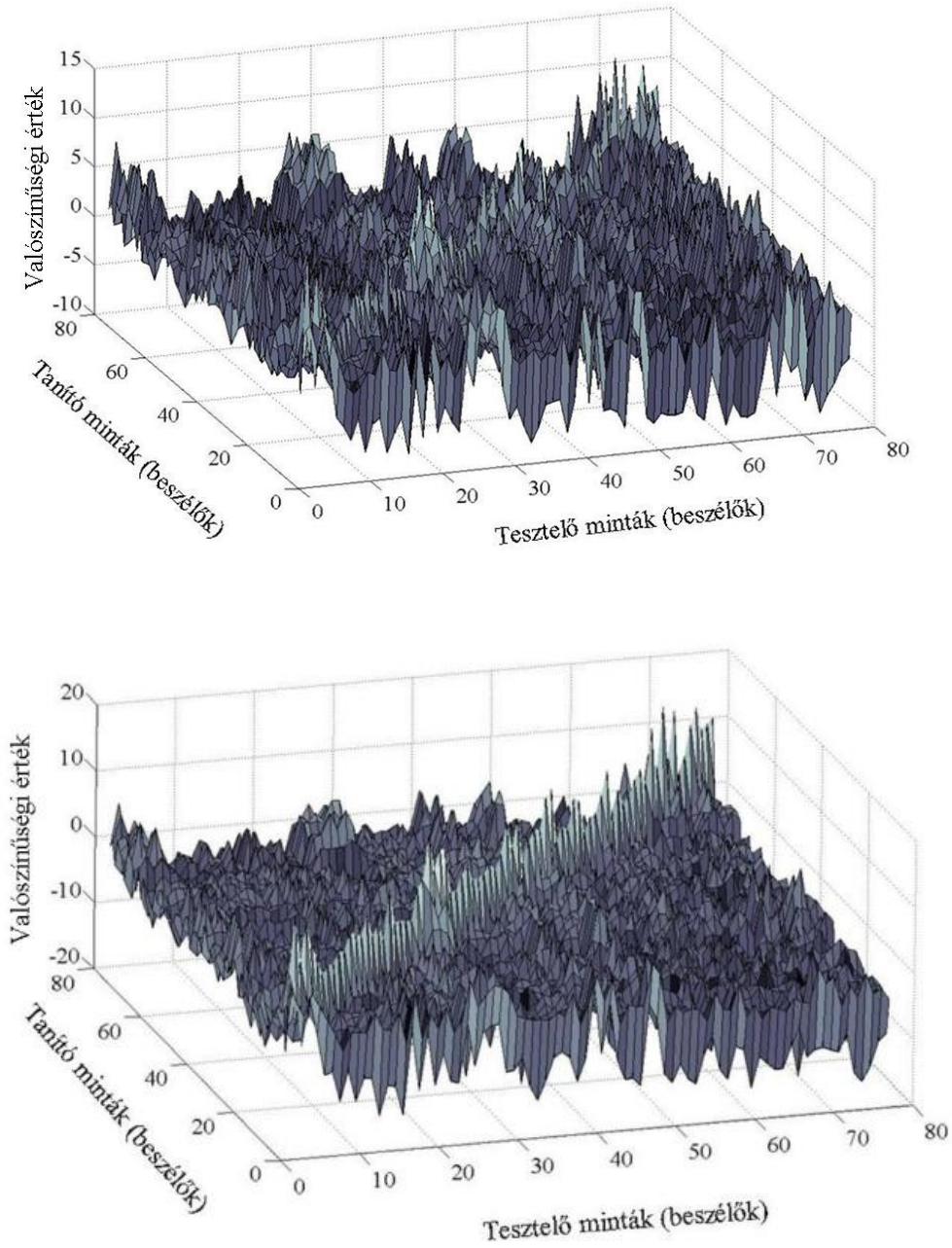
A felismerés eredménye a jellemző, az osztályozó és a Gauss-komponensek számának függvényében

Minden esetben az 24 MFCC-t használó osztályozó jobban teljesített a 13 MFCC-t használóval szemben (Wilcoxon-teszt: Z=-2,201; p=0,028). Az adatokból az is jól látszik, hogy a UBM használata a 13 MFCC-t használó osztályozó esetében javított az eredményeken.

Az **egyes beszélők** automatikus felismeréséhez különböző számú Gauss-komponenst tartalmazó GMM-eket használtunk a tanítás és a tesztelés során. A tanításhoz 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-adatbázistól különböző 20 adatközlő 25 s-

os beszédét használtunk fel. A kutatás során teszteltük azt is, hogy mely MFC-együtthatóval a legsikeresebb az osztályozás: (i) teljes spektrumot kódoló MFC; (ii) 1,5–2,5 kHz közötti MFC, (iii) 2,5–3,5 kHz közötti MFC; (iv) 3,5–4,5 kHz közötti MFC.

A Gauss-komponensek függvényében a valószínűségi érték az azonos beszélők esetében egyre magasabb értéket vesz fel, míg a különböző beszélők esetében ez az érték csökken (5.12. ábra).



5.12. ábra

A valószínűségi érték 2 komponensű Gauss (fent), és 256 komponensű Gauss esetén (lent)

Az ábrán jól látható, hogy a magasabb komponens számú 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 (5.2. táblázat), hogy ha a GMM-et általános háttérmodellel használjuk, akkor átlagosan jobb eredményeket kaptunk, mint a GMM általános háttérmodell nélkül. Megvizsgáltuk, hogy ez a teljesítmény szignifikánsan jobbnak mondható-e, amelyet nem parametrikus, összetartózó mintás elemzéssel végeztünk el, kiegészítve Monte Carlo szimulációval. A statisztikai elemzés szerint ez a különbség szignifikáns: Wilcoxon-próba: Z=-2,944; p=0,003. Megvizsgáltuk azt is, hogy mely akusztikai jellemzővel használt osztályozó adja a legjobb eredmény. A 12.2. táblázatból látható, hogy a legjobb osztályozási arányt a 2500-3500 Hz részsávra számolt MFC-együtthatókkal értük el mind a GMM, mind a GMM-UBM esetében. Ez azonban statisztikailag csak részben igazolható. Az $MFC_{(2,5–3,5)}$ jellemzővel elért eredmények szignifikánsan különböznek az $MFC_{(1,5–2,5)}$-vel (Z=-2,201; p=0,028) és az $MFC_{(3,5–4,5)}$-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 $MFC_{(2,5–3,5)}$ jellemző, mint az $MFCC_{(full\text{-}ban)}$. Ez az eredmény megerősíti a nemzetközi kutatások eredményeit, miszerint valóban a spektrum ezen régiója (2,5 kHz és 3,5 kHz) hordozza az egyéi beszédjellemzőket.

Elemeztük, hogy a felismerés pontossága hogyan alakul a Gauss-komponensek függvényében. Az eredményekből az látszik, hogy a Gauss-komponensek számának növekedésével javul a pontosság értéke is.

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

5.2. táblázat

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

		Gauss-komponensek száma					
Osztályozó	Jellemző	**8**	**16**	**32**	**64**	**128**	**256**
GMM	$MFC_{(full\text{-}ban)}$	28,81	34,01	56,08	69,46	72,46	75,66
	$MFC_{(1,5–2,5)}$	26,32	32,58	54,11	67,61	69,39	72,01
	$MFC_{(2,5–3,5)}$	**33,72**	**39,71**	**60,20**	**72,89**	**76,44**	**77,12**
	$MFC_{(3,5–4,5)}$	26,85	30,01	56,08	67,46	70,46	70,66
GMM-UBM	$MFC_{(full\text{-}ban)}$	29,15	34,35	55,42	74,8	77,81	76,76
	$MFC_{(1,5–2,5)}$	30,815	32,01	61,08	71,46	75,78	72,60
	$MFC_{(2,5–3,5)}$	**34,05**	**35,21**	**66,53**	**74,91**	**76,11**	**79,76**
	$MFC_{(3,5–4,5)}$	31,15	33,35	57,42	74,55	75,81	74,25

5.2.8. Következtetések

A jelen értekezésben egy szövegfüggetlen nem-, és beszélőfelismerőt készítettünk el MATLAB implementációban. A nem osztályozásában kimutattuk, hogy a 24 koefficienst tartalmazó MFCC-vel jobb eredményt lehet elérni, mint 13-mal. Mindez alátámasztja azt az elképzelést, hogy a magasabb kepsztrális együtthatók őrzik a nemre utaló akusztikai jegyeket.

A beszélőszemély-felismerésben az eredmények azt mutatják, 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, miszerint valóban a spektrum ezen régiója (2,5 kHz és 3,5 kHz) hordozza az egyéni beszédjellemzőket.

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, amelynek értéke 79,76% volt. Az 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.

Az eredményeink javítására újabb kísérletet tervezünk, amely több adatközlővel történik, más akusztikai jellemzőket és más mintaillesztési eljárást szeretnénk használni.

5.3. Az egyszerre beszélések automatikus osztályozása

5.3.1. Bevezetés

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

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 a többcsatornás felvételeket elemzik (Yamamoto 2006; Laskowski–Schultz 2006; Xiai 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 mutatták ki, hogy csökkenti a beszélődetektálási hiba arányát (DER) (Boakye et al. 2008; Boakye 2008; Trueba-Hornero 2008).

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, LPC-analízist 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: nem beszéd, 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 árfedő 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).

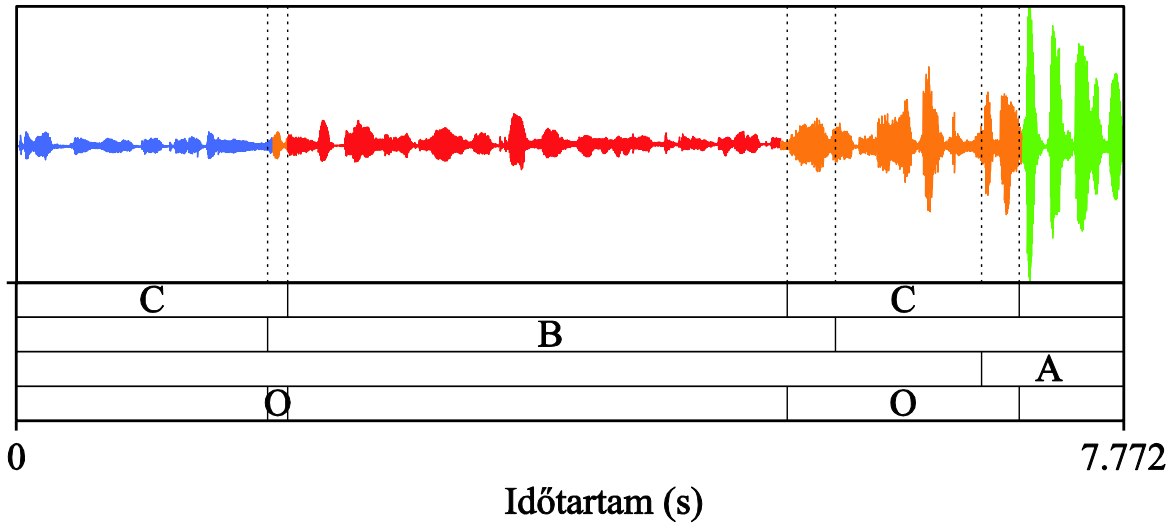
A jelen kutatásban egy ANN/SVM (Artificial Neural Network/Support Vector Machine, Mesterséges Neuron Háló/Szupport Vektor Gép) hibrid rendszert hoztunk létre az egyszerre beszélések automatikus osztályozásához.

Az osztályozás során az első lépés a lényegkiemelés, amelynek fő feladata, hogy a beszédjelből olyan információkat vonjuk 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.

Az osztályozás második fontos lépése a mintaillesztés, amelyben két fontos részfeladatot kell megoldani: (i) osztályozás, vagyis melyik beszédrészlet-modell 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 Gauss-keverék 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; Bourlard–Morgan 1993). Az elmúlt években az ANN egy új fajtája jelent meg: ún. mély neuronhálók, amelyek a vizsgálatok szerint igen jól alkalmazhatók többek között a beszédhang-felismerésben (Dahl et al. 2010; Grócz–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-el 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).

5.3.2. A vizsgálat anyaga

A társalgásokban manuálisan jelöltük azokat a részeket, ahol egyszerre több adatközlő beszél, illetve azokat a részeket, ahol csak egy beszélő beszél (5.13. ábra).



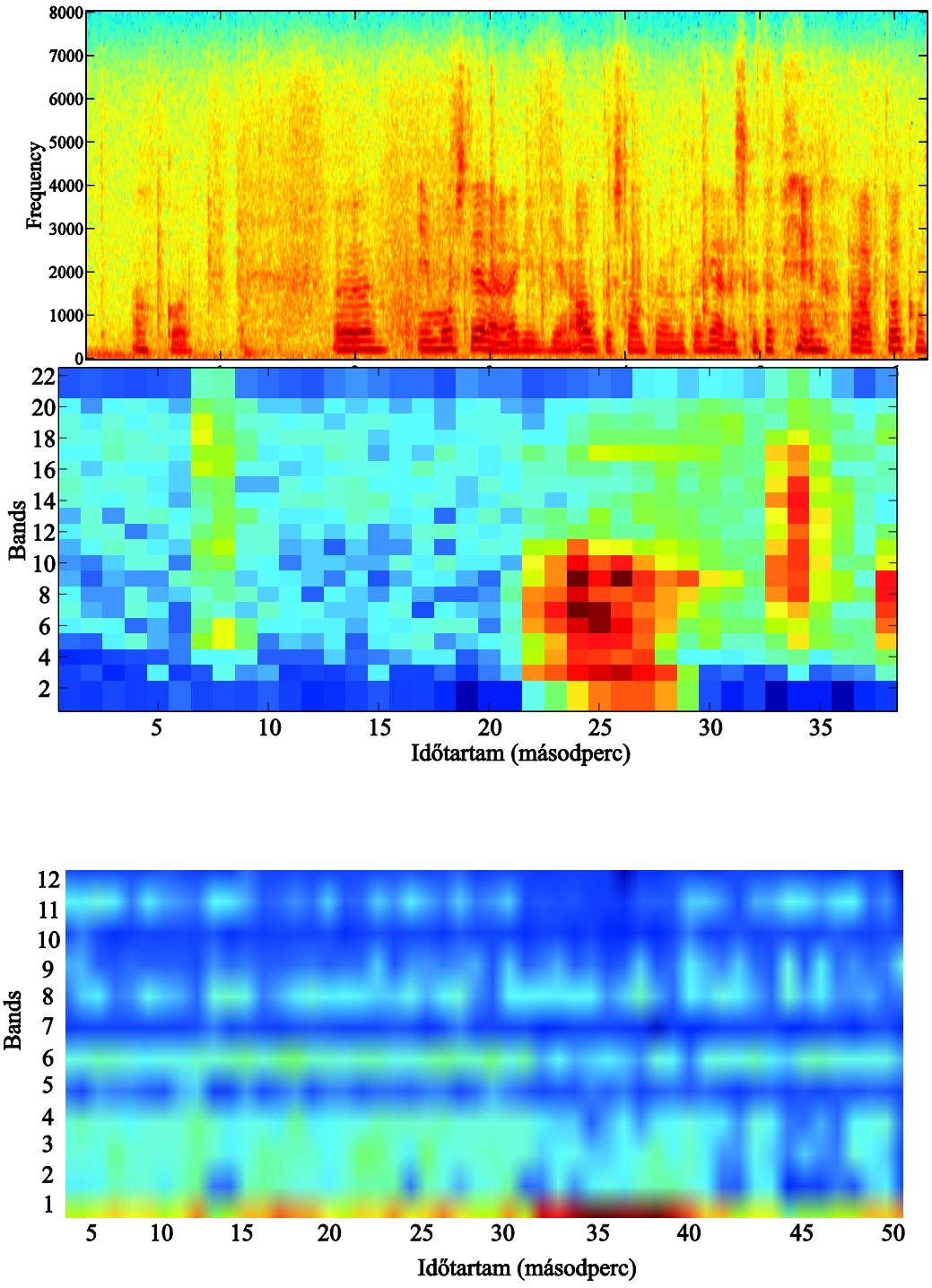
5.13. ábra

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 részvevő szólal meg egyszerre, vagyis ahol átfedő beszéd van. Ezen intervallumok összidőtartama közel 7 óra, ami a teljes korpusz 12%-a.

5.3.3. Jellemzőkinyerés

Az egyszerre beszélések jó megfeleltethetőségéhez az akusztikai beszédjelből különböző jellemzőket nyertünk ki. Az első kísérletben az akusztikai jelet FFT-analízissel felbontottuk spektrális jellemzőkké (5.14a). A második kísérletben egy igen elterjedten használt, az emberi hallást is modellező MFC-együtthatókat használtuk (5.14b). A harmadik kísérlet során a Mel-skála szerinti logaritmikus szűrőbank jellemzőt vizsgáltuk. A negyedik kísérletben a spektrumot részsávokra bontottuk, és az egyes részsávokban számoltuk ki a jel energiáját (5.14c).



5.14a–c ábra

Az akusztikai jelből számolt spektrum (fent), MFC-együtthatók (középen) és részsávenergia (lent)

(i) **A spektrum (SP)** kiszámolásához 256-ponos 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 (MFC) együtthatók** kinyeréséhez a PLP-RASTA csomagban található MATLAB szoftverkörnyezetre írt MFCCalgoritmust használtuk (vö. Daniel 2005). A jellemzők száma egy-egy időkeretben 39: a szokásos 12 MFCC koefficiens + az energia logaritmusa + ezek első két deriváltja (13*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 időkertre.

(iii) A **Mel-skála szerinti logaritmikus szűrőbank (MSL)** számítása ugyanúgy történik, ahogyan az MFC kiszámítása. A különbség abban áll, hogy a Mel-frekvenciás szűrés után vesszük annak logaritmusát, de nem végezzük el a kepsztális transzformációt. Ennek kiszámítása szintén: 12 koefficiens + az energia logaritmusa + ezek első két deriváltja (13*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 időkertre.

(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 koszinusz transzformáció) 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.

5.3.4. Lényegkiemelés

5.3.4.1. 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 spektrogrammok, és a leggyakrabban használt Mel-skálázott spektrogrammok, amelyek az emberi hallást modellezik.

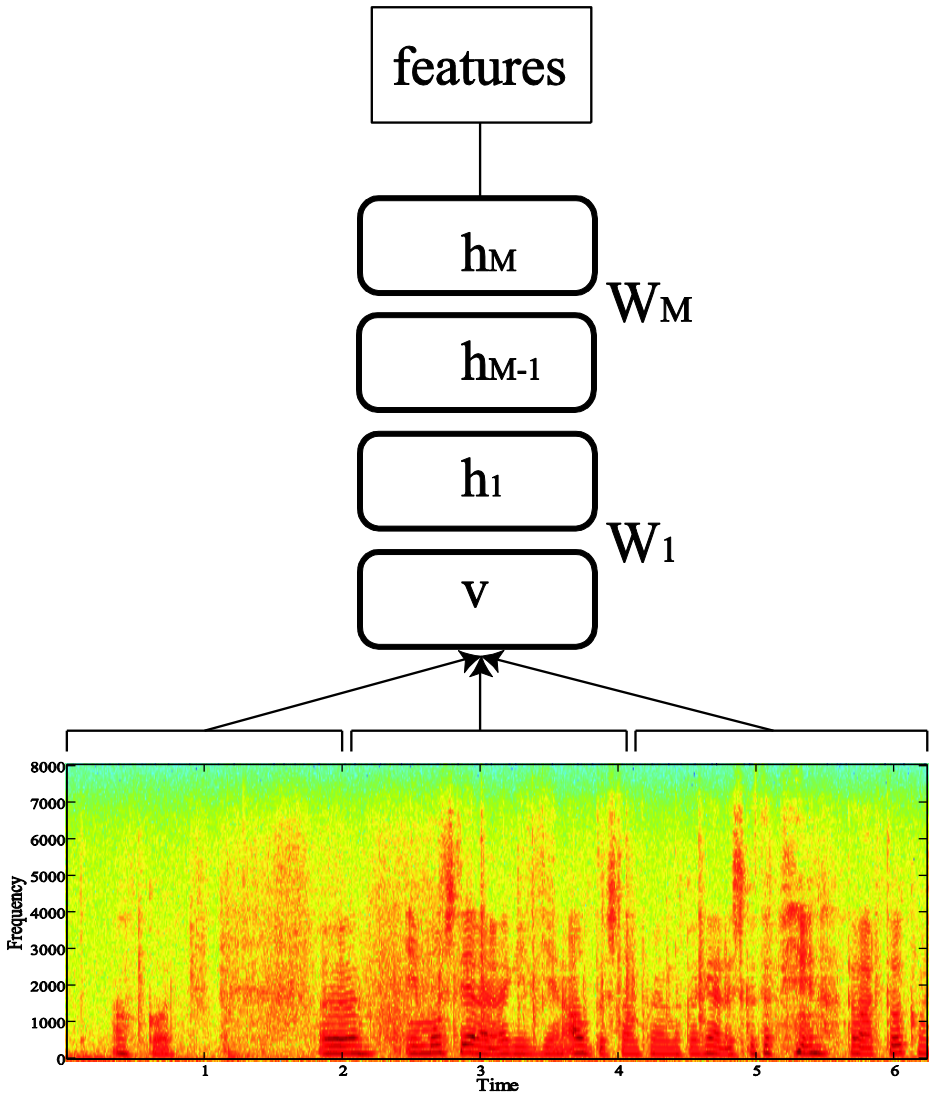
A RBM (Restricted Boltzmann Machine, korlátos Boltzmann-gép) alapvetően két különböző réteget tartalmaz: látható és rejtett réteg. A korlátos jelző arra utal, hogy a neuronok között csak akkor van összeköttetés, ha az egyik a látható, a másik pedig a rejtett réteghez tartozik. Az azonos rétegbe tartozó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

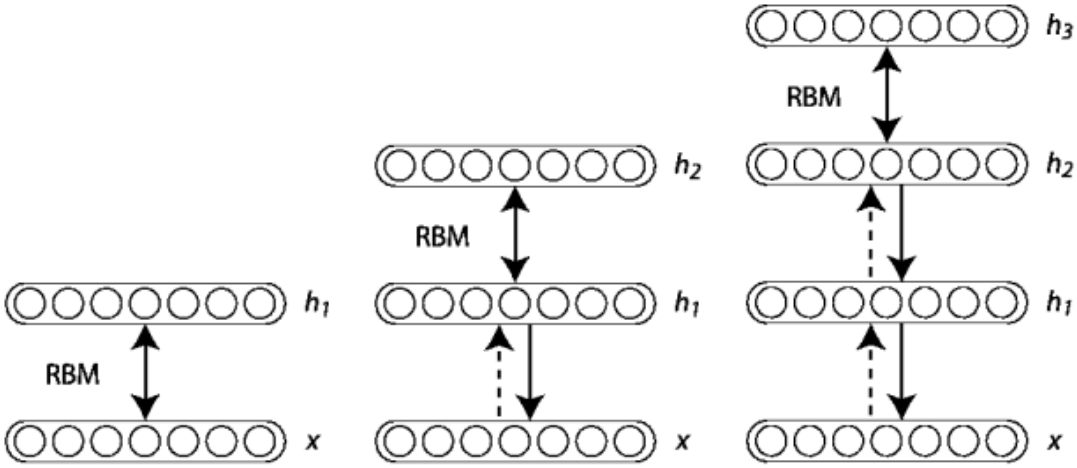
alkalmazni – főként a képfeldolgozásban –, amely ebben az esetben nemellenőrzött tanulási eljárással működik (5.15. ábra).



5.15. ábra

Jellemzőkinyerés korlátozott Boltzmann-géppel

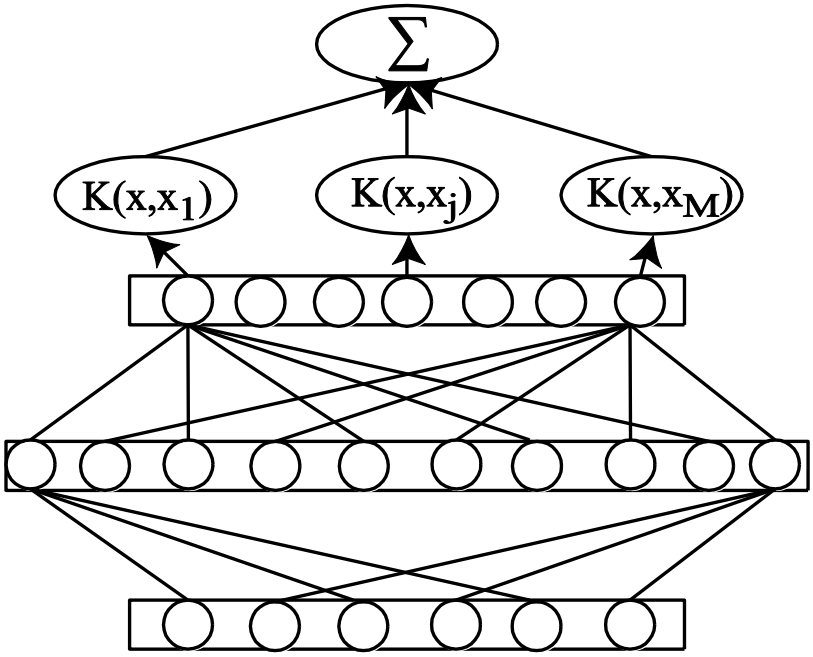
Az RBM előnye, hogy könnyedén mély neuronná lehet alakítani, ha az egyes RBM-eket összekötjük, előállítva ezzel egy hierarchikus tanulási láncot, így segítve a magasabb szintű struktúrák kinyerését az adatokból (5.16. ábra).



5.16. ábra

A korlátozott Bolztmann-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 (5.17. ábra).



5.17. ábra

Szupport Vektor Gép mély neuronhálóval előtanítva

5.3.4.2. Az RBM előtanítási paraméterei

Az RBM előtanításához az akusztikai paramétereket 15 keret hosszúságú csúszóablakkal nyerjük ki. Mindegyik összefüggő ablakot felhasználjuk az RBM tanításához. Az RBM látható egységeinek száma: a jellemzővektor dimenziószáma 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 (H1) a unitok száma 300.

A második RBM-ben (H2) a unitok száma 600.

A harmadik RBM-ben (H3) 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.

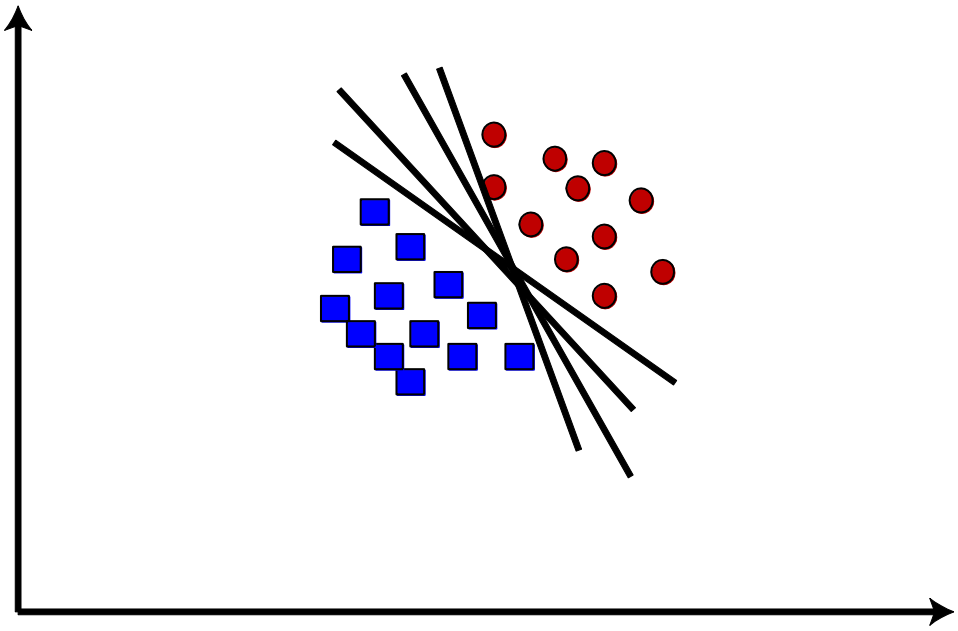
Az RBM megvalósítását a MATLAB-ban írt, Kyung Hyun Cho GitHub toolbox-át használtuk (Cho 2013).

5.3.5. Osztályozás

Az átfedő és nem átfedő beszédrészeket Szupport Vektor Géppel (SVM) kernelfüggvényként radiális bázisfüggvényt (RBF) alkalmazva osztályoztuk.

5.3.5.1. Szupport Vektor Gép (SVM: Support Vector Machine)

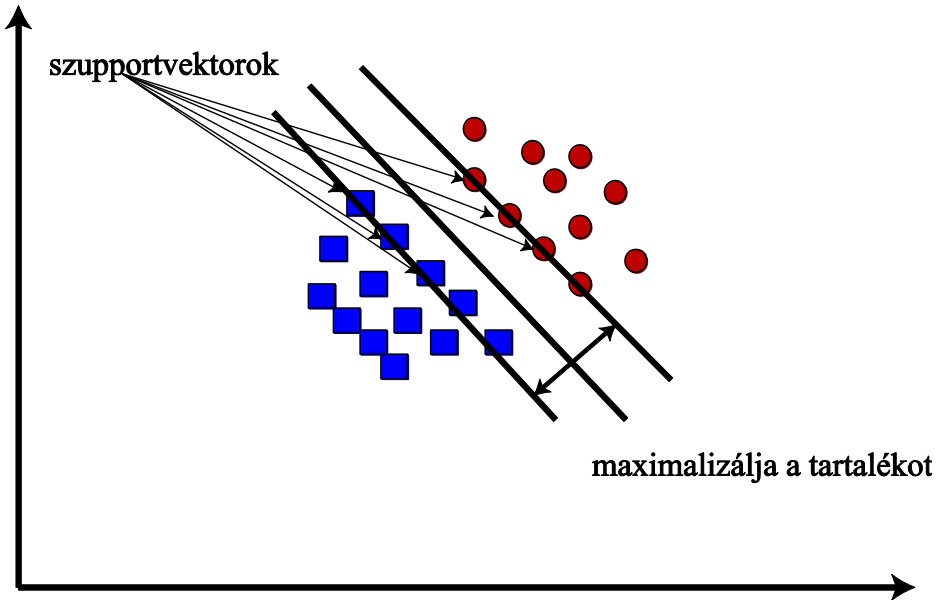
Az SVM olyan matematikai konstrukció, amelyet döntési problémák megoldásához szoktak alkalmazni. Alapverziója a lineáris osztályozók családjába tartozik, de bináris osztályozási problémák megoldására alkalmas. A többi lineáris osztályozóhoz képest az a fő ismérve, hogy nemcsak egyszerűen olyan hipersíkot (más néven vágási síkot) keres, amely elválasztja a pozitív és a negatív tanítómintákat, hanem ezek közül a legjobbat kutatja, vagyis intuitíve azt, amelyik a két osztály mintái között éppen „középen” fekszik (5.18. ábra).



5.18. ábra

Lehetséges hipersíkok lineárisan szeparálható adatok esetében

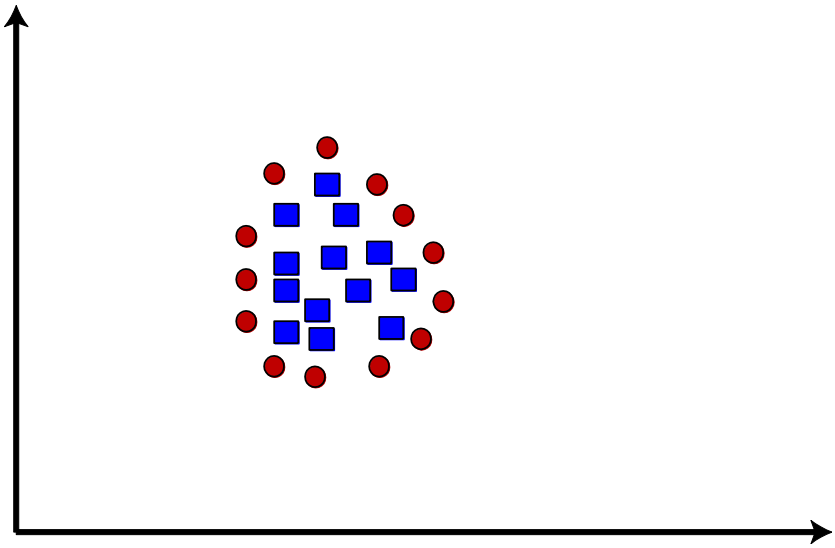
Az SVM tehát olyan döntési hipersíkot határoz meg, amely maximalizálja a tartalékot, azaz a hipersík és a hozzá legközelebbi pozitív és negatív tanítóadatok közti eltérést. Ezeket a tanítóadatokat szupport vektoroknak nevezzük. A hipersík meghatározásában a tanítóadatok közül csak a szupport vektorok 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 szeparáló sáv 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 (5.19. ábra).



5.18. ábra

Két osztály, amely egy hipersíkkal elkülöníthető: lineárisan szeparálható eset

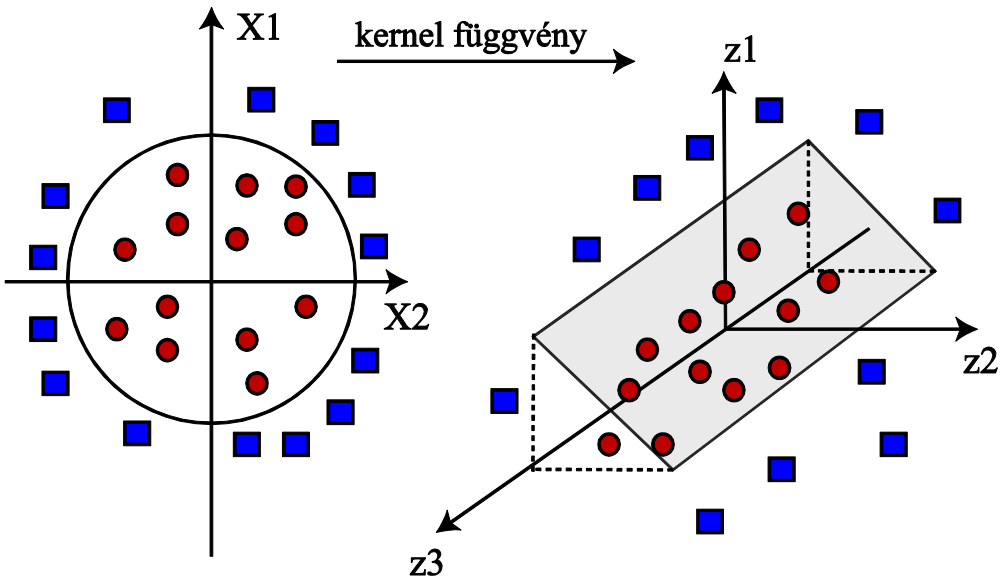
A valóságban azonban a legtöbb probléma nemlinearitása olyan nagyságrendű, hogy az osztályozó nem lesz hatékony (5.19. ábra).



5.19. ábra

Két osztály, amely egy hipersíkkal nem különíthető 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 (5.20. ábra).



5.20. 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 Mahine, 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; Chih-Chung–Chih-Jen 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, Chang és Lin (2003) szerint a C és γ értékeket az alábbi tartományokban érdemes keresni:

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

5.3.5.1.1. 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 tanítóhalmaz 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.

Ahhoz, hogy az SVM-et alkalmazni tudjuk, először az adatokat azonos dimenziójúra kell hoznunk. Mivel nem minden audioszegmens ugyanolyan hosszúságú, ezért a bemenő jellemzővektorok dimenziója nem egyenlő. Ennek kiküszöbölésére az egyes audioszegmensek kereteire statisztikai jellemzőket számolunk (átlag és szórás).

Az SVM bemeneti vektora tehát (i) a spektrumra: 2*80; (ii) az MFCC-re: 2*39; (iii) az MSL-re 2*39 (iv) részsávenergiára: 2*12.

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

5.3.6. Eredmények

Az egyszerre beszélések tehát 12%-át teszik ki a teljes korpusznak, míg a szünetek 10,9%-a, így a beszédrészek 77,1%-a a teljes korpusznak. Az átlagos átfedőbeszéd-arány 21,84% volt a korpuszban.

A jelen kutatásban teszteltük, hogy a 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 (5.3. táblázat), hogy a négy akusztikai paraméter (FFT spektrum (SP); Mel-frekvenciás kepsztrális (MFC) 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%.

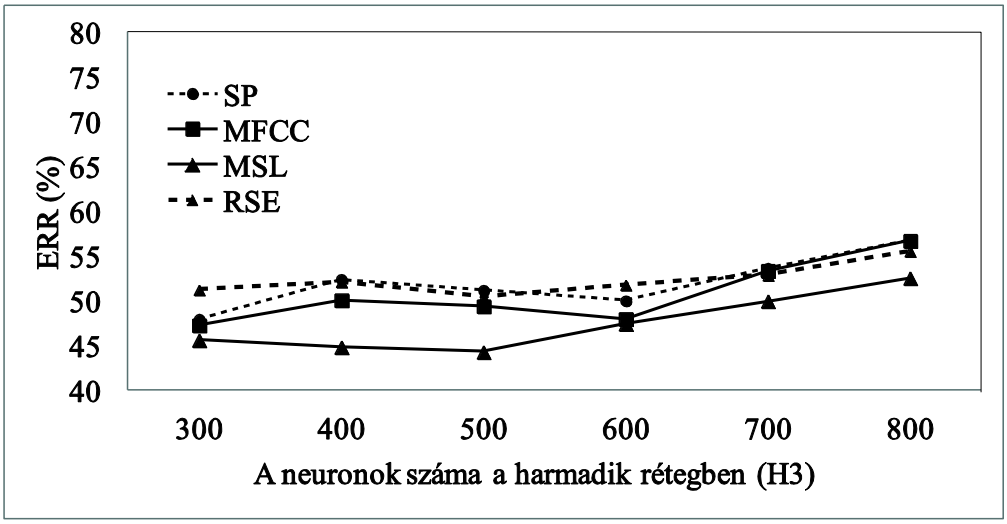
5.3. táblázat

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 hiba csökkenést tudtunk elérni a MSL jellemző alkalmazásával az MFCC-vel elér 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 H3-ban (5.21. ábra).



5.21. ábra

Az EER értéke a jellemzők és a H3-ban lévő neuronok számának függvényében

A statisztikai elemzések alátámasztják, hogy a MSL szignifikánsan jobban teljesít attól függetlenül, hogy hány neuront használunk a harmadik rétegben (5.4. 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.

5.4. táblázat

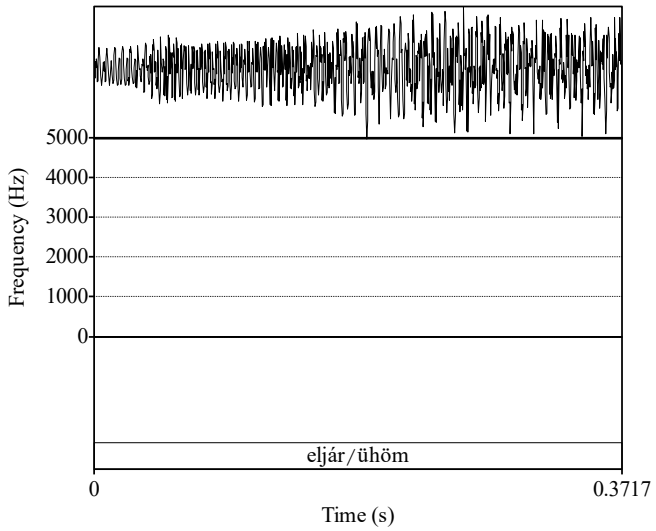
Az EER értéke a jellemzők és a H3-ban alkalmazott neuronok számának függvényében

		Akusztikai jellemzők			
	Neuronok száma a H3 rétegben	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 azt látszik, hogy két esetben (SP és MFCC) akkor volt a legkisebb a hiba értéke, ha a harmadik rétegben 300 neuront használtunk. Az MSL és a

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.

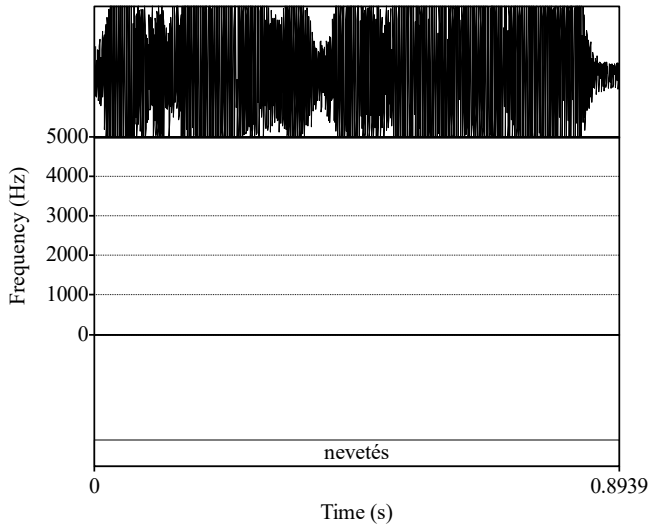
Az elért eredményeinket visszaellenőrizve elemeztük a hibák tulajdonságait. Az első és legnagyobb hibaforrás maga a kézi címkézés volt. Az egyszerre beszélések címkézése ugyanis sokszor igen nehéz feladat. A második hibaforrásként a háttércsatorna-jelzésekre vezethető vissza, a legtöbb hibát, 38,28%-ot ezek okozták. Ez a nagyszámú hiba annak tudható be, hogy a háttércsatorna-jelzések időtartama igen rövid (5.22. ábra), akár 60 ms-os is lehet, amely nem teszi lehetővé az elégséges számú jellemző kinyerését, így a belőlük származtatott statisztikai mutatók sem megbízhatók.



5.22. ábra

Háttércsatorna-jelzés (*ühüm*)

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 (5.23. ábra), hogy a nevetés közben az akusztikumban igen erős torzulás jelenik meg, a felvétel sokszor túlvezéreltté válik, így az akusztikai jellemzőkinyerés nehezítetté válik.



5.23. ábra

Példa a nevetésre mint nonverbális kommunikációs jelre

5.3.7. Következtetések

Az egyszerre beszélések magas, 12%-os az 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-score-t értek el az átfedő beszéd detektálására. Yella és Bourlard (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 indulnak 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 egy olyan algoritmust fejlesztett, amely ezt a jelenséget modellezi. Az általuk javasolt egyszerre beszélés detektálót beépítették 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 integrációs 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), amelyet igen hatékonyan alkalmaztak már más típusú problémák megoldására (Tang 2008).

A jelen kutatás során a legjobb eredményt a Mel-skála szerinti logaritmikus szűrőbank jellemző 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 szűrőbank jellemző 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 szűrőbank jellemzőt és H1(300)-H2(600)-H3(500) topológiájú DBN-t használtunk előfeldolgozásként, és 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%. Kimutattuk, hogy a mély neuronhálók alkalmasak a jelen problémában a jellemzők kialakításában nemellenőrzött tanulási folyamattal.

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 VAD 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). Ebben az esetben az EER értéke jóval alacsonyabb lehet. Az egyszerre beszélés és egyéb események esetleges elkülönítése további osztályozással is megvalósítható, erre azonban jelen munkában nem térünk ki.

Az átfedőbeszéd-osztályozót a beszélődetektáló rendszerünkbe integráljuk, és teszteljük, hogy a DER csökkenthető-e ezáltal.

5.4. Automatikus beszélődetektálás

5.4.1. Bevezetés

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 korpuszt 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.

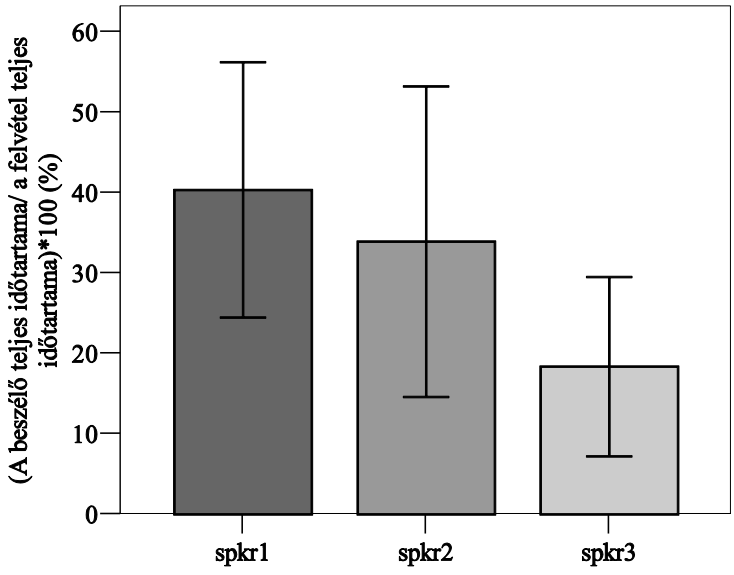
5.4.2.A vizsgálat anyaga

Száz társalgást használtunk a vizsgálatban; minden egyes beszédfelvételt manuálisan annotáltunk a beszélőváltások szerint. Minden társalgásban hárman vettek részt, ezért a jelölés beszélőnként: *spkr1*; *spkr2*; *spkr3*.

5.4.3. Az adatbázis leíró statisztikái a beszélődetektálás szempontjából

Az általunk random kiválasztott 100 társalgásban 7827 db beszédforduló volt. Egy felvételre átlagosan 70 db beszédforduló jut, amelynek szórása 41 db. A 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 (egyszempontos ANOVA).

Megvizsgáltuk, hogy a társalgásokon belül az egyes beszélők a teljes időtartamra nézve hány százalékában 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 (5.24. 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 feladat jellege, ismertségi fok).



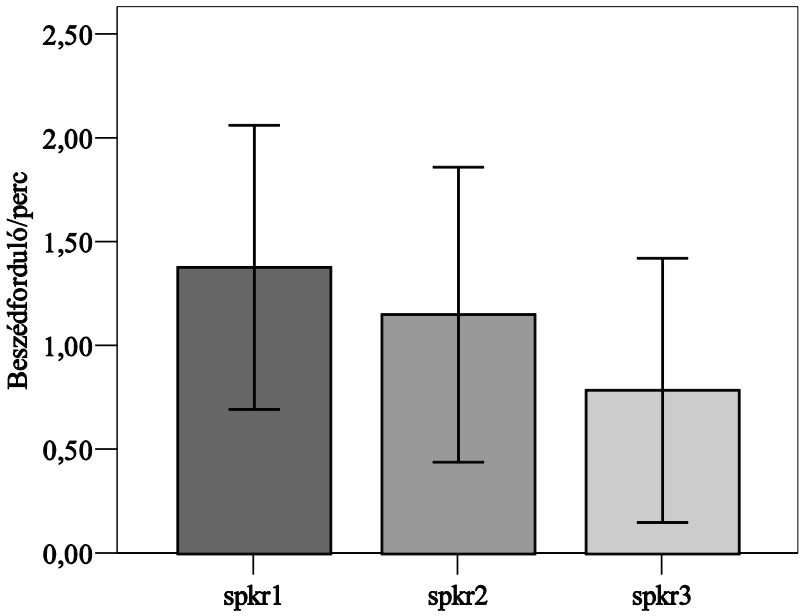
5.24. ábra

A beszélők teljes időtartama a felvétel teljes időtartamának függvényében

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. A férfiak átlagosan 37%-ban tartják magunknál a szót a teljes időtartamhoz képest, míg a nők 42%-ban. Azonban ez a különbség szintén nem szignifikáns (egytényezős ANOVA).

Továbbá kiszámoltuk, hogy az egyes résztvevőkre hány beszédforduló jut egy percre. Az adatközlőre átlagosan 1,38 beszédforduló jut egy percre, a felvételvezetőre 1,15, míg a harmadik személy esetében 0,78 (5.25. ábra). Ez szintén a társalgás részvevőinek aszimmetriáját mutatja.



5.25. ábra

Az egy percre jutó beszédfordulók száma a részvevők függvényében

Megvizsgáltuk, hogy a beszédidőtartamok és a beszédforduló/perc hogyan függnek össze az egyes részvevők 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 alaphelyzet az, hogy ő beszéljen, addig a felvételvezetőnek és a harmadik személynek igen, vagyis ahhoz, hogy minél hosszabb közléseket hozzanak létre, annál többször kell magukhoz venniük a szót.

5.4.4. Beszélőszegmentálás

5.4.4.1. Jellemzőkinyerés a beszélőszegmentáláshoz

Az általunk javasolt beszélőszegmentáláshoz a *Beszélőfelismerés a beszélődetektáláshoz* 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 2,5kHz és a 3,5kHz közötti részsáv az, amely a beszélőre vonatkozó akusztikai lenyomatokat tartalmazza. Az MFC-együtthatókat 32 ms-os ablakhosszra számoltuk, 10 ms-onként.

5.4.4.2. Bayesian Information Criterion (BIC: Bayes-féle Információs Kritérium)

A jelen munkában a Bayesian Information Criterion algoritmust használtuk a beszélők szegmentáláshoz. A BIC az egyik legtöbbször használt algoritmus a szegmentálásban, illetve a klaszterezésben, mivel számítása igen egyszerű és hatékony. A BIC a feltételes valószínűségszámítás alapjain nyugszik. A BIC-ben a modell kiválasztás úgy történik, hogy a valószínűségi kritérium értéke annál magasabb, minél magasabb a modell komplexitása, tehát bünteti a modell komplexitást (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, és amely a következőképpen definiálható:

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

Mivel a $logL(X_i,M_i)$ az adatok log-likelihood értéke (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 az $\#(M_i)$ 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 hozzák létre, és ahol N_i nagynak tekinthető.

Ahhoz, hogy a BIC-t 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 algoritmushoz hasonlóan –, amelyet a következőképpen számolunk:

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

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

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

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

$$\mathrm{P} = \frac{1}{2}(\mathrm{p} + \frac{1}{2}\mathrm{p}(\mathrm{p}+1)) \log(\mathrm{N})$$

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 $\Delta\#(\mathrm{i}, \mathrm{j})$ a különbség értéke a szabad paraméterekben a kombinált modell és a két különálló modell között.

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 (1998) 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 amellett érvel, hogy automatikusan kell a λ paramétert megválasztani (Tritschler–Gopinath 1999; Delacourt–Wellekens 2000; Delacourt–Kryze–Wellekens 1999a; Mori–Nakagawa 2001; Lopez–Ellis 2000; Vandecatseye et al. 2004).

Legyen *M1* és *M2* két modell:

- Az *M1* modell azt feltételezi, hogy *X* minden mintája független, és egyetlen többváltozós Gauss-szal leírható

$$Z = Z_1, Z_2, \dots, Z_N \sim \mathrm{N}(\mu_Z, \sum_Z)$$

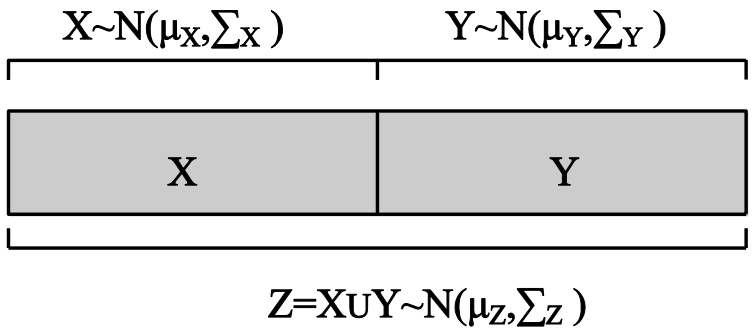
- Az *M2* modell esetében azt felételezzük, hogy *X* két többváltozós Gauss-szal írható le. Az egyik Gauss a szegmens elejétől *b* időkeretig, a másik Gauss a *b* időkerettől a szegmens végéig.

$$M_2: Z = X + Y$$

$$Z = Z_1, Z_2, \dots, Z_b \sim \mathrm{N}(\mu_X, \sum_X)$$

$$Z = Z_{b+1}, Z_{b+2}, \dots, Z_N \sim \mathrm{N}(\mu_Y, \sum_Y)$$

Ezen hipotéziseket a következő ábra szemlélteti (5.26. ábra):



5.26. ábra

Hipotetikus modellek egy keretre vonatkozóan a szegmentálásban

A fent leírt BIC-számítások alapján előáll *BIC(M1)*, *BIC(M2)*. A BIC-szegmentálás során a büntetőfaktor λ értékét 0-ra vesszük. Ekkor teszteljük a két hipotézisünket:

ha $\Delta BIC = BIC(M2) - BIC(M1) = 0$,

akkor ez azt jelenti, hogy a modellre érkező score alapján az adatokra jobban illeszkedik a két többváltozós Gauss-modell (*M2*), mint az egy többváltozós Gauss-modell (*M1*). Mindez azt jelenti, hogy a szegmens nem homogén, vagyis a szegmensben váltási pont van.

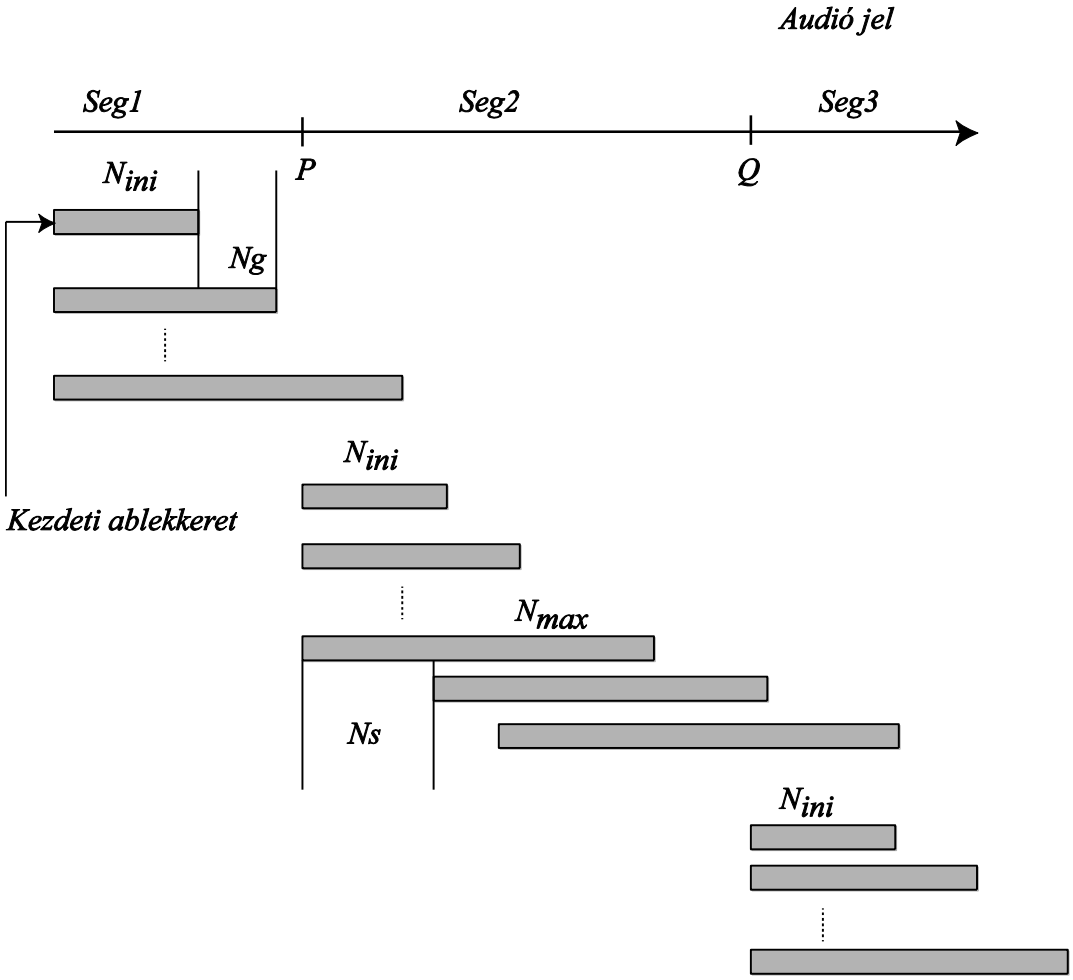
A ΔBIC a beszédben egy akusztikai váltási pont detektálására alkalmas. Nyílvánvalóan a társalgások során jóval több váltási pont létezik, ezért ennek megoldására szekvenciális detektáló algoritmust szokás használni (Cheng 2010). Ennek alapján a ΔBIC-et fel tudjuk írni, mint a *b* váltási pont függvényét.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

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

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 *M2* modell jobban leírja az adatokat, mint *M1* modell, és a váltási pont *b* tényleg valós.

5.4.4.2.1. Növekedő ablakhossz módszer a ΔBIC számításához

Ezt az eljárást a több váltási pont detektálására szokás alkalmazni. A 5.27. 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 a 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 2010). A hátránya ennek a folyamatnak, hogy ahogyan növeljük az ablakhosszt, sokkal nagyobb számítási kapacitásra van szükség.



5.27. ábra

A növekedő ablakolási eljárás sematikus ábrája (Cheng 2008)

5.4.4.2.2. 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 záró buffer mérete: 1 másodperc

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

5.4.4.3. 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 távolságalapú szegmentálási folyamattal vizsgáltuk felül a váltási pontokat (Siegler et al. 1997, Hung et al. 2000). 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 eltérését) a következőképpen tudjuk számolni:

$$KL(X,Y) = E_X(log\frac{P_X}{P_Y})$$

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

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

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

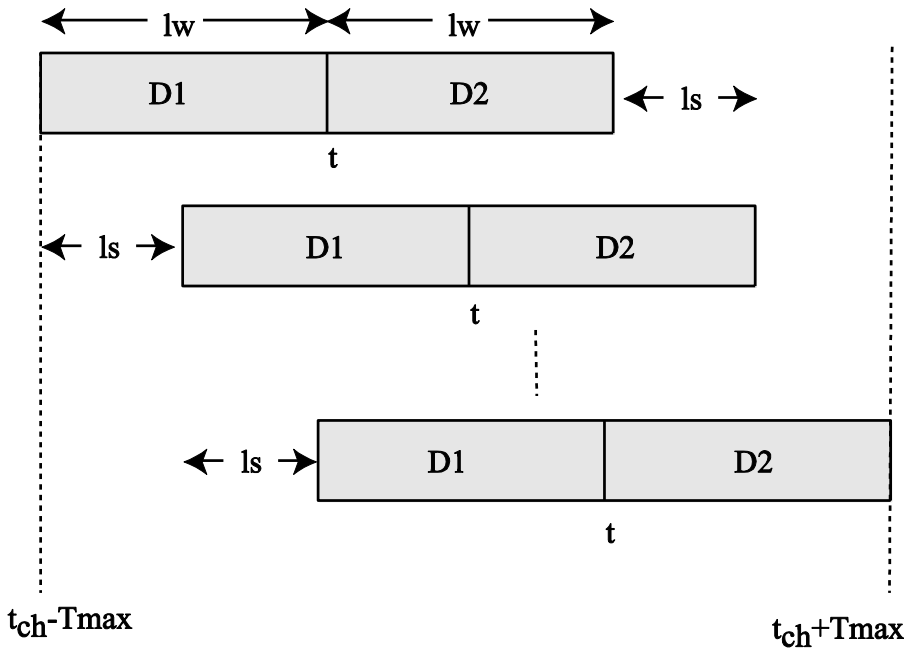
$$KL2(X,Y) = KL(X,Y) + KL(Y,X)$$

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 KL2-különbséget minden egyes időkeretre *t*, amelyet *ls* mérettel tolunk tovább, és amelyben a távolság ±*Tmax* 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 előredeifiniált faktor, amely kísérleti úton kell meghatároznunk (Ida 2011).

Mivel a beszélőszegmentálásból származó hiba 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 $\pm Tmax$ időkeretben számoljuk ki (5.28. ábra).



5.28. ábra

A KL2 utófeldolgozásának folyamata

5.4.4.3.1. 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 *Tmax*, amelyet a jelen munkában 200 keretnyire, vagyis 2 másodpercre állítottunk. A harmadikat, az α paramétert pedig 0,5-re állítottuk be.

5.4.5. Beszélőklaszterezés

5.4.5.1. Jellemzőkinyerés a beszélőklaszterezéshez

Az általunk javasolt beszélőszegmentáláshoz a *Beszélőfelismerés a beszélődetektáláshoz* fejezetben bemutatott MFCC-eljárást használtuk mint jellemzőkinyerő algoritmust. 12 dimenziós MFCC-vektor nyertünk ki 10 ms-onként 32

ms-os Hamming ablakoló függvénnyel. A határsávértékek a Mel-szűrő skálához először a teljes spektrumra, majd 2,5 kHz és a 3,5 kHz közöttire állítottuk, amely a beszélőre vonatkozó akusztikai lenyomatokat tartalmazza. 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.

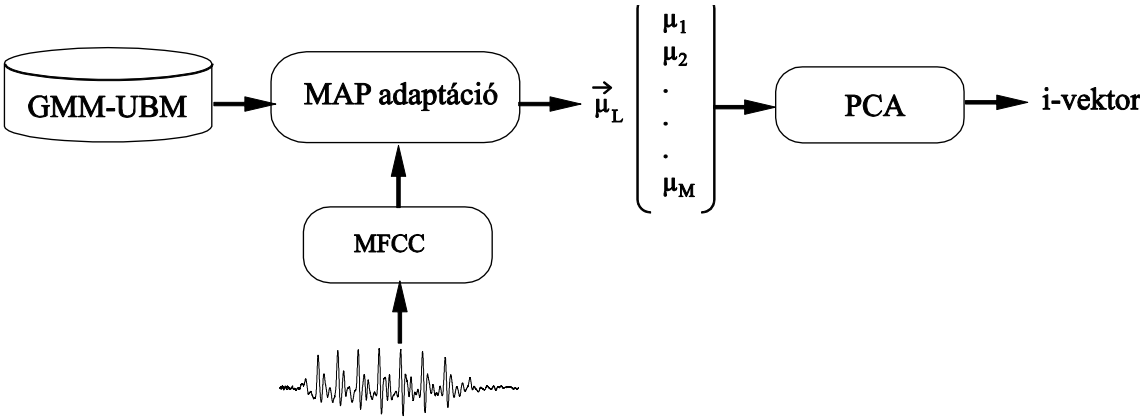
A beszélőklaszterezés során a beszélőszegmentálóból érkező szegmensek a bemenet, vagyis a két beszélőváltás között lévő beszédjel. Ezen beszédjelek modellezésére a kinyert akusztikai jellemzőkből GMM-szupervektorokat képeztünk.

5.4.5.2. GMM-szupervektor

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

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

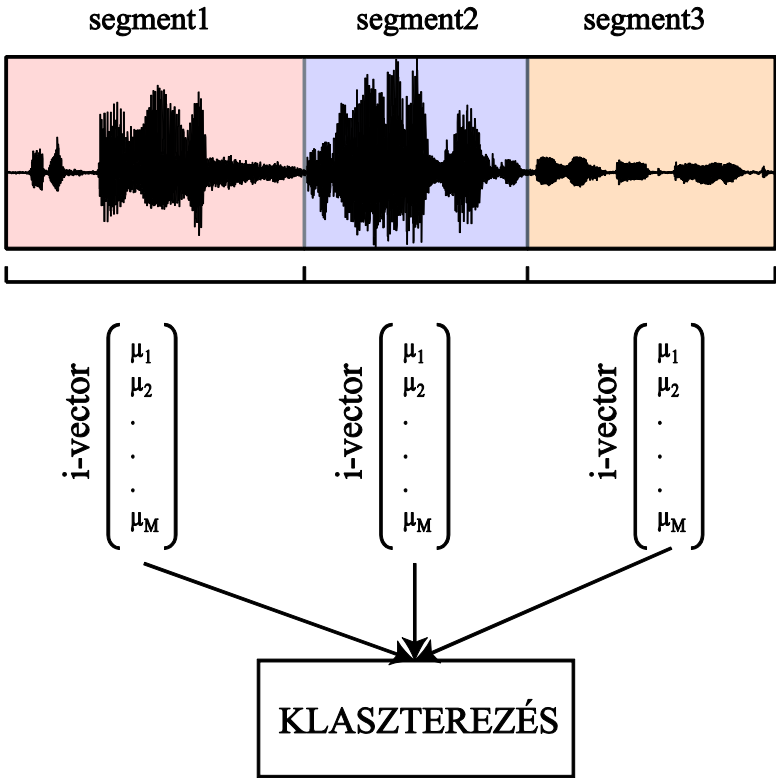
ahol λ_i a keverék súlyok, *N()* Guass komponens, és m_i és $\sum_i$ az középértéke és a kovarianciája a Gauss-eloszlásnak. A jelen kutatás során diagonális kovarianciájú GMM-et használtunk. Adott egy beszédszegmens a beszélőszegmentálásból, a GMM-UBM tanítása a középértékeknek MAP-adaptálásával hajtódik végre (Reynolds et a. 2000), m_i. Ezen adoptált középértékeket fűzzük össze GMM-szupervektorrá (5.29. ábra). A jelen kutatásban a GMM-UBM 256 kevert komponenst tartalmaz.



5.29. ábra

A jellemzőkinyerés sematikus blokk diagramja

Mivel így egy magas dimenziószámú jellemzővektort kapunk, ezért a jelen dolgozatban a dimenziószámot lecsökkentettük PCA-val. A dimenziócsökkentett jellemzővektor jelen esetben az i-vektror. Az i-vektor a bemenete a nemellenőrzött tanuláson alapuló beszélőklaszterezésnek (5.30. ábra).



5.30. ábra

Az i-vektor mint bemenet a beszélőklaszterezéshez

5.4.5.3. 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 milyen (i) metrikát használjuk 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) $x = \{x_1, x_2, \ldots, x_M\}$ és $y=\{y_1, y_2, \ldots, y_M\}$. 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 $_{Cz,}$ ahol $z = x \cup y$.

A fenti egyenletben a λ (teoretikusan 1) hangoló paraméter, N_{H0} és N_{H1} 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.

5.4.6. Eredmények

Ebben a fejezetben az általunk kialakított beszélődetektáló rendszert teszteljük különféle beállítások mellett. A beszélődetektálás működésének kiértékelésekor nagy különbségek lehetnek a tesztelésre használt korpusz függvényében. Erre az RT04f workshopon hívták fel a figyelmet (NIST Fall Rich Transcription Evaluation website 2006). 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 *Beszélődetektálás kiértékelése* című fejezet), amellyel minden tesztfájlra meghatároztuk a DER-értéket (*diarization error rate*).

A BEA-adatbázisból 12 társalgást választottunk ki random módszerrel. A 12 társalgás összidőtartama közel 2,8 óra. A 2,8 órányi társalgásban 480 beszédváltás történt (5.5. táblázat).

5.5. táblázat

A beszédfordulók száma és teljes időtartama az egyes tesztfájlokra

Felvétel sorszáma	Beszédfordulók száma (db)	Teljes időtartam (s)
bea071n037	55	919,528
bea072n038	46	1020,403
bea073n039	23	590,546
bea074n040	25	1053,364
bea075n041	16	887,631
bea094f039	31	799,54
bea150n091	32	769,777
bea166f066	50	982,415
bea174n105	46	773,054
bea184n111	48	599,437
bea189n114	68	973,18
bea192f077	50	816,264

5.4.6.1. A standard BIC beszélődetektáló kiértékelése különböző akusztikai jellemzők esetében

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. Amely azt jelenti, hogy 60,56%-ban helyesen szegmentál és klaszterez az alap kiinduló algoritmusunk.

Az előzetes kísérleteink szerint ha az MFCC-jellemzőt specifikusan a 2,5–3,5kHz-es részsávra számoljuk, akkor az eredmények javíthatók, hiszen az eredményeink szerint

ezen frekvenciatartomány tartalmazhatja a beszélőre specifikus akusztikai lenyomatokat. Ezért a standard BIC beszélődetektálóba ezt az akusztikai jellemzőt használtuk, mint a standard beszélődetektáló módosítását. Az eredmények (5.6. 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).

5.6. táblázat

A standard BIC beszélődetektálóval elért eredmények

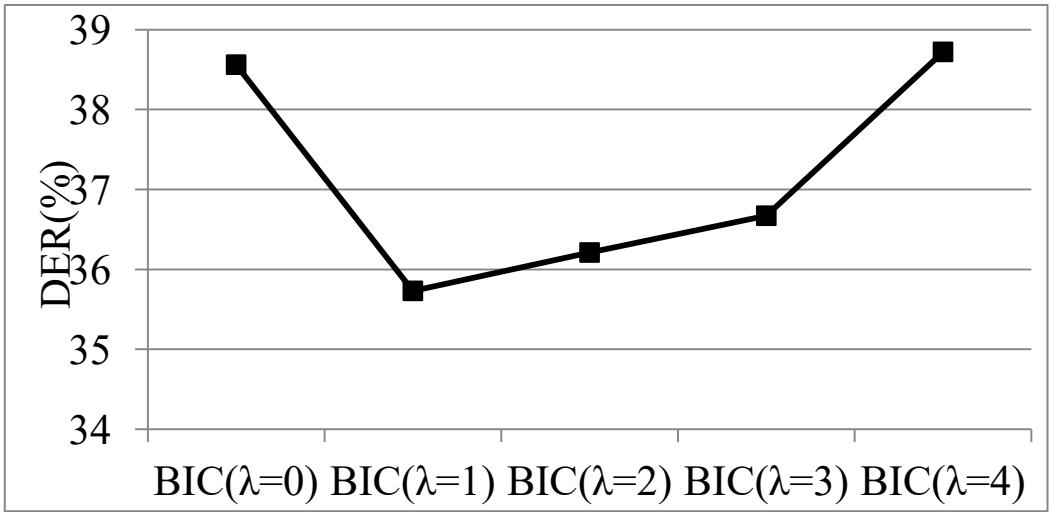
Felvétel sorszáma	**BF száma**	**Teljes időtartam**	**DER BIC(λ=0)**		**ΔDER**
			MFCC	***$MFCC_{(2,5-3,5)}$***	
bea071n	55	919,528	22,84%	21,21%	-1,63
bea072n	46	1020,403	35,92%	34,28%	-1,64
bea073n	23	590,546	43,38%	41,53%	-1,85
bea074n	25	1053,364	30,02%	30,33%	0,31
bea075n	16	887,631	41,25%	39,81%	-1,44
bea094f	31	799,54	39,25%	39,59%	0,34
bea150n	32	769,777	44,55%	42,63%	-1,92
bea166f	50	982,415	36,38%	34,14%	-2,24
bea174n	46	773,054	49,45%	47,48%	-1,97
bea184n	48	599,437	46,21%	44,46%	-1,75
bea189n	68	973,18	48,37%	46,03%	-2,34
bea192f	50	816,264	42,6%	41,24%	-1,36
Átlagos	**490**	**10185,14**	**39,43%**	**38,56%**	**-0,869**

A legjobb eredményt a bea071-es felvételére kaptunk, amely időtartamában és beszélőfordulók számában is magasnak mondható. A legrosszabbat pedig bea174-es

felvételre. A bea071-es felvételében egy idős nő az adatközlő, így hangja hallható módon különbözik a felvételvezetőétől, illetve a harmadik személyétől. A bea174-es felvételen pedig egy fiatal felnőtt nő az adatközlő, akinek hangszíne igen közeli a felvételvezetőéhez.

5.4.6.2. A BIC λ paraméterének optimális megválasztása

Teoretikusan a λ büntetőfaktor értéke zéró, amely 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. Az 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 hibaaránya 35,73% (5.31. ábra).



5.31. ábra

A DER értékének alakulása a BIC λ paraméterének függvényében

5.4.6.3. A beszéddetektáló implementációja a beszélődetektálóba

Számos kutatás kimutatta, hogy a beszéddetektáló implementációja beszélődetektálásba jelentősen csökkenti a DER értékét (Wei 2008). Ezért a jelen kutatásban az általunk létrehozott VAD-algoritmust implementáltuk a beszélődetektálóba. Az eljárás lényege, hogy a VAD által detektált szünetrészeket már

nem továbbítottuk a beszélődetektáló felé, vagyis töröltük felvételből. Tehát jelen esetben a VAD-ot mint előfeldolgozó egységként csatoltuk a beszélődetektáló elé.

Az eredmények azt mutatják (5.7. táblázat), hogy a VAD előfeldolgozásával az DER értékét átlagosan 4,535%-al tudtuk csökkenteni. Ez az átlagos javulás statisztikailag igazolható (Wilcoxon teszt Monte Carlo szimulációval kiegészítve: Z=-3,059; p<0,001).

5.7. táblázat

A DER értéke beszéddetektáló nélkül és beszéddetektálóval

Felvétel sorszáma	DER		ΔDER
	VAD nélkül BIC(λ=1)	VAD-ot használva	
bea071n037	18,6%	14,98%	-3,62
bea072n038	31,86%	27,83%	-4,03
bea073n039	39,94%	35,64%	-4,3
bea074n040	26,69%	23,89%	-2,8
bea075n041	38,14%	34,21%	-3,93
bea094f039	36,19%	33,63%	-2,56
bea150n091	39,63%	36,26%	-3,37
bea166f066	31,7%	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
Átlagos	**35,73%**	**31,21%**	**-4,535**

5.4.6.4. Az egyszerrebeszélés-detektáló integrálása 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 alfejezetbe ennek a lehetőségét kívánjuk megvizsgálni, ezért az egyszerrebeszélés-detektálót implementáltuk az általunk létrehozott beszélődetektálóba. Hasonlóan a VAD-hoz, 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 VAD egység után.

Az átfedő beszédek automatikus detektációjával átlagosan 2,49%-os relatív javulást tudtunk elérni, vagyis a DER értékét 31,21%-ról le tudtuk csökkenteni 28,713%-ra (5.8. táblázat). Ez a javulás szignifikáns (Wilcoxon teszt Monte Carlo szimulációval kiegészítve: Z=-3,06; p=0,002).

5.8 táblázat

A DER értéke az átfedő beszéddetektáló nélkül és az átfedőbeszéd-detektálóval

Felvétel sorszáma	DER		ΔDER	Átfedő beszéd és a társalgás hosszának aránya
	Átfedő beszéd			
	tartalmaz	nem tartalmaz		
bea071n037	14,98%	12,357%	-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%
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%
Átlagos	**31,21%**	**28,713%**	**-2,49%**	**30,94%**

Elemeztük, hogy a teszteléskor használt társalgásokban milyen arányban fordulnak elő egyszerre beszélések (14.4. táblázat). A táblázatban látható, hogy elég gyakoriak az átfedő részek ezen felvételekben. Jóllehet az egyszerre beszéléseket detektáló algoritmus eredményei nem voltak túl magasak, mégis statisztikailag igazolható relatív javulást tudtunk elérni a beszélődetektálóba való implementációjával.

5.4.7. Következtetések

A jelen fejezetben az általunk javasolt beszélődetektáló felépítését és működését mutattuk be. A létrehozott beszélődetektálót a BEA-adatbázisból random kiválasztott társalgásokon teszteltük. Ebben a fejezetben integráltuk az egyes fejezetekben

bemutatott különálló blokkokat, fejezeteket. Bemutattuk, hogy az egyes fejezetek, különálló blokkokban kikísérletezett eredményeket hogyan használhatjuk fel a beszélődetektálásban.

A *Beszélőfelismerés a beszélődetektáláshoz* című fejezetben bemutattuk, hogy ha az MFCC jellemzőkinyerést 2,5 kHz és 3,5 kHz-es részsávban nyerjük, akkor a beszélőszemély felismerésének eredménye növelhető. 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 teljes spektumra számolva. A $MFCC_{(2,5–3,5)}$ jellemzővel 38,56%-os DER-értéket kaptunk, amely átlagosan 0,869%-os DER-javulást eredményezett.

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 beszélődetektálási hibaaránya 35,73%-os volt.

A *Beszéddetektálás* című fejezetben létrehozott VAD-ot implementáltuk a beszélődetektálóba. Az eredmények azt mutatták, hogy a VAD előfeldolgozásával az DER értéke átlagosan 4,535%-kal csökkenthető.

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

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 egyszerre beszéléseket detektáló algoritmusokat. Ekkor a DER értéke 28,71% volt.

6. KÖVETKEZTETÉSEK

A jelen kutatás fő célja az volt, hogy magyar nyelvre elsőként hozzon létre spontán társalgásokra nemellenőrzött 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 a képi feldolgozásban használt mély neuronhálókat alkalmazni az egyszerrebeszélés-detektáló jellemzőkinyeréseként. Elemeztük, hogy a beszélőszegmentálásban milyen beállítások mellett kapjuk a legjobb eredmény.

Az értekezésben részletezett kutatássorozat legfőbb eredménye, hogy magyar nyelvű spontán társalgások beszédfordulóinak automatikus detektálását valósított meg. Kísérleti úton bizonyítottuk, hogy a beszélőspecifikus akusztikai jellemzőkkel csökkenthető a beszélődetektálási hiba értéke. Igazoltuk továbbá, hogy a mély neurális hálók alkalmazhatók az egyszerre beszélések automatizált detektálásában. Kísérletekkel, illetve az adatfeldolgozással bizonyítottuk azt a felismerést, hogy a beszéddetektálónak és az egyszerre beszélés detektálónak a beszélődetektálóba történő integrálása csökkenti a DER értékét.

Az egyes kísérletek következtetéseit külön alfejezetekben ismertetjük (lásd 6.1-től a 6.4-ig).

Az eredmények alapján az alábbi elméleti konklúziókat fogalmazhatjuk meg a spontán társalgások sajátosságaira vonatkozóan. Noha adataink elsődlegesen a BEA hármas társalgásaira jellemzők, a nagy mennyiségű elemzett anyag feljogosít arra, hogy óvatos általánosításokat fogalmazzunk meg. Egy kb. 30 perces társalgásra átlagosan 70 beszédforduló jut. Az a tény, hogy az adatközlő nő-e vagy férfi, nem mutatott összefüggést a beszédfordulók számával. Megállapíthatjuk tehát, hogy a beszélő neme nem meghatározó a beszédfordulók számát tekintve. A társalgásban résztvevők beszédideje alapján jól kirajzolódó szerepviszonyokat határozhattunk meg. A

legfontosabb szerepe az adatközlőnek van, hiszen az ő beszédének rögzítése a cél az adatbázisban, amelyet megerősít az az objektív tény is, hogy a teljes felvételi idő több, mint egyharmadában a mindenkori adatközlő tartja magánál a szót. A felvétel elkészítésében értelemszerűen igen fontos szerepet jut a felvételvezetőnek, aki az adatközlőhöz hasonló arányban tartja magánál a szót. Kettőjükkel ellentétben a harmadik beszélő mintegy háttérbe szorul, amit alátámaszt, hogy a felvételi idő csupán 18%-ban szólal meg, átlagosan. Mindezekből az a következtetés vonható le, hogy a társalgásban a szerepek nem kiegyenlítettek, egy-egy személy sokszor háttérbe szorul (ennek oka többféle lehet, például feladat jellege, ismertségi fok), és ezt a helyzetet minden beszélő természetesnek, a társalgás velejárójának ítéli.

Vizsgáltuk a beszédidőtartamok és a beszédforduló/perc összefüggését az egyes résztvevők függvényében. Az eredmények azt mutatták, hogy míg az adatközlőnek nem kell törekednie a szóátvételre, hiszen az alaphelyzet az, hogy ő beszéljen, vagyis ez az elvárás vele szemben, addig a felvételvezetőnek és a harmadik személynek ahhoz, hogy minél hosszabb közléseket hozhassanak létre, többször kell magukhoz venniük a szót.

Az automatikus beszéddetektáló használatával bármilyen társalgás vagy párbeszéd vizsgált jellemző objektív paraméterekkel alátámaszthatók, és ez a diskurzuselemzés számára fontos információk alkalmazását teszi lehetővé.

6.1. Beszéddetektálás

Ebben a vizsgálatban Giannakopoulos (2009) által kidolgozott és MATLAB-ba implementált beszéddetektáló algoritmusát használtuk, illetve módosítottuk. Ez az algoritmus rövid idejű energiafüggvény (short-term energy) és spektrális centroid (spectral centroid) akusztikai jellemzőket és adaptív küszöbölést alkalmaz a beszéd és nem-beszéd szegmensek automatikus meghatározására. Az általunk ajánlott módszer annyiban tér el ettől, hogy a küszöb meghatározását (beszéd és nem beszéd) nemellenőrzött 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 nemellenőrzött tanulási módszer javít-e az eredményeken.

100 társalgásban (55 órányi társalgást jelent) manuálisan jelöltük azokat a részeket, ahol valamelyik adatközlő beszél, illetve azokat a részeket, ahol nincs beszédjel, vagyis némaszünet van. A korpusz 49 órányi beszédrészt és 6 órányi szünetet tartalmaz, vagyis a teljes korpusz 10,9%-át a szünetek teszik ki. A VAD kiértékelését 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. 10%-os jel/zaj arányig még közel 65,28%-os eredménnyel, 5%-os jel/zaj aránytól viszont már csak 38,8%-os helyes találati aránnyal működik a rendszer. Ez azzal magyarázható, hogy a VAD-algoritmusban nem használtunk zajszűrőt. Ezért tervezzük, hogy zajszűrőkkel is kísérletezni fogunk. Az elkészített VAD egy általunk fejlesztett beszélődetektálóba lesz integrálva, ami feltehetőleg javítani fogja annak működését.

6.2. Beszélőfelismerés a beszélődetektáláshoz

A kutatás egyik célja az volt, hogy megvizsgálja, a magyar nyelvű beszédben mely spektrális régiók beszélőspecifikusak. A második célja az volt, hogy a beszélőket MFC-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 kívánjuk integrálni.

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 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 kutatás során elkészítettünk egy férfi/nő osztályozó algoritmust is, vagyis egy nem szerinti osztályozót. Az osztályozáshoz akusztikai jellemzőként 13 koefficienst tartalmazó MFCC mellett, 24 koefficienst tartalmazó MFCC-t is használtunk. A legjobb eredményt a 24 együtthatót tartalmazó MFCC-t használó, GMM-UBM-et és 16 Gauss komponenst tartalmazó osztályozóval kaptuk. Ennek eredménye 87,01%-os volt. Az eredményeink alátámasztják azt az elképzelést, hogy a magasabb kepsztrális együtthatók őrzik a nemre utaló akusztikai jegyeket.

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. Az 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.

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

6.3. Az egyszerre beszélések automatikus osztályozása

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 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 annál több részvevő szólal meg egyszerre, vagyis ahol átfedő beszéd van. Ezen intervallumok összidőtartama közel 7 óra, amely a teljes korpusz 12%-a.

A jelen kutatásban egy ANN/SVM hibrid rendszer (Artificiel Neural Network/Support Vector Machine: Mesterséges Neuron Hálók/Szupport Vektor Gépek) segítségével hoztunk létre az egyszerre beszélések automatikus osztályozását. Mivel nem ismert, hogy mely akusztikai paraméter mentén különülnek el az átfedő beszédrészek és a nem átfedő beszédrészek, több akusztikai jellemzőt is teszteltünk, mint például az FFT spektrum, részsáv-energia (subband-energy), MFCC. A jellemzők jobb reprezentálásához főkomponens-analízist (PCA) használtunk. A jelen kutatásban a mély neuronhálókat az akusztikai jellemzők előfeldolgozásához használtuk. A tényleges osztályozást LS-SVM-el végeztük el, amely az SVM egyik változata.

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

Teszteltük azt is, hogy a 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 (az EER értéke 44,33%), ha Mel-skála szerinti logaritmikus szűrőbank jellemzőt, és H1(300)-H2(600)-H3(500) topológiájú DBN-t használtunk előfeldolgozásként, valamint SVM-RBF-et osztályozóként.

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.

6.4. 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 beszédfordult adatoltunk. Egy felvételre átlagosan 70 db beszédforduló jut, amelynek szórása 41 db. A legtöbb beszédforduló 240 db volt, míg a legkevesebb 11 db. A nemek tekintetében nem találtunk szignifikáns különbséget a beszédfordulók számában (férfi adatközlő átlagosan 79 db beszédfordulót produkált, míg női 65 db-ot). A társalgásban betöltött szerepek 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%-ba. Ezek az arányok azt mutatják, hogy a társalgások során a szerepek nem kiegyenlítettek, a harmadik személy sokszor háttérbe szorul (ennek oka többféle lehet, pl. ismertségi fok). Az beszédidőtartamban sem tudtunk szignifikáns különbséget kimutatni a nemek között (férfiak 367%, nők 42%-ban tartják maguknál a szót a teljes időtartamhoz képest). Megvizsgáltuk, hogy a beszédidőtartamok és a beszédforduló/perc hogyan függnek össze az egyes részvevők 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 az alaphelyzet az, hogy ő beszéljen, míg a felvételvezetőnek és a harmadik személynek ahhoz, hogy minél többet hosszabb közléseket hozzanak létre, annál többször kell magukhoz venniük a szót.

A beszélődetektálón belül a beszélőszegmentálást a Bayesian Information Criterion (BIC: Bayes-féle Információs Kritérium) algoritmust használtuk. Akusztikai jellemzőként az MFCC-t kétféleképpen használtuk. Az első megközelítésben a teljes spektrumra kiszámoltuk. A másodikban pedig részsávra; kimutattuk ugyanis, hogy a 2,5kHz és a 3,5kHz közötti részsáv az, amely a beszélőre vonatkozó akusztikai lenyomatokat tartalmazza. Az MFC együtthatókat 32 ms-os ablakhosszra számoltuk, 10 ms-onként. A téves riasztások kezelésére egy utófeldolgozó lépést használtunk, amely Kullbak–Leibler távolságon alapul. A beszélőklaszterezéshez szintén a BIC algoritmust

használt mind a klaszterek közötti hasonlóság mérésére, mind megállási feltételként. A beszélőklaszterezésben GMM-szupervektort PCA transzformáltját használtuk mint a beszélőklaszterezésnek bementi jellemzőjét.

Ebben a fejezetben az általunk javasolt beszélődetektáló felépítését és működését mutattuk be. A létrehozott beszélődetektálót a BEA-adatbázisból random kiválasztott társalgásokon teszteltük. Ebben a fejezetben integráltuk az egyes fejezetekben bemutatott különálló blokkokat, fejezeteket. Leírtuk a különálló blokkokban kikísérletezett eredményeket hogyan használtuk fel a beszélődetektálásban.

A *Beszélőfelismerés a beszélődetektáláshoz* című fejezetben bemutattuk, hogy ha az MFCC jellemzőkinyerést 2,5 és 3,5kHz-es részsávban nyerjük, akkor a beszélőszemély felismerés eredménye növelhető. 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.

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 hiba aránya 35,73% volt.

A *Beszéddetektálás* c. fejezetben létrehozott VAD-ot implementáltuk a beszélődetektálóba. Az eredmények azt mutatták, hogy a VAD előfeldolgozásával az DER értéke átlagosan 4,535%-kal csökkenthető.

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

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 egyszerre beszélés detektáló algoritmusokat. Ekkor a DER értéke 28,71% volt.

7. ÖSSZEGZÉS

A beszédtudomány alapvető kutatási célja a beszédkommunikáció komplex körfolyamatának leírása. A beszédtechnológiában a beszédkommunikáció egyes moduljainak a 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 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 történt 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 értékezés ezt a sokszínűséget kívánta bemutatni, rendezni és felhasználni a beszélődetektálás megvalósításában.

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

8. TOVÁBBI TERVEK

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 létrehozását, amellyel a beszédfordulók egy része egyértelműsíthető lenne, csökkentve ezzel a téves riasztások számát.

Tervezzük továbbá, hogy az általunk kidolgozott rendszert más standard korpuszokon is teszteljük, így a rendszerünk összevethető más, már létező beszélődetektáló algoritmusok eredményeivel.

8.1. A beszélődetektálás felhasználási területei

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

Mindezek mellet a beszélődetektálás fontos szerepet játszhat a dokumentum visszakeresésben, tartalomkinyerésben vagy a kérdés-válasz rendszerekben.

Az ilyen fajta megközelítés új ismereteket nyújthatnak a társalgások felépítéséről, és a társas viszonyokról.

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

A kutatás eredményei hozzájárulnak az emberi viselkedés megértéséhez, illetve tovább muttanak az ember–gép kommunikáció gépi modellezése felé.

9. A DISSZERTÁCIÓ TÉZISEI

Az eredményeink alapján öt tézist fogalmaztunk meg; jellegüktől függően rövidebben vagy hosszabban kifejtve.

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

Az újdonság a kutatásban az, hogy a magyar nyelvű spontán társalgásokban a standard BIC beszélődetektálóval, MFCC teljes spektrumot lekódoló jellemzőt használva, λ paraméterét 0-ra állítva, sem szünetmodellt, sem egyszerrebeszélés-modellt nem használva 39,43%-os DER eredménnyel működik.

II. tézis: *Igazoltuk, hogy a beszédfelismerésben a spektrum célzott részsávjára (2,5–3,5kHz) történő akusztikai jellemzőkinyerés jobb eredményeket ad, mint a teljes spektrumot lekódoló eljárások.*

Jóllehet az akusztikai jellemzők közül az MFC (Mel Frequency Cepstral) együtthatók az egyik legtöbbet használt paraméteregyüttes, továbbra is kérdés maradt, hogy a spektrumban mely frekvenciasáv tartalmazza a beszélőspecifikus jegyeket. A beszélőfelismerésben az MFCC együtthatókat először a teljes spektrumra számoltuk (full-band spectral based MFCC). Más megközelítésben az MFCC együtthatókat a spektrumból egy-egy tartományra számoltuk: 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.

Kísérleti úton igazoltuk, hogy a legjobb osztályozási eredményt a 2,5–3,5kHz részsávra számolt MFC együtthatókkal értük el mind a GMM, mind a GMM-UBM esetében. Ez azonban statisztikailag csak részben igazolható. Az $MFC_{(2,5–3,5)}$ jellemzővel elért eredmények szignifikánsan különböznek az $MFC_{(1,5–2,5)}$-vel (Z=-2,201; p=0,028) és az $MFC_{(3,5–4,5)}$-vel (Z=-2,201; p=0,028) elért eredményektől, de a teljes spektrumot lekódoló eljárástól nem. Az adatokból azonban látszik, hogy szisztematikusan jobban teljesít az $MFC_{(2,5–3,5)}$ jellemző, mint az $MFCC_{(full-ban)}$. 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.

III. tézis: *Igazoltuk, hogy az egyszerre beszélések detektálásában jól alkalmazható a mély neurális hálózat (DBN) mint az akusztikai jellemzők reprezentációja, illetve előnyös összekapcsolni hibrid rendszerré szupport vektor géppel (SVM).*

Kísérletileg igazoltuk, hogy a más (főleg vizuális) információk feldolgozására használt mély neurális hálózatok alkalmasak az egyszerre beszélések akusztikai előfeldolgozására, jellemzőkinyerésre. A legjobb eredményt a Mel-skála szerinti log szűrőbank jellemzővel és H1(300)-H2(600)-H3(500) topológiájú DBN-nel és SVM osztályozóval értük el, amelynek EER értéke 44,33% volt. Jóllehet az egyszerre beszélések automatikus osztályozása igen fontos feladat a beszélődetektálásban, mégis csak néhány tanulmány foglalkozik ezzel a kérdéssel (pl. Mowlaee et al. 2010; Saeidi et al. 2010). Boakye és munkatársai (2008) az AMI korpuszon (amely 18%-ban tartalmaz átfedő beszédet) 38%-os F-score-t értek el az átfedő beszéd detektálására. Yella és Bourlard (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%-al tudták csökkenteni. Prozódiai jellemzőket is tartalmazó eljárással Zelenak és Hernando (2011) hasonló F-score-t tudtak elérni az átfedőbeszéd-detektálásra, ez közel 40% 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. Yella és Bourlard (2013) Shriberg 2001-es kutatási eredményeiből indulnak 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. Yella és Bourlard egy olyan algoritmust használt, amely ezt a jelenséget modellezi. Az általuk javasolt egyszerrebeszélés-detektálót beépítették beszélődetektálóba, amellyel 5%-os relatív DER javulást tudtak elérni.

A fent leírt eredményekből látszik, hogy az általunk javasolt mélyrétegű jellemzőkinyerő algoritmussal közel azonos eredményt tudtunk elérni, mint más eljárásokkal.

IV. tézis: *Igazoltuk, hogy a II. tézisben kikísérletezett akusztikai jellemzők javuló eredménnyel implementálhatók a beszélődetektálásba.*

Ha az MFCC jellemzőt specifikusan a 2,5–3,5kHz-es részsávra számoljuk, akkor az eredmények javíthatók, hiszen az eredményeink szerint ezen frekvenciatartomány tartalmazhatja a beszélőre specifikusan jellemző akusztikai lenyomatokat. Ezért a standard BIC beszélődetektálóba ezt az akusztikai jellemzőt használtuk mint a standard beszélődetektáló módosítását. Az eredmények 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. Jóllehet az átlagos javulás mértéke csupán 0,8%, ez a különbség szignifikáns (Wilcoxon teszt Monte Carlo szimulációval kiegészítve: Z=-2,824; p=0,005).

V. tézis: *Kísérletileg igazoltuk, hogy a beszélődetektáló eredményei 2,49%-kal javíthatók, ha az egyszerrebeszélés-detektálót implementáljuk a rendszerbe.*

Kísérletileg igazoltuk, hogy az *Egyszerre beszélések automatikus osztályozása spontán magyar társalgásokban* című fejezetben létrehozott algoritmus implementálásával a beszélődetektáló rendszerünkbe átlagosan 2,49%-os relatív javulást tudtunk elérni, vagyis a beszélődetektálás DER értékét 31,21%-ról le tudtuk csökkenteni 28,71%-ra.

10. IRODALOM

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

Ajmera, J. – Bourlard, H. – Lapidot, I. 2002. *Improved unknown-multiple speaker clustering using HMM*. Technical report, IDIAP.

Ajmera, J. – McCowan, I. – Bourlard, H. 2003. *Robust speaker change detection*. Technical report, IDIAP.

Ajmera, J. – McCowan, I. – Bourlard, H. 2004. Robust speaker change detection. In: *IEEE Signal Processing Letters* 11/8: 649–651.

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

Ajmera, J. 2004. *Robust audio segmentation*. PhD thesis, Ecole Polytechnique Federale de Lausanne.

András Beke, György Szaszák: Unsupervised Clustering of Prosodic Patterns in Spontaneous Speech. TSD 2012: 648-655

Anguera X, 2006. Robust speaker diarization for meetings. PhD thesis. Universitat Politecnica De Catalunya.

Anguera, X. – Aguilo, M. – Wooters, C. – Nadeu, C. – Hernando, J. 2006. Hybrid speech/nonspeech detector applied to speaker diarization of meetings. In: Proceedings of Speaker Odyssey Workshop, Puerto Rico, USA.

Anguera, X. – Hernando, J. 2004. Evolutive speaker segmentation using a repository system. In: Proceedings of International Conference on Speech and Language Processing, Jeju Island, Korea.

Anguera, X. – Wooters, C. – Hernando, J. 2005. Speaker diarization for multi-party meetings using acoustic fusion. In: IEEE Automatic Speech Recognition and Understanding Workshop, Puerto Rico, USA.

Anguera, X. – Wooters, C. – Pardo, J. M. 2006. Robust speaker diarization for meetings: ICSI RT06s evaluation system. In: *Proceedings of INTERSPEECH 2006.*

Anguera, X.: 2005. Xbic: Real-time cross probabilities measure for speaker segmentation. Technical report, ICSI.

Appel, U. – Brandt, A. 1982. Adaptive sequential segmentation of piecewise stationary time series. Information Sciences 29/1: 27–56.

Armani, L.; Matassoni, M.; Omologo, M.; Svaizer, P. (2003). Use of a CSP-based voice activity detector for distant-talking ASR, Proc. EUROSPEECH 2003, Geneva, Switzerland, pp. 501–504.

Atal, B. 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.

Atal, B. S. 1976. Automatic recognition of speakers from their voices. In: Proceedings of the *Institute* of *Electrical* and *Electronic Engineers (IEEE)* 64: 460–475.

Attili, J. – Savic, M. – Campbell, J. 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. 599–602.

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

Austin, J. L. 1962/1990. *Tetten ért szavak*. Akadémiai Kiadó, Budapest.

Bakis, R. – Chen, S. – Gopalakrishnan, P. – Gopinath, R. 1997. Transcription of broadcast news shows with the IBM large vocabulary speech recognition system. In: Proceedings of the Speech Recognition Workshop, 67–72.

Bangerter, A. – Clark, H. H. – Katz, A. R. 2004. Navigating joint projects in telephone conversations. Discourse Processes 37: 1–23.

Barras, C. – Zhu, X. – Meignier, S. – Gauvain, J.-L. 2004. Improving speaker diarization. In: Proceedings of Fall 2004 Rich Transcription Workshop (RT04), Palisades, NY.

Bartha Csilla – Hámori Ágnes 2010. Stílus a szociolingvisztikában, stílus a diskurzusban. Nyelvi variabilitás és társas jelentések konstruálása a szociolingvisztika „harmadik hullámában”. Magyar Nyelvőr 134/3: 298–321.

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

Bata Sarolta 2009a. Beszélőváltások a beszédpartnerek személyes kapcsolatának függvényében. In: 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 beszélı beszédének szupraszegmentális jellegzetességeire. In Keszler Borbála – Tátrai Szilárd (szerk.): Diskurzus a grammatikban, grammatika a diskurzusban. Budapest, Tinta Kiadó, 74–83.

Bavelas, J. B. – Coates, L. – Johnson, T. 2002. Listener responses as a collaborative process: The role of gaze. Journal of Communication 52/4: 566–580.

Bazzanella, C. 1990. Phatic connectives as interactional cues in contemporary spoken Italian. Journal of Pragmatics 14: 629–647.

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

Beattie, G. 1977. The dynamics of interruption and the filled pause. The British journal of Social and Clinical Psychology 16/3: 283–284.

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

Beke András – Szaszák György 2009. A svávariációk automatikus felismerése magyar nyelvű spontán beszédben. Beszédkutatás 2009. 148–169.

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

Beke András 2009. A veláris magánhangzók stabilitása a spontán beszédben. In: Gecső Tamás – Sárdi Csilla (szerk.) A kommunikáció nyelvészeti aspektusai. Kodolányi János Főiskola, Tinta Kiadó, Székesfehérvár, Budapest. 27–31.

Beke András 2012. Automatic identification of discourse markers in spontaneous speech for speaker diarization. SJUSK 2012. Copenhagen.

Belin, P. – Fecteau, S. – Bédard, C. 2004. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences 8/3: 129–135.

Bell, A. – Johnson, G. 1997. Towards a sociolinguistics of style. University of Pennsylvania Working Papers in Linguistics 4.1: A Selection of Papers from NWAVE 25. 1–21.

Bell, A. 1984. Language style as audience design. Language in Society 13: 145–204.

Bell, A. 2001. Back in style: reworking audience design. In: Eckert, P. – Rickford, J. R. 2001. Style and sociolinguistic variation. Cambridge University Press. Cambridge. 139–69.

Bellili A. et al. 2001. An hybrid mlp-svm handwritten digit recognizer. In Proc. of the Intl. Conf. on Document Analysis and Recognition (ICDAR), 2001, pp. 28–32.

Ben, M. – Betser, M. – Bimbot, F. – Gravier, G. 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.

Ben-Harush O. – Guterman H. – Lapidot I. 2009. Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization. In Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on, 2009. 1–6.

Bennett, M. – J. Jarvis 1991. The communicative function of minimal responses in everyday conversation. Journal of Social Psychology 131/4: 519–523.

Bergmann J. R. 1988. Ethnomethodologie und Konversationsanalyse. Kurseinheit 1–3. Fernuniversität – Gesamthochschule – in Hagen. Fachbereich Erziehungs-, Sozial- und Geisteswissenschaften.

Bishop, C. M. 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Bishop, C. M. 1995. Neural networks for pattern recognition. Oxford University Press, Oxford.

Bishop, C. M. 1996. Theoretical foundations of neural networks. In: Borcherds, P. – Bubak, M. – Maksymowicz, A. (eds.) Proceedings of Physics Computing, Academic Computer Centre, Krakow. 500–507.

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

Blakemore, D. 1988. 'So' as a constraint on relevance. In: Kempson, R. M. (ed.) Mental representations. The interface between language and reality. Cambridge University Press, Cambridge. 183–195.

Boakye K. 2008. Audio Segmentation for Meetings Speech Processing. Ph.D. dissertation, University of California at Berkeley, 2008.

Boakye K. – Vinyals O. – Friedland G. 2008. Two'sa crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech. In Proc. Interspeech 2008. 32–35.

Boakye K. –Trueba-Hornero B. –Vinyals O. –Friedland G. 2008. Overlapped speech detection for improved speaker diarization in multiparty meetings. Proc. ICASSP. 4353–4356, 2008.

Boakye, K. – Trueba-Hornero, B. – Vinyals, O. – Friedland, G. 2008. Overlapped speech detection for improved speaker diarization in multiparty meetings. In: Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, Nevada. 4353–4356.

Boakye, K. – Vinyals, O. – Friedland, G. 2011. Improved Overlapped Speech Handling for Speaker Diarization. In: Proceeding of INTERSPEECH 2011. Firenze, Olaszország. 941–944.

Boersma, P. – Weenink, D. 2009. Praat: doing phonetics by computer (Version 5.3.) http://www.praat.org/

Bóna Judit 2006. A megakadásjelenségek akusztikai és percepciós sajátosságai. Beszédkutatás 2006. 101–113.

Bonastre, J.-F. – Delacourt, P. – Fredouille, C. – Merlin, T. – Wellekens, C. 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.

Bondarko, L. V. – Zagorujko, N. G. – Kozhevnikov, V. A. – Molchanov, A. P. – Chistovich, L. A. 1970. A model of speech perception by humans. Working papers in linguistics

6. Technical report. Computer and Infomation Science Research Center, Ohio State University, 70–92.

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

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.

Browman, C. P. – Goldstein, L. M. 1992. Articulatory phonology: an overview. Phonetica 49: 155–180.

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

Bucholtz, M. 1995. From Mulatta to Mestiza. Language and the Reshaping of Ethnic Identity. In: Hall, K. – Bucholtz, M. (eds.) Gender Articulated: Language and the socially constructed self. Routledge. London–New York. 351–374.

Bucholtz, M. 2004. Styles and Stereotypes: The Linguistic Negotiation of Identity among Laotian American Youth. Pragmatics 14/2–3: 127–47.

Bunt, H. – Petukhova, V. 2009. Towards a Multidimensional Semantics of Discourse Markers in Spoken Dialogue. In: *Proceedings of the Eighth International Workshop on Computational Semantics.* 157–168.

Bunt, H. 2000. Dynamic interpretation and dialogue theory. In: Taylor, M. – Bouwhuis, D. – Neels, F. (eds.) *The structure of multi-modal dialogue, Vol. 2.* John Benjamins, Amsterdam.

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

Campbell, W. M. – Campbell, J. P. – Reynolds, D. – Singer, E.– Torres-Carrasquillo, P. A. 2006. Support Vector Machines for Speaker and Language Recognition. *Computer Speech and Language* 20/2-3: 10–29.

Carlson, L. 1984. *'Well' in dialogue games. A discourse analysis of the interjection 'well' in idealized conversation*. John Benjamins, Amsterdam.

Çetin O. – Shriberg E. 2006. Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap. In: Proc. ICASSP, 2006, pp. 357–360, Toulouse, France.

Çetin, Ö. – Shriberg, E. 2006. Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition. In: *Proceedings of INTERSPEECH 2006*. 293–296.

Cettolo, M. – Vescovi, M. 2003. Efficient audio segmentation algorithms based on the BIC. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*.

Chafe, W. 1994. *Discourse, Consciousness and Time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press, Chicago.

Che, C. – Lin, Q. 1995. Speaker recognition using HMM with experiments on the YOHO database. In: *Proceedings of EUROSPEECH*, Madrid, Italy. 625–628.

Chen, S. S. – Gales, M. J. F. – Gopinath, R. A. – Kanvesky, D. – Olsen, P. 2002. Automatic transcription of broadcast news. *Speech Communication* 37: 69–87.

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

Cheng, S.-S. – Wang, H.-M. 2003. A sequential metric-based audio segmentation method via the bayesian information criterion. In: *Proceedings of Eurospeech*, Geneva, Switzerland.

Cheng, S.-S. – Wang, H.-M. 2004. METRIC-SEQDAC: A hybrid approach for audio segmentation. In: *Proceedings of International Conference on Speech and Language Processing*, Jeju Island, Korea.

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

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, Letöltés ideje: 2013.06.05. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cho, Y.D.; Kondoz, A. (2001). Analysis and improvement of a statistical model-based voice activity detector, IEEE Signal Processing Letters, vol. 8, no. 10, pp. 276–278.

Chomsky, N. – Halle, M. 1968. *The sound pattern of English*. Harper & Row, New York.

Clancy, P. M. – Thompson, S. A. – Suzuki, R. – Tao, H. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics* 26/3: 355–387.

Clark, A. 1996. *A megismerés építőkövei*. Osiris Kiadó, Budapest.

Clark, H. H. – Clark, E. C. 1987. Hogyan tervezzük meg, hogy mit mondjunk? In: Pléh Csaba (szerk.): *Szöveggyűjtemény a pszicholingvisztika tanulmányozásához*. Tankönyvkiadó. Budapest. 333–374.

Clark, H. H. – Clark, E. V. 1977. *Psychology and language. An Introduction to Psycholinguistics*. Harcourt Brace Jovanovich, New York.

Clark, H. H. 1994a. Managing problems in speaking. *Speech Communication* 15: 243–250.

Clark, H. H. 1994b. Discourse in Production. In: Gernsbacher, M. A. (ed.) *Handbook of Psycholingiustics*. Academic Press, San Diego. 985–1021.

Cohen, R. 1984. A computational theory of the function of clue words in argument understanding. In*: Proceedings of Coling,* Standford, 251–258.

Cohen, R. 1984. A computational theory of the function of clue words in argument understanding. In: *Proceedings of the 10th International Conference on Computational Linguistics*. 251–255.

Colombi, J. M. – Ruck, D. W. – Rogers, S. K. – Oxley, M. – Anderson, T. 1996. Cohort selection and word grammar effects for speaker recognition. In: *Proceedings of IEEE International Conference on Acoustics*, Speech and Signal Processing, Atlanta, GA, 85–88.

Coupland, N. 1980. Style-shifting in a Cardiff work-setting. *Language in Society* 9: 1–12.

Coupland, N. 1985. Hark, hark the Lark: Social Motivations for Phonological Style-shifting. *Language and Communication* 5/3: 153–171.

Coupland, N. 2001. Dialect stylization in radio talk. *Language in Society* 30/3: 345–375.

Coupland, N. 2007a. *Style*. Cambridge University Press. Cambridge.

Coupland, N. 2007b. Aneurin Bevan, class wars and the styling of political antagonism. In: Auer, P. (ed.) *Style and Social Identities: alternative approaches to linguistic heterogenity*. Mouton, Berlin. 213–247.

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

Cutler, A., – Norris, D.G. 1979. Monitoring sentence comprehension. In: Cooper, W. E. – Walker, E. C. T. (eds.) *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Erlbaum, Hillsdale, NJ. 113–134.

Dahl, G. E., Ranzato, M., Mohamed, A., Hinton, G.: Phone recognition with the mean-covariance restricted boltzmann machine. In: NIPS (2010) 469–477

Daniel P. W. Ellis 2005. PLP and RASTA and MFCC, and inversion in Matlab, http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

Daniel S. – Madan Raja – Shanmugam A. 2011. Hybrid Feature Based War Scene Classification using ANN and SVM: A Comparative Study. In: International Journal of Engineering Science & Technology;2011, Vol. 3 Issue 5, 869-873.

Denes, P.B. – Pinson, E.N. 1993. The Speech Chain: Physics and Biology of Spoken Language. 2nd edition. New York: Freeman

De Rulter, J. P. − Mitterer, H. − Enfield, N. J. 2006. Projecting the end of a speakers turn: A cognitive cornerstone of conversation. *Language 82/3:* 515−535.

DeFrancisco, V. L. 1991. The sounds of silence: How men silence women in marital relations. *Discourse and Society* 2/4: 413–423.

Delacourt, P. – Kryze, D. – Wellekens, C. J. 1999a. Detection of speaker changes in an audio document. In: *Proceedings of Eurospeech 1999*. 1195–1198.

Delacourt, P. – Kryze, D. – Wellekens, C. J. 1999b. Speaker-based segmentation for audio data indexing. In: *Proceedings of the ESCA Workshop Accessing Information in Spoken Audio*.

Delacourt, P. – Wellekens, C. J. 1999. Audio data indexing: Use of second-order statistics for speaker-based segmentation. In: IEEE International Conference on Multimedia, Computing and Systems, Florence, Italy. 959– 963.

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

Dell, G. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93: 283–321.

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

Denes P.B. – Pinson E.N. 1993. The speech chain – the physics and biology of spoken language. W.H.Freeman and Co., New York, 1993.

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

Dittmann, A. T. – Llewellyn, L. G. 1967. The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology* 6: 341–348.

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

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

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

Drummond, K. – Hopper, R. 1993b. Acknowledgment tokens in series. *Communication Reports* 6/1: 47–53.

Duncan, S. – Fiske, D. 1977. *Face-to-face interaction: Research, methods, and theory*. Lawrence Erlbaum , Hillsdale, New Jersey.

Duncan, S. – Fiske, D. 1985. *Interaction structure and strategy*. Cambridge University Press, Cambridge.

Duncan, S. – Fiske, D. W. 1985. The turn system. In: S. Duncan, Jr. and D.W. Fiske (eds.) *Interaction structure and strategy.* Cambridge University Press, Cambridge. 43–64.

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

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

Duncan, S. 1973. Toward a grammar for dyadic conversation. *Semiotica* 9: 29–46.

Eckert, P. 2000. *Linguistic variation as social practice*. Blackwell. Oxford.

Edelsky, C. 1981. Who's got the floor? *Language in Society* 10: 383–421.

Elizabeth Shriberg – Andreas Stolcke – Don Baron 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In Eurospeech, Aalborg, Denmark, 2001. 1359–1362

Eskenazi, M. 1993. Trends in speaking styles research. In: *Proceedings of Eurospeech 1993*. 1. 501–509.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27/8: 861–874.

Fék Márk 1997. Beszélőfelismerés neurális hálózatokkal és vektorkvantálóval. *OTDK konferencia*. Szeged 1997.

Ferrer, L. – Shriberg, E. – Kajarekar, S. – Sönmez, K. 2007. Parameterization of prosodic feature distributions for SVM modeling in speaker recognition. In: Proceedings of IEEE *International Conference on Acoustics, Speech, Signal Processing*. 233–236.

Fischer, K. 2000. Discourse particles, turn-taking, and the semantics-pragmatics interface. *Revue Semantique et Pragmatique* 8: 111–137.

Fishman, P. 1978. Interaction: The work women do. *Social Problems* 24*:* 397–406.

Fodor, J. A. 1981. *Representations: Philosophical Essays on the Foundations of Cognitive Science.* The MIT Press, Cambridge, MA. 257.

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

Forster, K. 1990. Lexical Processing. In Osherson, D. N. – Lasnik, H. (eds.) *Language: An Invitation to Cognitive Science*, Bradford Books, MIT Press, Cambridge MA. 95–131.

Fowler, C. 1986. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14: 3–28.

Fowler, C. A. – Saltzman, E. 1993. Coordination and coarticulation in speech production. *Language and speech* 36/2-3: 171–195.

Fox Tree, J. E. – Schrock, J. C. 1999. Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language* 40: 280–295.

Fox Tree, J. E. –Schrock, J. C. 2002. Basic meanings of you know and I mean. *Journal of Pragmatics* 34: 727–747.

Franke, W. 1990. *Elementare Dialogstrukturen: Darstellung, Analyse, Diskussion*. Niemeyer, Tübingen.

Fraser, B. 1999. What are discourse markers? *Journal of Pragmatics* 31: 931–952.

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

Fries, C.G. 1952. *The Structure of English*. Harcourt Brace, New York.

Fromkin, V. A. 1999. Gondolatok az agy, az elme és a nyelv közti kapcsolatokról. In: Bánréti Zoltán (szerk.) *Nyelvi struktúrák és az agy – Neurolingvisztikai tanulmányok*. Corvina Kiadó, Budapest. 59–91.

Fry, D. B. 1973. The linguistic evidence of speech errors. In: Fromkin, V. A. (ed.) *Speech errors as linguistic evidence*. Mouton, The Hague. 157–163.

Furui, S. 1981. Cepstral analysis technique for automatic speaker verification. In: *Proceedings of the Institute of Electrical and Electronic Engineers, International Conference on Acoustics, Speech and Signal Processing* 29/2: 254–272.

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

Gallois, C. – Ogay, T. – Giles, H. 2005. Communication Accommodation Theory: A look back and a look ahead. In: Gudykunst, W. B. *Theorizing about intercultural communication*. Sage, Thousand Oaks, CA. 121–148.

Gangadharaiah, R. – Narayanaswamy, B. – Balakrishnan, N. 2004. A novel method for twospeaker segmentation. In: *Proceedings of the International Conference on Speech and Language Processing*, Jeju Island, Korea.

Garfinkel, H. 1967. *Studies in Ethnomethodology*. Prentice Hall, Englewood Cliffs, NJ.

Garman, M. 1990. *Psycholinguistics*. Cambridge University Press, Cambridge.

Garrett, M. 1988. Processes in language production. In: Newmeyer, F. (ed.) *Linguistics: The Cambridge Survey III. Language: Psychological and Biological Aspects.* Cambridge University Press. Cambridge. 69–96.

Garrett, M. F. 1982. Production od speech: Observations from normal and pathological language use. In: Ellis, A. W. (ed.) *Normally and pathology in cognitive functions*. Academic Press, London, 19–76.

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

Gernsbacher, M. A. – Faust, M. 1991. The role of suppression in sentence comprehension. In: Simpson, G. B. (ed.) *Understanding word and sentence*. North-Holland, Amsterdam. 97–128.

Gernsbacher, M. A. (ed.) 1994. *Handbook of psycholinguistics*. Academic Press, San Diego, CA.

Giles, H. – Coupland, J. – Coupland, N. 1991. Accommodation Theory: Communication, context, and consequence. In: Giles, H. – Coupland, J. – Coupland, N. (eds.) *The contexts of accommodation*. Cambridge University Press, New York. 1–69.

Giles, H. – Coupland, N. 1991. *Language: Contexts and consequences*. Open University Press, Milton Keynes.

Giles, H. 1973. Accent mobility: A model and some data. *Anthropological Linguistics* 15: 87–105.

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

Gish, H. – Siu, M.-H. – Rohlicek, R. 1991. Segregation of speakers for speech recognition and speaker identification. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing 2*, Toronto, Canada, 873–876.

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

Goffman, E. 1983. The Interaction Order. *American Sociological Review* 48: 1–17.

Goldman-Eisler, F. 1968. *Psycholinguistics. Experiments in Spontaneous Speech*. Academic Press, London–New York.

Goodwin, Ch. 1979. The interactive construction of a sentence in natural conversation. In: Psathas, G. (ed.) *Everyday Language: Studies in Ethnomethodology*. Irvington Publishers, New York, 97–121.

Górriz, J.M.; Ramírez, J.; Segura, J.C.; Puntonet, C.G. (2006b). An effective cluster-based model for robust speech detection and speech recognition in noisy environments, Journal of the Acoustical Society of America, vol. 120, No. 1, pp. 470-481.

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

Gósy Mária 1998. A beszédtervezés és beszédkivitelezés paradoxona. *Magyar Nyelvőr* 122/1: 3–15.

Gósy Mária 2000. *A hallástól a tanulásig*. Nikol, Budapest.

Gósy Mária 2001. A testalkat és az életkor becslése a beszéd alapján. MAGYAR NYELVŐR 125:(4). 478-487.

Gósy Mária 2003. Virtuális mondatok a spontán beszédben. *Beszédkutatás 2003*. 19–44.

Gósy Mária 2004. *Fonetika, a beszéd tudománya*. Osiris Kiadó. Budapest.

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

Gósy Mária 2006. A semleges magánhangzó nyelvi funkciói. *Beszédkutatás 2006.* 8–23.

Gósy Mária 2008. A zaj hatása a beszédre. *Beszédkutatás 2008*. 5–21.

Gósy Mária 2008. Magyar spontánbeszéd-adatbázis − BEA. *Beszédkutatás 2008*. 194−207.

Gósy Mária–Horváth Viktória 2009. Hogyan tükrözi a kiejtés a nyelvi funkció változását? In Keszler Borbála, Balázs Géza (szerk.): Diskurzus a grammatikában, grammatika a diskurzusban. Tinta Kiadó, Budapest. 37–45.

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.

Gráczi Tekla Etelka – Bata Sarolta 2010. Megszólalási formák és funkciók az összeszokottság függvényében. In: Gecső Tamás – Sárdi Csilla (szerk.) *Új módszerek az alkalmazott nyelvészeti kutatásban.* Kodolányi János Főiskola, Tinta Könyvkiadó, Székesfehérvár, Budapest. 28–32.

Gráczi Tekla Etelka – Horváth Viktória 2010. A magánhangzók realizációja spontán beszédben. *Beszédkutatás 2010.* 5–16.

Gregory, S. W. – Green, B. E. – Carrothers, R. M. – Dagan, K. A. 2001. Verifying the primacy of voice fundamental frequency in social status accommodation. *Language and Communication* 21: 37–60.

Grice, H. P. 1975. Logic and conversation. In: Cole, P. – Morgan, J. L. (eds.) *Syntax and Semantics 3. Speech Acts*. Academic Press, New York. 41–58. [Magyarul: Grice, H. P. 1997. A társalgás logikája. In: Pléh Csaba – Síklaki István – Terestyéni Tamás (szerk.) *Nyelv – kommunikáció – cselekvés*. Osiris kiadó, Budapest. 213–228.]

Grosz, B. J. – Sidner, C. L. 1986. Attention, intentions, and the structure of discourse. Computational Linguistics 12/3: 175–204.

Grósz, T., Tóth, L.: A Comparison of Deep Neural Network Training Methods for Large Vocabulary Speech Recognition, Proc. TSD2013, pp. 36–43, 2013.

Gyarmathy Dorottya 2007. A beszédpercepciós és beszédprodukciós folyamat összefüggései a megakadásjelenségek tükrében. In: Heltai Pál (szerk.) *Nyelvi modernizáció. A XVI. Magyar Alkalmazott Nyelvészeti Kongresszus előadásai*. Szent István Egyetemi Kiadó, Gödöllő, 449–454.

Gyarmathy Dorottya 2008. Különböző zajok hatása a beszédprodukcióra. *Alkalmazott Nyelvtudomány* VIII/1-2: 135–147.

Gyarmathy Dorottya 2009. A beszélő bizonytalanságának jelzései: ismétlések és újraindítások. *Beszédkutatás 2009*. 196–216.

Gyarmathy Dorottya 2011. *A magakadások javításának stratégiái a spontán beszédben*. PhD-disszertáció. ELTE, Budapest.

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

Hardcastle, W. J. – Marchal, A. (eds.) *Speech production and speech modelling*. Kluwer, Dordrecht.

Harley, T. 2001. *The Psychology of Language. From Data to Theory*. Taylor & Francis, New York.

Harnad, S. 1992/1993. A szimbólum-lehorgonyzás problémája. *Magyar Pszichológiai Szemle* 32/3: 365–383. Újraközlése: In: Pléh Csaba (szerk.) *Kognitív tudomány*. Osiris Kiadó, Budapest. 207–222.

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

Hayashi, T. – Hayashi, R. 1991. Back channel or main channel: A cognitive approach based on floor and speech acts. *Pragmatics and Language Learning Monograph Series* 2: 119–138.

Heck, L. – Sankar, A. 1997. Acoustic clustering and adaptation for robust speech recognition. In: *Proceedings of Eurospeech 1997*, Rhodes, Greece.

Heeman, P. A. – Allen, J. F. 1995. *The Trains spoken dialog corpus*. CD-ROM, Linguistics Data Consortium.

Heeman, P. A. – Allen, J. F. 1999. Speech repairs, intonational phrases and discourse markers: Modelling speakers utterances in spoken dialogue. *Computational Linguistics* 12/3: 1–45.

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

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

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

Higgins, A. L. – Bahler, L. – Porter, J. 1991. Speaker verification using randomized phrase prompting. *Digital Signal Processing* 1/2: 89–106.

Higgins, A. L. – Wohlford, R. 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.

Hirschberg, J. – Litman, D. 1993. Empirical studies on the disambiguqtion of cue phrases. *Computational Linguistics* 25/4: 501–530.

Honbolygó, Ferenc 2009. *A beszéd prozódiai jellemzőinek észlelése*. PhD-disszertáció. Témavezető: Csépe Valéria. ELTE PPT, Budapest.

Horváth Viktória 2004. Megakadásjelenségek a párbeszédekben. *Beszédkutatás 2004*. 187–199.

Horváth Viktória 2007a. Vannak-e „női" és „férfi" megakadásjelenségek a spontán beszédben? *Magyar Nyelvőr* 131/3: 315–323.

Horváth Viktória 2007b. A dysarthria tünetei a spontán beszédben. In: Heltai Pál (szerk.) *Nyelvi modernizáció. A XVI. Magyar Alkalmazott Nyelvészeti Kongresszus előadásai*. Szent István Egyetemi Kiadó. Gödöllő. 455–461.

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

Howell, P. 2007. A model of serial order problems in fluent, stuttered and agrammatical speech. *Human Movement Science* 26: 728–741.

Hölker, K. 1991. Französisch: Partikelforschung. In: Holtus, G. – Metzeltin, M. – Schmitt, Ch. (eds.) *Lexikon der Romanistischen Linguistik*. Niemeyer, Tübingen. 77–88.

Hsu, Chih-Wei – Chang, Chih-Chung – Lin, Chih-Jen 2003. A practical guide to support vector classification. *Technical report, Department of Computer Science*, National Taiwan University

Huixuan Tang 2008. A Comparative Evaluation of Deep Belief Nets in Semi-supervised Learning, Report for CSC2515, 2008.

Hung, J. – Wang, H. – Lee, L. 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.

INTERSPEECH 2011. Olaszország. Firenze. http://www.interspeech2011.org/

ITU 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, 1996.

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

J. Saunders 1996. Real-time discrimination of broadcast speech/-music. In: Proceedings of the Acoustics, Speech, and Signa Processing (ICASSP96), 993–996.

J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002 (ISBN 981-238-151-1).

James, A. R. 1983. 'Well' in reporting clauses: Meaning and form of a 'lexical filler'. *Arbeiten aus Anglistik und Amerikanistik* 8/l: 33–40.

Janin, A. – Baron, D. – Edwards, J. – Ellis, D. – Gelbart, D. – Morgan, N. – Peskin, B. – Pfau, T. – Shriberg, E. – Stolcke, A. – Wooters, C. 2003. The ICSI meeting corpus. In: *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing*. 364–367.

Jefferson, G. 1983. *Two explorations of the organization of overlapping talk in conversation: Notes on some orderlinesses of overlap onset*. Tilburg Papers in Language and Literature, No. 28. Tilburg, The Netherlands.

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

Jefferson, G. 1993. Caveat speaker: Preliminary notes on recipient topic-shift implicature. *Research on Language and Social Interaction* 26: 1–30. (Original work published 1983).

Jin, H. – Kubala, F. – Schwartz, R. 1997. Automatic speaker clustering. In: *DARPA Speech Recognition workshop*, Chantilly, USA.

Jin, Q. – Laskowski, K. – Schultz, T. – Waibel, A. 2004. Speaker segmentation and clustering in meetings. In: *Proceedings of NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada. 112–117.

Johnson, K. 1997. Speech perception without speaker normalization: an exemplar model. In: Johnson, K. – Mullennix, J. W. (eds.) *Talker variability in speech processing.* Academic Press, San Diego, 145–166.

Johnson, S. – Woodland, P. 1998. Speaker clustering using direct maximization of the MLLR adapted likelihood. In: *Proceedings of International Conference on Speech and Language Processing* 5. 1775–1779.

Johnson, S. 1999. Who spoke when? Automatic segmentation and clustering for determining speaker turns. In: *Proceeding of Eurospeech 1999*, Budapest, Hungary.

Jokinen, K. – Harada, K. – Nishida, M. – Yamamoto, S. 2010. Turn-alignment using eye-gaze and speech in conversational interaction. In: *Proceedings of INTERSPEECH 2010*: 2018–2021.

Joshi, S. – Prahallad, K. – Yegnanarayana, *B.* 2008. AANN-HMM models for speaker verification and speech recognition. In: *Proceedings of International Joint Conference on Neural Networks 2008 (IJCNN)*. 2681–2688.

Juang, B. – Rabiner, L. 1985. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal* 64/2: 391–408.

Jurafsky, D. – Bates, R. – Coccaro, N. – Martin, R. – Meteer, M. – Ries, K. – Shriberg, E. – Stolcke, A. – Taylor, P. – Van Ess-Dykema, C. 1997. Automatic detection of discourse structure for speech recognition and understanding. In: *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara.

Kajarekar, S. – Ferrer, L. – Sönmez, K. – Zheng, J. – Shriberg, E. – Stolcke, A. 2004. Modeling NERFs for speaker recognition. In: *Proceedings of Speaker Odyssey Workshop*. 51–56.

Knapp, M.L. – Hall, J.A. 2001. Nonverbal communication in human action. Belmont, CA: Wadsworth.

Kass, R. E. – Raftery, A. E. 1995. Bayes factors. *Journal of the American Statistics Association* 90: 773–795.

Kátainé Koós Ildikó 1998. Kommunikációs keret az első évben: intonáció – gőgicsélés. *Beszédkutatás '98*. 58–67.

Kawahara, T. – Hasegawa, M. – Shitaoka, K. – Kitade, T. – Nanjo, H. 2004. Automatic indexing of lecture presentations using unsupervised learning of presumed discourse

markers. In: *Proceedings of IEEE Transactions on Speech and Audio Processing* 12/4: 409–419.

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

Kendon, A. 1994. Do gestures communicate? A review. *Research on Language and Social Interaction* 27/3: 175–200.

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

Kharroubi, J. – Petrovska-Delacretaz, D. – Chollet, G. 2001. Text-independent speaker verification using support vector machines. In: *Proceedings of Speaker Odyssey Workshop*, 51–54.

Kingsbury B. E. D. – Morgan N. – Greenberg S. 1998. Robust speech recognition using the modulation spectrogram," Speech Commun., vol. 25, no. 1–3.117–132.

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

Klatt, D. H. 1989. Review of selected models of speech perception. In: Marslen-Wilson, W. (ed.) *Lexical representation and process*. MIT Press, London. 169–226.

Knott, A. – Sanders, T. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics* 30: 135–175.

Koh Chin Wei, Eugene 2008. Speaker Diarization of News Broadcasts and Meeting Recordings. Thesis. Nanyang Technological University, Singapore.

Kotti, M. – Benetos, E. – Kotropoulos, C. 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.

Kotti, M. – Moschou, V. – Kotropoulos, C. 2008. Speaker Segmentation and Clustering. *Signal Processing* 88/5: 1091–1124.

Krauss, R. M. – Weinheimer, S. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal interaction. *Journal of Personality and Social Psychology* 4: 342–346.

Kraut, R. E. – Lewis, S. H. – Swezey, L. W. 1982. Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology* 43/4: 718–731.

Kubala, F. – Jin, H. – Matsoukas, S. – Gnuyen, L. – Schwartz, R. – Machoul, J. 1997. The 1996 BBN byblos HUB-4 transcription system. In: *Proceedings of Speech Recognition Workshop*, 90–93.

Lakoff, R. 1973. Questionable answers and answerable questions. In: B.B. Kachru, B. B. – Lees, R. B. – Malkiel, Y. – Pietrangeli, A. – Saporta, S. (eds.) *Issues in linguistics. Papers in honor of Henry and Rente Kahane*. University of Illinois Press, Urbana, IL. 453–467.

Langleben, M. 1983. On the structure of dialogue. In: Petőfi, J. S. – Sözer, E. (eds.) *Micro and macro connexity of texts*. Buske, Hamburg. 220–286.

Laskowski, K. – Schultz, T. 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, N. J. – Hughes, K. R. – Bowyer, M. D. – Waters, L. T. – Bourne, V. T. 1976. Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America*, 59/3: 675–678.

Leet-Pellegrini, H. M. 1980. Conversational dominance as a function of gender and expertise. In: H. Giles – Robinson, W. P. – Smith, P. M. (eds.) *Language: Social psychological perspectives*. Pergamon Press, Oxford, 97–104.

Lerch Ágnes 2011. *A szintaxis, az intonáció és a pragmatika szerepe a beszélőváltás megvalósulásában magyar nyelvű konverzációkban*. Előadás. Grammatika és kontextus: új szempontok az uráli nyelvek kutatásában III. Budapest, 2011. április 21.

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

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

Li J. – Yu D. – Huang J.-T. – Gong Y. 2012. Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM. In Proc. IEEE Workshop on Spoken Language Technology, 2012.

Li, K.-P. – Wrench, Jr. E. H. 1983. 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.

Liberman, A. M. – Cooper, F. S. – Shankweiler, D. P. – Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychological Review* 74: 431–461.

Liberman, A. M. – Mattingly, I. G. 1985. The motor theory of speech perception revised. *Cognition* 21: 1–36.

Liberman, A. M. – Mattingly, I. G. 1985. The motor theory of speech perception revised. *Cognition* 21/1: 1–36.

Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W. K. – Marchal, A. (eds.) *Speech Production and Speech Modelling*. Kluwer Academic, The Netherlands. 403–439.

Liu, D. – Kubala, F. 1999. Fast speaker change detection for broadcast news transcription and indexing. In: *Proceedings of Eurospeech 1999*. Budapest, Hungary, 1031–1034.

Liu, H. M. – Tsao F. M. – Kuhl P. K. 2003. Speech input to infants: The acoustic-phonetic characteristics of infant-directed speech in Mandarin Chinese. In: Solé, M. J. – Recasens, D. – Romero, J. (eds.) *Proceedings of the 15th International Congress of Phonetic Sciences.* Barcelona. 901–904.

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

Lopez, J. F. – Ellis, D. 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.

Louwerse, M. M. – Mitchell, H. H. 2003. Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse Processes* 35/3: 199–239.

Lu, L. – Li, S. Z. – Zhang, H.-J. 2001. Content-based audio segmentation using support vector machines. In: *Proceedings of ACM International Conference on Multimedia*, 203–211.

Lu, L. – Zhang, H.-J. – Jiang, H. 2002. Content analysis for audio classification and segmentation. In: *Proceedings of IEEE Transactions on Speech and Audio Processing* 10/7: 504–516.

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

Lu, L. – Zhang, H.-J. 2002b. Speaker change detection and tracking in real-time news broadcasting analysis. In: *Proceedings of ACM International Conference on Multimedia*, 602–610.

Maclay, H. – Charles E. O. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15: 19–44.

Mády Katalin 2008. Beszédpercepció és pszicholingvisztika, Pszicholingvisztikai kézikönyv. http://www.phonetik.uni-muenchen.de/~mady/pub/mady_percepcio.pdf

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

Mann, W. – Thompson, S. 1988. *Rhetorical structure theory: toward a functional theory of text organisation*. The MIT Press, Cambridge, MA.

Markel, J. D. – Davis, S. B. 1979. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. In: *Proceedings of the Institute of Electrical and Electronic Engineers, International Conference on Acoustics, Speech and Signal Processing*, 74–82.

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.

Markó Alexandra – Gráczi Tekla Etelka – Bóna Judit 2009. Zöngésségi hasonulás a spontán beszédben és a felolvasásban (esettanulmányok). *Beszédkutatás 2009*. 5–27.

Markó Alexandra 2004. Megakadások vizsgálata különféle monologikus szövegekben. *Beszédkutatás 2004*. 209–222.

Markó Alexandra 2005. *A spontán beszéd néhány szupraszegmentális jellegzetessége. Monologikus és dialogikus szövegek összevetése, valamint a hümmögés vizsgálata*. PhD-értekezés. ELTE, Budapest.

Markó Alexandra 2006. A megakadásjelenségek hatása a beszédészlelésre. *Alkalmazott Nyelvtudomány*. VI/1–2: 103–117.

Markó Alexandra 2006. *Beszélőváltások a társalgásban*. Előadás. IX. Pszicholingvisztikai és Alkalmazott Nyelvészeti Nyári Egyetem. Balatonalmádi, május 21−24.

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

Marslen-Wilson, W. D. – Tyler, L. K. 1975. Processing structure of sentence perception. *Nature* 257: 784–786.

Marslen-Wilson, W. D. – Tyler, L. K. 1980. The temporal structure of spoken language understanding. *Cognition* 8/1: l–71.

Marslen-Wilson, W. D. – Welsh, A. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* 10: 29–63.

Marslen-Wilson, W. D. 1980. Speech understanding as a psychological process. In: Simon, J. C. (ed.) Spoken language generation and understanding. Reidel, New York. 39–67.

Martin, A. F. et al. 1997. The DET Curve in Assessment of Detection Task Performance. Proc. Eurospeech 1997, Rhodes, Greece, September 1997, Vol. 4, 1899–1903.

Massaro, D. W. – Chen, T. H. 2008. The motor theory of speech perception revisited. *Psychonomic bulletin & review* 15/2: 453–457; Discussion: 457–462.

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

Matza, A.– Bistritz, Y. 2011. Skew Gaussian mixture models for speaker recognition. Presentation. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2011*. 28–31.

Maynard, S. K. 1997. Analyzing interactional management in native/non-native English conversation: A case of listener response. *International Review of Applied Linguistics in Language Teaching* 35: 37–60.

McClelland, J. L. – Elman, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18: 1–86.

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

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

Meinedo, H. – Neto, J. 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.

Mermelstein, P. 1976. Distance measures for speech recognition, psychological and instrumental. In: Chen, C.-H. (ed.) *Pattern Recognition and Artificial Intelligence*. Academic Press, New York. 374–388.

Metze, F. – Fugen, C. – Pan, Y. – Schultz, T. – Yu, H. 2004, The ISL RT-04S meetings transcription system. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada.

Meyer, A. S. 1993. Investigation of phonological encoding through speech error analyses: Achievements, limitations, and alternatives. In: Levelt, W. J. M. (ed.) *Lexical access in speech production*. Blackwell. Cambridge, Oxford. 181–211.

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

Miller, J. L. 1990. Speech Perception. In: Osherson, D. N. – Lasnik, H. (eds.) *Language: An Invitation to Cognitive Science*. Bradford Books, MIT Press. Cambridge, MA. 69–93.

Miller, L. C. – R.E. Lechner – D. Rugs 1985. Development of conversational responsiveness: Preschoolers' use of responsive listener cues and relevant comments. *Developmental Psychology* 21: 473–480.

Miró, A. 2006. *Robust speaker diarization for meetings*. PhD thesis. UPC University, Barcelona, Spain.

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

Moh, Y. – Nguyen, P. – Junqua, J.-C. 2003. Towards domain independent speaker clustering. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing,* Hong Kong.

Mohamed, G. Hinton, and G. Penn, 2012. Understanding how deep belief networks perform acoustic modelling. In: Proc. ICASSP, pp. 4273-4276, 2012.

Moraru, D. – Ben, M. – Gravier, G. 2005. Experiments on speaker tracking and segmentation in radio broadcast news. In: *Proceedings of International Conference on Acoustics, Speech and Language Processing*, Lisbon, Portugal.

Mori, K. – Nakagawa, S. 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 P. – Christensen M. G – Tan Z. H. – Jensen S. 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 (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on, 2010. 538–541.

Nakagawa, S. – Suzuki, H. 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.

Nánay Bence 1996. Új divat a tudatfilozófiában: a konnekcionizmus. Andy Clark: A megismerés építőkövei. *Budapesti Könyvszemle* 8: 262–269.

Nánay Bence 1997. A természettudományok és a filozófia találkozása. Pléh Csaba: Kognitív tudomány. *Budapesti Könyvszemle* 9: 148–157.

Nánay Bence 2000. *Elme és evolúció. Az elmefilozófia és a kognitív tudomány tudományos evolúciós megközelítése*. Kávé Kiadó, Budapest.

Nelson Morgan – Hervé Bourlard 1993. Bourlard, H. and Morgan, N. (1993), "Continuous Speech Recognition by Connectionist Statistical Methods," IEEE Trans. on Neural Networks, vol. 4, no. 6, pp. 893-909.

Nemer, E.; Goubran, R.; Mahmoud, S. (2001). Robust voice activity detection using higherorder statistics in the lpc residual domain, IEEE Trans. Speech Audio Processing, vol. 9, no. 3, pp. 217–231.

Németh Géza – Olaszy Gábor (szerk.) 2010. A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek (8–12. fejezet). Akadémiai Kiadó, Budapest.

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

Neuberger Tilda 2012. Virtuális mondatok gyermekek spontán beszédében. *Beszédkutatás 2012*. 217–233.

Neuberger Tilda – Beke András 2013. Automatic Laughter Detection in Spontaneous Speech Using GMM-SVM Method. In: TSD. 2013. 113–120.

Ng, Shu-Kay – Krishnan, T. – McLachlan, G. J. 2011. The EM algorithm. (Second Edition). In: Gentle, J. – Hardle, W. – Mori, Y. (eds.) *Handbook of Computational Statistics: Concepts and Methods* 1. Springer-Verlag, New York, 137–172.

Nguyen, P. 2003. SWAMP: An isometric frontend for speaker clustering. In: *Proceedings of NIST 2003 Rich Transcription Workshop*, Boston, USA.

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

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.

Nishida, M. – Kawahara, T. 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.

Nosofsky, R. M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 39–57.

Nygaard, L.C. – Pisoni, D.B. 1995. Speech Perception: New Directions in Research and Theory. In: Miller, J L. – Eimas, P. D. *Handbook of Perception and Cognition: Speech, Language, and Communication*. Academic Press, San Diego.

Ogden, R. 2004. Non-modal voice qualtity and turm-taking in Finnish. In.: Couper-Kuhlen, E. − Ford, C. (eds.) *Sound patterns in interaction*. Benjamins, Amsterdam, 29−62.

Onshus Ida 2011. Indexing of Audio Databases : Event Log of Broadcast News. PhD thesis. Norwegian University of Science and Technology, Department of Electronics and Telecommunications.

Otterson S. – Ostendorf M. 2007. Efficient use of overlap information in speaker diarization, In Proc. ASRU, Kyoto, Japan, 2007. 686–6.

Owen, M. 1981. Conversational units and the use of 'well'. In: Werth, P. (ed.) *Conversation and discourse*. Croom Helm, London, 99–116.

Pauka Károly 1982. A beszéd megértése. In: Bolla Kálmán (szerk.) *Fejezetek a magyar leíró hangtanból*. Akadémiai Kiadó, Budapest, 175–232.

Perez-Freire, L. – Garcia-Mateo, C. 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, B. – Navratil, J. – Abramson, J. – Jones, D. – Klusacek, D. – Reynolds, D. 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.

Petukhova, V. – Bunt, H. 2009. Towards a multidimensional semantics of discourse markers in spoken dialogue. In: *Proceedings of the Eight International Conference on Computational Semantics* 157–168.

Pickering, M. J. 1999. Sentence comprehension. In: Garrod, S. – Pickering, M. J. (eds.) *Language processing*. Psychology Press, Hove, UK. 123–153.

Pickett, J. M. (szerk.) 1999. The acoustics of speech communication. Allyn and Bacon, Boston, London, Toronto.

Pisoni D. B. – Sawush J. R. 1975. Some stage of processing in speech perception. In: Cohen A. – Nooteboom, S. G. (eds.) *Structure and Process in Speech Perception*. Springer Verlag, Berlin, Heidelberg, New York, 16–35.

Placencia, M. 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.

Pomerantz Anita 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In: Atkinson, J. M. – Heritage, J. (eds.) *Structures of Social Interaction: Studies in Conversation Analysis*. Cambridge University Press, Cambridge. 57–101.

Popescu-Belis, A. – Zufferey, S. 2006. Automatic identification of discourse markers in multiparty dialogues. In: *Working paper 65, ISSCO*, University of Geneva.

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

Qin Jin 2007. Robust Speaker Recognition. PhD thesis. Language Technologies Institute School of Computer Science, Carnegie Mellon University.

Ramírez, J.; Górriz, J.M.; Segura, J.C. (2007). Statistical Voice Activity Detection Based on Integrated Bispectrum Likelihood Ratio Tests, to appear in Journal of the Acoustical Society of America.

Ramírez, J.; Górriz, J.M; Segura, J.C.; Puntonet, C.G; Rubio, A. (2006a). Speech/Non-speech Discrimination based on Contextual Information Integrated Bispectrum LRT, IEEE Signal Processing Letters, vol. 13, No. 8, pp. 497-500.

Ramírez, J.; Segura, J.C.; Benítez, C.; de la Torre, Á.; Rubio, A. (2004a). Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information, Speech Communication, vol. 42, No. 3-4, pp. 271-287.

Ramírez, J.; Segura, J.C.; Benítez, C.; de la Torre, Á.; Rubio, A. (2005). An effective OSF-based VAD with Noise Suppression for Robust Speech Recognition, IEEE Transactions on peech and Audio Processing, vol. 13, No. 6, pp. 1119-1129.

Redeker, G. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14: 367–381.

Reynolds, D. – Andrews, W. – Campbell, J. – Navratil, J. – Peskin, B. – Adami, A. – Jin, Q. – Klusacek, D. – Abramson, J. – Mihaescu, R. – Godfrey, J. – Jones, D. –

Bing, X. 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, D. – Torres-Carrasquillo, P. 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.

Reynolds, D. A. – Quatieri, T. F. – Dunn, R. 2000. Speaker verifcation using adapted Gaussian mixture models. *Digital Signal Processing* 10/1–3: 19–41.

Reynolds, D. A. – Singer, E. – Carlson, B. A. – O'Leary, G. C. – McLaughlin, J. J. – Zixxman, M. A. 1998. Blind clustering of speech utterances based on speaker and language characteristics. In: *Proceedings of International Conference on Speech and Language Processing*, Sidney, Australia.

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

Reynolds, D. A. 1996. M.I.T. Lincoln Laboratory site presentation. In: *Speaker Recognition Workshop* ftp://jaguar.ncsl.nist.gov/speaker/

Reynolds, D. A. 1997. Comparison of background normalization methods for text-independent speaker verification, In: *Proceedings of 5th European Conference on Speech Communication and Technology (Eurospeech).* 963–966.

Reynolds, D. A. 2009. Universal Background Models. In: *Encyclopedia of Biometrics 2009*. 1349–1352.

Reynolds, Douglas A. – Rose, R. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. In: *Proceedings of the Institute of Electrical and Electronic Engineers Transactions on Speech and Audio Processing* 3/1: 72–83.

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

Rosenberg, A. E. – DeLong, J. – Lee, C.-H. – Juang, B.-H. –Soong, F. K. 1992. The use of cohort normalized scores for speaker verification. In: *Proceedings of International Conference on Spoken Language Processing*. 599–602.

Rosenfeld, H. M. 1966. Approval-seeking and approval-inducing functions of verbal and nonverbal responses in the dyad. *Journal of Personality and Social Psychology* 6: 597–605.

Rosenfeld, H. M. 1967. Nonverbal reciprocation of approval: An experimental analysis. *Journal of Experimental Social Psychology* 3: 102–111.

Rougui, J. – Rziza, M. – Aboutajdine, D. – Gelgon, M. – Martinez, J. 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.

Furui S. 1986. Research on individuality features in speech waves and automatic speaker recognition techniques. Speech Communication, vol. 5. 183–197, 1986

Gazor S. – Zhang W. 2003. Speech probability distribution," Signal Processing Letters, IEEE, vol. 10, no. 7. 204–207.

Parthasarathi S. H. K. – Magimai.-Doss M. – Gatica-Perez D. – Bourlard H. 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, 2009.

Sacks, H. – Schegloff, E. A. – Jefferson, G. 1974. A simplest systematics for the organization of turntaking for conversation. *Language* 50: 696–735.

Sacks, H. – Schegloff, E. A. – Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50: 696–735.

Sacks, H. 1992. *Lectures on Conversation*. Blackwell, Oxford.

Saeidi R. – Mowlaee P. – Kinnunen T. – Tan Z. H. – Christensen M. G. – Jensen S. H. – Franti P. 2010. Improving monaural speaker identification by double-talk detection. In Eleventh Annual Conference of the International Speech Communication Association, 2010

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

Salembier, P. – Sikora, T. – Manjunath, B. S. (eds.) 2002. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley & Sons, New York.

Sankar, A. – Beaufays, F. – Digalakis, V. 1995. Training data clustering for improved speech recognition. In: *Proceedings of Eurospeech 1995*, Madrid, Spain.

Sankar, A. – Weng, F. – Stolcke, Z. R. A. – Grande, R. R. 1998. Development of SRI's 1997 broadcast news transcription system. In: *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, USA.

Schegloff, E. 1992. Introduction. In: Sacks, H. *Lectures on Conversation*. Vol.1. Blackwell, Oxford. 9–12.

Schegloff, E. A. – Jefferson, G. – Sacks, H. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53/2: 361–382.

Schegloff, E. A. 1968. Sequencing in conversational openings. *American Anthropologist* 70/6: 1075–1095.

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

Schiffrin, D. 1985. Conversational coherence: The role of 'well'. *Language* 61/3: 640–667.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge University Press, Cambridge.

Schiffrin, D. 1996. Narrative as self portrait: The sociolinguistic construction of identity. *Language in Society* 25/2: 167–203.

Schilling-Estes, N. 2004. Investigating stylistic variation. In: Chambers, J. K. – Trudgill, P.– Schilling-Estes, N. (eds.) *The handbook of language variation and change*. Blackwell. Malden/Oxford. 375–402.

Schirm Anita 2011. A diskurzusjelölők funkciói: a hát, az -e és a vajon elemek története és szinkrón státusa alapján. PhD disszertáció, Szegedi Tudomány Egyetem, Szeged.

Schourup, L. 1999. Discourse markers. *Lingua* 107: 227–265.

Schourup, L. C. 1982. *Common discourse particles in English conversation*. Garland, New York.

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

Schwarz, G. 1971. A sequential student test. *The Annals of Statistics* 42/3: 1003–1009.

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

Selting, M. 1998. TCUs and TRPs: The construction of units in conversational talk. *InLiSt - Interaction and Linguistic Structures 4*. Potsdam.

Selting, M. 2008. Interactional stylistics and style as a contextualization cue. In: Fix, U.– Gardt, A. – Knape, J. (eds.) *Rhetorik und Stilistik/Rhetoric And Stylistics*. Halbband 1. De Gruyter, Berlin/New York. 1039–1053.

Serrano-López, A. J. – Soria-Olivas, E. – Martín-Guerrero, J. D. – Magdalena-Benedito, R. – Gómez-Sanchis, J. 2010. Feature selection using ROC curves on classification problems. In: *Proceedings of IEEE World Congress on Computational Intelligence – International Joint Conference on Neural Networks*, 1980–1985.

Shih-Sian Cheng – Hsin-Min Wang – Hsin-Chia Fu 2008. BIC-based audio segmentation by divide-andconquer," in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, Las Vegas, NV, Apr. 2008. 4841–4844.

Shih-Sian Cheng – Hsin-Min Wang – Hsin-Chia Fu 2010. BIC-Based Speaker Segmentation Using Divide-and-Conquer Strategies With Application to Speaker Diarization," IEEE Transaction on Audio, Speech, and Language Processing, vol. 18, no. 1, pp. 141–157, Jan. 2010.

Shriberg, E. E. (2001). To "Errrr" is Human: Ecology and Acoustics of Speech Disfluencies. Journal of the International Phonetic Association 31(1),Cambridge University Press, 153-169.

Shriberg E. –Stolcke A. – Baron D. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversations," in Proc. Eurospeech 2001, 2001, pp. 1359–1362, aalborg, Denmark.

Shriberg, E. – Ferrer, L. – Kajarekar, S. – Venkataraman, A. – Stolcke, A. Modeling prosodic feature sequences for speaker recognition. 2005. *Speech Communication* 46/3–4: 455–472.

Shriberg, E. – Stolcke, A. – Baron, D. 2001. Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation. In: Proceedings of EUROSPEECH, Aalborg, Denmark. 1359–1362.

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

Sigmund, M. 2008. Gender Distinction Using Short Segments of Speech Signal. In: *International Journal of Computer Science and Network Security* 8/10: 159–162.

Sinha, R. – Tranter, S. E. – Gales, J. J. F. – Woodland, P. 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, M.-H. – Yu, G. – Gish, H. 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, P. – Fortuna, J. – Ariyaeeinia, A. 2001. On the use of the Bayesian information criterion in multiple speaker detection. In: *Proceedings of Eurospeech 2001*, Scandinavia.

Smith, C, 2007. Prosodic accommodation by French speakers to a non-native interlocutor. In: *Proceedings of the 16th International Conference of the Phonetic Sciences*, Saarbruecken, Germany. 1081–1084.

Smolensky, P. 1996. A konnekcionizmus helyes kezeléséről. In: Pléh Csaba (szerk.) *Kognitív tudomány*. Osiris Kiadó, Budapest. 87–135.

Sohn, J. – Kim N. S. – Sung, W. 1999. A statistical model-based voice activity detection. In: *IEEE Signal Processing Letters* 6/1: 1–3.

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

Soong, F. K. – Rosenberg, A. E. – Rabiner, L. R. – Juang, B.H. 1985. A vector quantization approach to speaker recognition. In: *Proceedings of International Conference Acoustics, Speech, and Signal Processing*, Tampa, FL, 387–390.

Sönmez, K. – Shriberg, E. – Heck, L. – Weintraub, M. 1998. Modeling dynamic prosodic variation for speaker verification. In: *Proceedings of International Conference on Spoken Language Process* 3189–3192.

Sporleder, C. – Lascarides, A. 2008. Using Automatically Labelled Examples to Classify Rhetorical Relations: An Assessment. *Natural Language Engineering* 14/3: 369–416.

Sree Harsha Yella, Fabio Valente 2012. Speaker Diarization of Overlapping Speech based on Silence Distribution in Meeting Recordings. In: Proceedings of Interspeech, 2012.

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

Stevens, K. N. 1989. On the quantal theory of speech. *Journal of Phonetics* 17: 3–45.

Stevens, K. N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America* 111: 1872–1891.

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

Streeck, J. 1983. Konversationsanalyse. Ein Reparaturversuch. *Zeitschrift für Sprachwissenschaft* 2/1: 72–104.

Sundberg, U. 1999. Quantity in infant-directed speech. In: *Proceedings of the ICPhS 1999*. 2189–2191.

Svartvik, J. 1980. 'Well' in conversation. In: Greenbaum, S. – Leech, G. – Svartvik, J. (eds.) *Studies in English linguistics for Randolph Quirk*. Longman, London. 167–177.

Szaszák, György and Beke, András: Exploiting Prosody for Syntactic Analysis in Automatic Speech Understanding, Journal of Language Modelling, 0(1) pp. 143-172. (2012)

Szaszák, György; Nagy, Katalin; Beke, András: Analysing the correspondence between automatic prosodic segmentation and syntactic structure. In Proceedings of Interspeech 2011. Florence, Italy, pp. 1057-1060.

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

Tanenhaus M. K. – Carlson G. N. 1989. Lexical structure and language comprehension. In: Marslen-Wilson, W. (ed.). *Lexical Representation and Process*. The MIT Press, Cambridge, MA. 529–561.

Tannen, D. 2005. *Conversational Style: analyzing talk among friends*. Oxford University Press, New York.

Tannen, D. 2007. *Taking voices: repetition, dialogue, and imagery in conversational discourse*. Cambridge University Press. Cambridge–New York.

The RBM code for Matlab developed by KyungHyun Cho is used from http://users.ics.tkk.fi/kcho/

Theodoridis, S. – Koutroumbas, K. 2008. Pattern Recognition, Third Edition. Orlando, FL, USA: Academic Press, Inc.

Theodoros Giannakopoulos 2009. Study and application of acoustic information for the detection of harmful content, and fusion with visual information. Department of Informatics and Telecommunications,University of Athens, Greece, PhD theis.

Tierney, J. 1980. A study of LPC analysis of speech in additive noise. In: Proceedings of IEEE *Transactions on Acoustics, Speech, Signal Processing*, 389–397.

Tishby, N. 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: 563–570.

Tomasello M. 1999b. *The cultural origins of human cognition*. Harvard University Press, Cambridge, MA.

Tomasello, M. – Carpenter, M. – Call, J. – Behne, T. – Moll, H. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* 28: 1–17.

Tomasello, M. 1999a. The cultural ecology of young children's interactions with objects and artifacts. In: Winograd, E. – Fivush, R. – Hirst, W. (eds.) *Ecological approaches to cognition: Essays in honor of Ulric Neisser*. Erlbaum.

Tritschler, A. – Gopinath, R. 1999. Improved speaker segmentation and segments clustering using the bayesian information criterion. In: Proceedings of Eurospeech 1999. 679–682.

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

Thurlow, C. 2006. From Statistical Panic to Moral Panic: The Metadiscursive Construction and Popular Exaggeration of New Media Language in the Print Media, Journal of Computer-Mediated Communciation, 11, 667–701.

Turner, K. (ed.) 1999. *The semantics/pragmatics interface from different points of view, CRISPI Series 1*, Elsevier Science, Oxford, Amsterdam.

Turner, L.H. – West, R. 2010. *Communication Accommodation Theory. Introducing Communication Theory: Analysis and Application* (4th ed.). McGraw-Hill, New York.

van Dijk, T. A. 2006. Introduction: discourse, interaction and cognition. *Discourse Studies* 8: 5–7.

van Dommelen, W. A. – Moxness, B. H. 1995. Acoustic parameters in Speaker Height and Weight Identification: Sex-Specific Behaviour. *Language and Speech* 38/3: 267–287.

Vandecatseye, A. – Martens, J.-P. – Neto, J. – Meinedo, H. – Garcia-Mateo, C. – Dieguez, J. – Mihelic, F. – Zibert, J. – Nouza, J. – David, P. – Pleva, M. – Cizmar, A. – Papageorgiou, H. – Alexandris, C. 2004, *The COST278 pan-European Broadcast News Database*. LREC'04, Lisbon, Portugal.

Vandecatseye, A. – Martens, J.-P. 2003. A fast, accurate and stream-based speaker segmentation and clustering algorithm. In: Proceedings of Eurospeech 2003, Geneva, Switzerland, 941–944.

Váradi Viola 2010. A felolvasás és a spontán beszéd temporális sajátosságainak összehasonlítása. *Beszédkutatás 2008*. 100–109.

Váradi Viola 2012. Beszédhangok törlődése a magyar spontán beszédben. *Beszédkutatás 2012*. 58–69.

Verschueren, J. 1999. *Understanding Pragmatics*. Arnold, London–New York–Sydney–Aukland.

Vescovi, M. – Cettolo, M. – Rizzi, R. 2003. A DP algoritm for speaker change detection. In: *Proceedings of Eurospeech 2003*.

Vinciarelli, A. – Pantic, M. – Bourlard, H. 2009. Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing* 27/12: 1743–1759.

Vipperla R. – Geiger J. – Bozonnet S. – Wang D. – Evans N. – Schuller B. – Rigoll G. 2012. Speech overlap detection and attribution using convolutive non-negative sparse coding. In ICASSP-12, 2012. 4181–4184.

Voulgaris, Z. – Magoulas, G. D. 2008. Dimensionality reduction for feature and pattern selection in classification problems. In: *The Third International MultiConference on Computing in the Global Information Technology*, 160–165.

Vousden, J. I. – Brown, G. D. A. – Harley, T. A. 2000. Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology* 41: 101–175.

Wactlar, H. – Hauptmann, A. – Witbrock, M. 1996. News on-demand experiments in speech recognition. In: *ARPA STL Workshop*.

Wang, R. – Tang, K. 2009. Feature Selection for Maximizing the Area Under the ROC Curve. In: *Proceeding of Data Mining Workshops, ICDMW '09. IEEE International Conference*, 400–405.

Wang, R. – Tang, K. 2009. Feature Selection for Maximizing the Area Under the ROC Curve. In: *Proceeding of Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference.* 400–405.

Ward, N. – Tsukahara, W. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics* 32: 1177–1207.

Ward, N. 1997. Responsiveness in dialog and priorities for language research. *Systems and Cybernetics* 28: 521–533.

Watts, R. J. 1986. Relevance in conversational moves: A reappraisal of 'well'. *Studia Anglica Posnaniensia* 19: 37–59.

Watts, R. J. 1989. Taking the pitcher to the 'well'. Native speakers' perception of their use of discourse markers in conversation. *Journal of Pragmatics* 13 : 203–237.

Wegmann, S. – Scattone, F. – Carp, I. – Gillick, L. – Roth, R. – Yamron, J. 1998. Dragon system's 1997 broadcast news transcription system. In: *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, USA.

Wells, B. – Peppé, S. 1996. Ending up in Ulster: Prosody and turn-taking in English dialects. In: Couper-Kuhlen, E. – Selting, M. *Prosody in Conversation: Interactional studies*. 101–130.

Wennerstrom, A. – Andrew F. S. 2003. Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes* 36/2: 77–107.

Wilcox, L. – Chen, F. – Kimber, D. – Balasubramanian, V. 1994. Segmentation of speech using speaker identification. In: *Proceedings of IEEE International*

Conference on Acoustics, Speech and Signal Processing, Adelaide, Australia, 161–164.

Willsky, A. S. – Jones, H. 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.

Wingfiled, A. – Titone, D. 1998. Sentence Processing. In: Geason B. J. – Ratner, B. N. (eds.). *Psycholinguistics*. Harcourt Brace College Publishers, New York, London. 227–274.

Woodland, P. – Gales, M. – Pye, D. – Young, S. 1997. The development of the 1996 HTK broadcast news transcription system. In: *Speech Recorgnition Workshop*, 73–78.

Wooters C. – Huijberts M. 2007. The ICSI RT07s speaker diarization system. In Proceedings of of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop, 2007. Baltimore, MD.

Wooters, C. – Fung, J. – Peskin, B. – Anguera, X. 2004. Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system. In: *Fall 2004 Rich Transcription Workshop*, Palisades, NY.

Wu, T. – Lu, L. – Chen, K. – Zhang, H.-J. 2003a. UBM-based incremental speaker adaptation. In: *Proceedings of IEEE International Conference on Multimedia & Expo*, 721–724.

Wu, T. – Lu, L. – Chen, K. – Zhang, H.-J. 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, T. – Lu, L. – Chen, K. – Zhang, H.-J. 2003c. Universal background models for real-time speaker change detection. In: *International Conference on Multimedia Modeling*.

Xiao, B. – Ghosh, P. K. – Georgiou, P. – Narayanan, 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, M. – Yamashita, M. – Matsunaga, S. 2005. Spectral cross-correlation features for audio indexing of broadcast news and meetings. In: *Proceedings of International Conference on Speech and Language Processing*.

Yamamoto, K. – Asano, F. – Yamada, T. – Kitawaki, N. 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 S.H. – Valente F. 2012. Speaker diarization of overlapping speech based on silence distribution in meetings recordings. In: Proc. Interpseech, Portland, USA, September 2012.

Yella, Sree Harsha – Bourlard, Hervé 2013. Improved Overlap Speech Diarization of Meeting Recordings using Long-term Conversational Features, In: ICASSP, 2013

Ying, D. – Yan, Y. – Dang, J. – Soong, F. 2011. Voice Activity Detection Based On An Unsupervised Learning Framework. In: *IEEE Transactions on Audio, Speech and Language Processing* 19/8: 2624–2633.

Yngve, V. 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 M. – Hernando J. 2011. The detection of overlapping speech with prosodic features for speaker diarization. In Proc. Interspeech 2011, 2011, pp. 32–35.

Zelenák, M.– Segura, C.– Hernando, J. 2010. Overlap detection for speaker diarization by fusing spectral and spatial features. In: *Proceeding of IINTERSPEECH 2010*, Makuhari, Japan, 2302–2305.

Zhang, Y. – Jin, R. – Zhou, Z.-H. 2010. Understanding Bag-of-Words Model: A Statistical Framework. *International Journal of Machine Learning and Cybernetics* 1/1: 43–52.

Zhou, B. – Hansen, J. 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, X. – Barras, C. – Lamel, L. – Gauvain, J.-L. 2006. Speaker diarization: from broadcast news to lectures. In: Renals, S. – Bengio, S. – Fiscus, J. G. (eds.) *Machine*

Learning for Multimodal Interaction. Lecture Notes in Computer Science. Spinger. 396–406.

Zhu, X. – Barras, C. – Meignier, S. – Gauvain, J.-L. 2005. Combining speaker identification and bic for speaker diarization. In: *Proceedings of International Conference on Speech and Language Processing*, *INTERSPEECH 2005.* Lisbon, Portugal. 2441–2444.

Zochova, P. – Radova, V. 2005. Modified DISTBIC algorithm for speaker change detection. In: *Proceedings of International Conference on Speech and Language Processing*, Lisbon, Portugal.

11. RÖVIDÍTÉSEK JEGYZÉKE

VAD	Voice Activity Detection	Beszéddetektálás
MFCC	Mel Frequency Cepstral Coefficients	Mel-frekvenciás kepsztrális együttható
PLP	Perceptual Linear Prediction	Perceptuális Lineáris Becslés
GMM	Gaussian Mixture Models	kevert Gauss-modellek
SVM	Suport Vector Machine	Szupport vektor gépek
SWAMP	Sweeping Metric Parameterization	
MPEG-7	Multimedia content description interface	Multimédia tartalmakat leíró szabvány
CoG	Central of Gravity	Súlypont
BIC	Bayesian Information Criterion	Bayes-féle Információs Kritérium
ML	Maximum Likelihood	Maximum valószerűségi becslés
GLR	Generalized Likelihood Ratio	Általános valószerűség arány
MAP	Maximum A Posteriori	Maximum A Poszteriori becslés
KL	Kullback-Leibler distance	Kullback-Libler távolság
KL2	symmetrized Kullback-Leibler distance	szimmetrikus Kullback-Libler távolság
NLLR	Normalized Log-Likelihood Ratio	Normalizált valószerűségi arány logaritmusa
LLR	Log-Likelihood Ratio	Valószerűségi arány logaritmusa
DSD	Divergence Shape Distance	Eltérő alakú távolság

XBIC	cross-likelihood of BIC	A BIC kereszt-valószerűsége
HMM	hidden Markov-model	rejtett Markov-model
CuSum	Cumulative Sum algorithm	Kumulatív Összeg
VQ	Vector Quantization	Vektorkvantáló
MAD	Mean Absolute Deviation Statistic	Átlagos abszolút eltérés statisztika
RT	Rich Transcription	Információgazdag átirat
UBM	Universal Background Model	Általános háttér modell
ASR	Automatic Speech Recognition	Automatikus beszédfelismerés
MLLR	Maximum Likelihood Linear Regression	Maximum valószerűségi lineáris regresszió
NIST	National Institute of Standards and Technology	Nemzetközi, Szabványok és Technológiák Nemzetközi Intézete
M4	Multimodal Meeting Manager	
IM2	Swiss Interactive Multimodal Information Management	
AMI	Augmented Multi-party Interaction	
AMIDA	Augmented Multi-party Interaction projekt a Distant Access	
CHIL	Computers in the Human Interaction Loop	
F0	fundamental frequency	Alaphangmagasság
SAT	Speech Accommodation Theory	Beszéd akkomodációs elmélet
CAT	Communication Accommodation Theory	Kommunikáció akkomodáció elméletet
VOCSTAT	Vocal Channel Social Status Model	
DJ	Discours Marker	DiskurzusJelölő

DIT	Dynamic Interpretation Theory	Dinamikus Interpretáció Elmélet
DER	Diarization Error Rate	Beszélődetektálási hiba
FAR	False Acceptance Rat: False Positives	Az elsőfajú hiba: a téves elfogadás
FRR	False Rejection Rate: FRR; False Negatives rate	A másodfajú hiba: a téves elutasítás
HIT	Hit	Találat
NOHIT	No hit	Nincs találat
ROC	Receiver Operating Caracteristic	
EER	Equal Errror Rate	Egyenlő arányú hiba
ITU	International Telecommunication Union	Nemzetközi Telekommunikációs Unió
LPC	Liear Prediction Coeffiecients	Lineáris Predikciós Együttható