EGYSZERRE BESZÉLÉSEK DETEKTÁLÁSA A BESZÉLŐDETEKTÁLÁS JAVÍTÁSÁHOZ

Beke András

Bevezetés

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), ezért a beszélődetektáláskor valamennyiük modellezésére szükség van (Boakye et al. 2008, 2011; Zelenák et al. 2010).

A társalgás alapegysége a beszédforduló (angol terminusban *turn).* A beszédforduló során a társalgás egyik résztvevője beszél, amíg át nem adja, vagy amíg át nem veszik tőle a beszéd jogát (Sacks et al. 1974). A társalgásban az elméletek szerint egyszerre csak egy beszélő beszél, ezért az átfedések, vagyis az egyszerre beszélések és a hosszabb szünetek hibának minősülnek, amelyeket a beszélők igyekeznek javítani (Boronkai 2008a, b). Heldner és Edlund (2010) tanulmányában a beszélőváltások lehetséges módozataival foglalkozott akusztikai aspektusból. Az elméletek szerint három lehetséges módon történhet beszélőváltás: a) szünet van a két beszélő megnyilatkozása között, b) a két beszélőtől származó megnyilatkozás időben átfedi egymást, vagy c) sem szünet, sem átfedő beszéd nem történik. Sacks, Schegloff és Jefferson (1974) munkájukban megkülönböztetnek időközt (hosszabb vagy megnyúlt szünet), hiányosságot (rövidebb szünet) és átfedéseket (átfedő beszéd), valamint nincs-szünet-nincs-átfedést.

Az átfedő beszéd aránya a spontán társalgásokban meglehetősen nagynak mondható (Gráczi–Bata 2010). Beattie (1983) a beszélőváltásokat elemezve 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 (2005) 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 társalgásokat elemzett. Ez a magas előfordulási szám az át-

fedő beszéd funkciójából adódik. A társalgás során ugyanis az egyszerre beszélés kettős funkciót tölt be. Egyrészt megerősítő szerepe van (pl. *igen*, *aha*, *ühüm*, amelyek háttércsatorna-jelzések), 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).

A társalgásban a beszélők tapasztalati úton érzékelik a beszélőváltásra alkalmas helyeket, amelyekhez bizonyos szupraszegmentális, szemantikai, pragmatikai kulcsok kombinációját alkalmazzák/alkalmazhatják. A beszélőváltások vizsgálatára számos kutatás történt. Vizsgálták a beszélőváltások dallammenetét (Wells−Peppé 1996), előre jelezhetőségét (De Rulter et al. 2006), a zönge minőségét (Ogden 2004), a magyar nyelvben a fonetikai megvalósulásukat és szintaktikai szerkezetüket (Markó 2005; Bata 2009; Gráczi–Bata 2010).

A diskurzuselemzés felől jelentős mennyiségű munka foglalkozott már az egyszerre beszélésekkel (vö. Cetin–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 – illetve lényegében egyáltalán nem – feldolgozható szakaszai a beszédnek (Boakye et al. 2008). 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 modell egységességét gyengítheti, csökkentve ezzel a végleges beszélődetektálási eredményt. Ezért elengedhetetlen, hogy az átfedő részek gépi úton automatikusan azonosíthatók legyenek.

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 beszélődetektálás hiba arányának 17%-át a téves elutasítások száma adja, amelyet az átfedő beszédrészek okoznak.

Az egyszerre beszéléseket modellező munkák száma relatíve kevés, és azok közül is csak néhány kutatásban mutatták ki, hogy csökkenti a beszélődetektálási hibaarányát (DER: Diarization Error Rate) (Boakye et al. 2008; Boakye 2008; Trueba-Hornero 2008).

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 zöngeminőségét 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. Jóllehet ezek az eredmények messze elmaradnak a várttól, 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 (Wooters–Huijbert 2007).

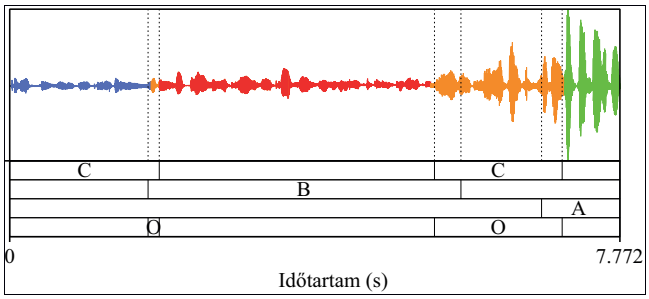
A jelen kutatás célja, hogy a spontán társalgásokban modellezze az egyszerre beszéléseket, és automatikus osztályozó algoritmussal különítse el azoktól a beszédszakaszoktól, ahol csak egy társalgó beszél. A kutatás további célja, hogy az egyszerre beszélések detektálásával javítsuk a beszélődetektálás eredményeit. 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. Feltételeztük továbbá azt is, hogy az egyszerre beszéléseket detektáló algoritmusnak a beszélődetektálóba való implementálásával a DER értéke csökkenthető.

Anyag, kísérleti személyek, módszer

A vizsgálatban a BEA adatbázisból (Gósy 2012) 100 társalgást választottunk ki, amely 55 órányi társalgást jelent. A társalgásokban minden esetben három személy vett rész. Ebből két társalgó állandó volt (2 nő, életkoruk 33 év). A harmadik személy 43 férfi és 57 nő közül került ki, átlagos életkoruk 35 év.

A felvétel minősége laboratóriumi körülményekhez hasonló. A felvételt egy AT–4040 irányított 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 beszédje-le volt a legerősebb, míg a kísérletvezető, illetve egy másik bevont személy beszédjele gyengébb volt. Ez megnehezítette az egyszerre beszélések auto-matikus osztályozását. Lehetőség lett volna normalizációs eljárásokat hasz-nálni, de feltehetően a zajt is felerősítette volna.

A társalgásokban manuálisan jelöltük azokat a részeket, ahol egyszerre több adatközlő beszél, illetve azokat a részeket, ahol csak egy beszélő beszél (1. ábra).



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

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

Módszer

Jóllehet az egyszerre beszélések automatikus osztályozása egyszerű fela-datnak tűnik, megvalósítása korántsem triviális. A beszélődetektálás egyik alapfeladata, mégis csak néhány olyan tanulmány ismert, amely megfelelő eredménnyel tudta megvalósítani az egyszerre beszélések automatikus osztá-lyozását (vö. Boakye et al. 2008).

A jelen kutatásban egy ANN/SVM hibrid rendszert (Artificial Neural Network/Support Vector Machine: Mesterséges Neuronháló/Szupportvektor-gép) 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, amellyel jól megragadha-tók az egyszerre beszélések. Mivel nem ismert, hogy mely akusztikai para-méter mentén különülnek el az átfedő beszédrészek és a nem átfedő beszéd-részek, több akusztikai jellemzőt is teszteltünk, mint az FFT-spektrum, rész-

sáv-energia (subband-energy), MFCC és Mel-skála szerinti logaritmikus szűrőbank. 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 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 előzetes feltételeknek kell megfelelniük az adatoknak a becslést megelőzően – ilyen követelmény a normál eloszlás. A GMM alternatívájaként léteznek más mesterséges neuronhálók, mint a MLP (Multilayer Perceptron; Bourlard–Morgan 1993). Az elmúlt években az ANN egy új fajtája jelent meg: ún. mély neuronhálók, amely vizsgálatok szerint igen jól alkalmazhatók a beszédhang-felismerésben (Dahl et al. 2010; Tóth–Grósz 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 kombinációi jól alkalmazhatók automatikus osztályozásához (Bellili et al. 2001).

1. Jellemzőkinyerés. Az egyszerre beszélések jó megfeleltethetőségéhez az akusztikai beszédjelből különböző jellemzőket nyertünk ki. A mély neuronhálókat a hangfelismerésben oly módon szokás alkalmazni, hogy előfeldolgozásként a hangot valamilyen képformátumúvá alakítjuk. Ennek egyik legegyszerűbb formája a különféle spektrumok, vagy az emberi hangok esetén a Mel-skálázott spektrogram. A jelen kutatásban négy különböző eljárást teszteltünk a beszédhang képpé alakítására.

(i) A **spektrum (SP)** kiszámolásához 256 pontos FFT-analízist használtunk Hamming-ablakkal, az ablak hossza 32 ms volt (8000 Hz-es mintavételezés esetén), 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 MFCC-algoritmust használtuk (vö. Ellis 2005). A jellemzők száma egy-egy időkeretben 39: a szokásos 12 MFCC koefficiens + az energia logaritmusa + ezek el-

ső 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 kepsztrális 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 **részsáv-energiát** (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) dimenziócsökkentettük 12-re (vö. Sarikaya et al. 1998).

Mindegyik jellemző esetén a különféle zajok – elsősorban a konvolúciós zajok (pl. 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.

2. Lényegkiemelés

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, amelyet a felsőbb szinteknek továbbítanak, és egyre feljebb jutva az egyes szinteken egyre általánosabb jellemzőket állapítanak meg.

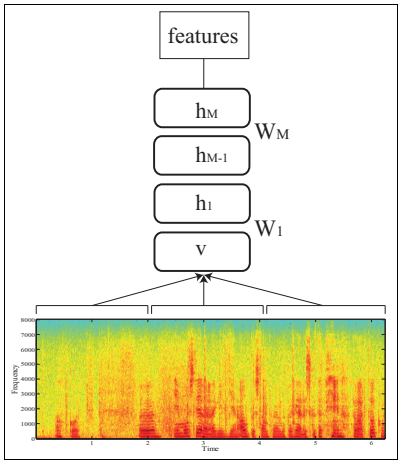
A korlátozott Boltzmann-gép (RBM: Restricted Boltzmann Machine) alapvetően két különböző réteget tartalmaz: látható és rejtett réteget. A korlátos jelző arra utal, hogy a neuronok között csak akkor van összeköttetés, ha az egyik a látható, a másik pedig a rejtett réteghez tartozik. Az azonos rétegbe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

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

A korlátozott Boltzman-gép tanító algoritmusa a CD-algoritmus (kontrasztív divergencia). A CD-algoritmus felügyelet nélküli tanulást végez,

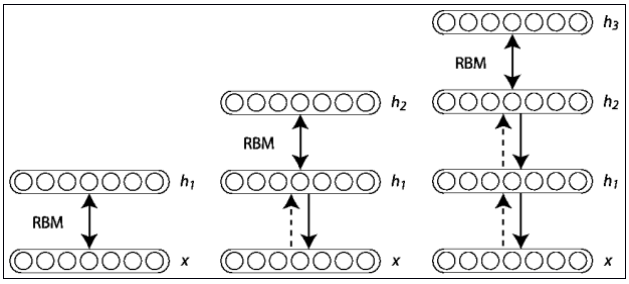
amely a „maximum likelihood”-tanítás közelítését adja. Ezt a folyamatot az RBM előtanításának nevezzük (Tóth–Grósz 2013).

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 (2. ábra).



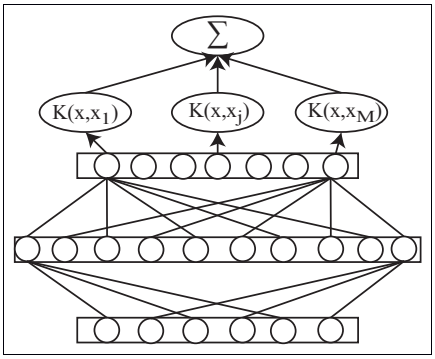
2. ábra
Jellemzőkinyerés korlátozott Boltzman-géppel

A korlátozott Boltzmann-gép igen jónak bizonyult a képi feldolgozásban. 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 (3 ábra).



3. ábra
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 szupportvektorgéppel valósítottunk meg (4 ábra).



4. ábra
Szupportvektorgép mély neuronhálóval előtanítva

Az RBM előtanítási paraméterei. Az RBM előtanításához az akusztikai paramétereket 15 keret hosszúságú csúszóablakkal nyertük ki. Mindegyik összefüggő ablakot felhasználtuk az RBM tanításához. Az RBM látható egységeinek száma a jellemzővektor dimenziószámának a keret hosszával képzett szorzata. Minden egyes audio szegmensre az érvényes konvolúcióval kifejezve $m - n + 1$ összefüggő ablak adódik, ahol m a keretek száma, n a csúszóablak hossza. A mély rétegű neurális hálózatok (DBN: Deep Belief Network) 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 iterációt használtunk, a többi rétegben 25-öt.

Az RBM megvalósításához Kyung Hyun Cho a Matlab-ban írt GitHub toolbox-át használtuk (Cho 2014).

3. Osztályozás

Az átfedő és nem átfedő beszédrészeket szupportvektorgéppel (SVM) kernelfüggvényként radiális bázisfüggvényt (RBF) alkalmazva osztályoztuk.

Szupportvektorgép (Support Vector Machine). Az SVM olyan matematikai konstrukció, amelyet döntési problémák megoldásához szoktak alkalmazni. Alapverziója a lineáris osztályozók családjába tartozik, de bináris osz-

tályozási 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 (Burges 1998). Az SVM-et alapvetően lineárisan szeparálható esetekre találták ki. A valóságban azonban a legtöbb probléma nemlinearitása olyan nagyságrendű, hogy az osztályozó nem lesz hatékony.

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

A jelen kutatásban az SVM egy változatát használtuk, amely 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 a teszthalmazt használhatjuk. 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 szupportvektorgépnek két szabadon állítható paramétere van: C a hibázási paraméter (penalty parameter) és γ az RBF kernelfüggvény (Gauss-függvény) szóráspa-ramétere. 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). A fentieket elvégezve az SVM számos lehetséges C és γ paraméterpárjára (kimerítő keresés, grid-search) megtalálhatjuk az optimális beállítást, vagyis amikor az SVM a legnagyobb felismerési arányokat éri el. Hsu, Chang és Lin (2003) szerint a C és γ értékeket az alábbi tartományokban érdemes keresni:

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

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

Az SVM tanítási paraméterei. Az átfedő és nem átfedő beszédrészek osztályozásához 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) részsáv-energiára 2×12.

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

Az osztályozás kiértékelése

(i) DET (Detection Error Tradeoff). Az osztályozásra alkalmazott algoritmusok működésének kiértékelésére és összehasonlítására a DET (Detection Error Tradeoff, Martin et al. 1997) kiértékelő algoritmust használtuk. A DET kiértékeléséhez először bemutatjuk a bináris osztályozás esetén a tévesztési mátrixot (1. táblázat).

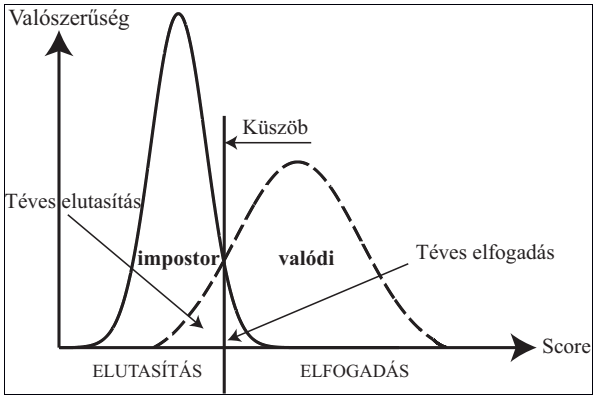
1. táblázat: A bináris osztályozás esetén a tévesztési mátrix

		Aktuális feltétel	
		Pozitív	**Negatív**
A 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) (5. ábra). A jelen munka során a téves elutasításról akkor beszélnénk, ha a beérkező szegmens átfedő beszéd, de nem fogadja el annak a gép.

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 dönt 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

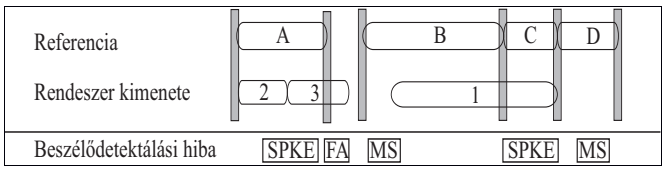
technikák, mint a ROC (Receiver Operating Caracteristic), vagy a DET (Detection Error Tradeoff). A DET-ben úgy választjuk meg a küszöbértéket, hogy az elsőfajú hiba és a másodfajú hiba egyenlő legyen. Ezt úgy hívják, hogy Equal Error Rate.



5. ábra
A bináris osztályozáskor fellépő hibák sematikus ábrázolása

(ii) 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, sem a nem beszédnek. Ennek mérésére az MD-eval-v12.pl-t, a NIST MD-eval-v12 DER kiértékelő szkriptjét használtuk.

Mivel a váltási pontok meghatározása a feladat, a rendszer hipotéziseként a beszélődetektálás kimenetében nem kell explicite meghatározni a beszélő nevét vagy identitását, ezért a beszélőkhöz rendelt azonosító címkéknek nem kell azonosnak lenniük a bemeneti (kézi) címkében és a kimeneti (automatikus) címkében. Ez a feladat tehát nem olyan, mint a beszéd-nem beszéd automatikus címkézése, amely során a szegmenst azonosító címkének egyezni kell a bementi és a kimeneti címkében (6. ábra).



6. ábra
A DER kiértékelési metódus sematikus ábrázolása

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

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

ahol az S az összes beszélői szegmens száma, ahol mind a hipotetikus, mind a referencia címke tartalmazza ugyanazokat a beszélői párokat. Ezt úgy kapjuk meg, hogy összehasonlítjuk a hipotetikus, illetve a referencia-beszédfordulókat. A N_{ref} és a N_{sys} kifejezések a beszélők számát jelölik a beszédszegmensben, az $N_{correct}$ 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 beszéd-, 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:

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

ahol a T_{score} 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:

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

amit csak azon szegmensekben mérünk, amely a referenciacímkében nem-beszédrészként 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:

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

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 kiértékelő script megadja minden egyes referencia-hipotetikus szegmentációra a DER értékét, illetve az összes kiértékeléshez használt fájlra ad egy súlyozott átlagot.

Eredmények

Az egyszerre beszélések időtartama 12%-át teszi ki a teljes korpusznak, míg a szünetek időtartama 10,9%-át; a beszédrészek tehát a teljes korpusz 77,1%-a. Az átlagos átfedőbeszéd-arány a felvételekben 21,84%.

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 mély rétegű neuronhálózat harmadik rétegében hány neuront használunk.

Az eredmények azt mutatják (2. táblázat), hogy a négy akusztikai paraméter [FFT spektrum (SP); Mel-frekvenciás kepsztrális (MFC) együtthatók; Mel-skála szerinti logaritmikus szűrőbank (MSL); részsáv-energia (RSE)]

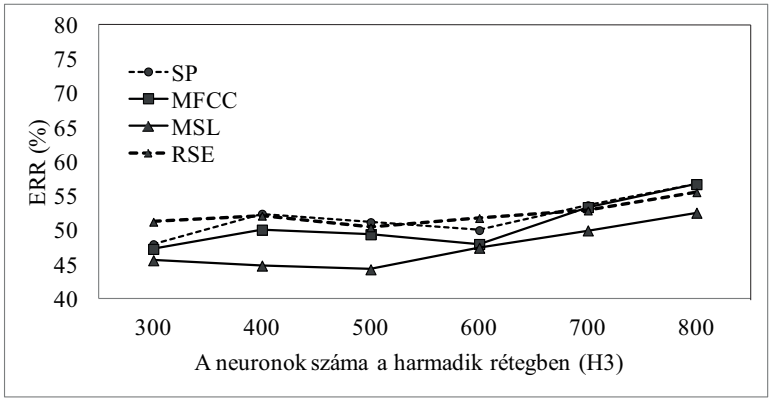
közül a legjobb teljesítményt akkor kaptuk, ha jellemzőként a Mel-skála szerinti logaritmikus szűrőbankot alkalmaztuk. Ekkor az Equal Error Rate (EER) átlagos értéke 47,49%, vagyis a helyesen felismert szegmensek aránya átlagosan 52,51%.

2. táblázat: Az átlagos EER értéke az akusztikai paraméterek függvényében

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

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

Megvizsgáltuk, hogy az EER értéke hogyan függ a jellemzők és a harmadik rétegben használt neuronok számától. Az eredmények azt mutatják, hogy a legjobb eredményt akkor kapjuk, ha MSL-jellemzőt és 500 neuront használunk a H3-ban (7. ábra).



7. 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 (3. táblázat): MSL-MFCC: $Z = -2{,}201$; $p = 0{,}028$; MSL-SP: $Z = -2{,}201$; $p = 0{,}028$; MSL-RSE: $Z = -2{,}201$; $p = 0{,}028$.

3. táblázat: Az EER értéke a jellemzők és a H3-ban alkalmazott neuronok számának függvényében

	Neuronok száma a H3 rétegben	Akusztikai jellemzők			
		SP	**MFCC**	**MSL**	**RSE**
EER (%)	300	**48,00**	**47,31**	45,65	51,27
	400	52,45	50,12	44,87	52,15
	500	51,22	49,44	**44,33**	**50,45**
	600	50,05	48,02	47,48	51,81
	700	53,68	53,41	50,01	52,91
	800	56,77	56,76	52,59	55,58
	900	52,03	50,84	47,49	52,36

Az EER-értékekből 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 elmondható, hogy 500 neuron felett mindegyik jellemző esetében nőtt az EER értéke.

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

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

Az egyszerrebeszélés-detektáló implementálása a beszélődetektálóba

A beszélődetektáló felépítése

A beszélődetektáló rendszerben megvalósítottunk egy beszélőszegmentáló és egy beszélőklaszterező eljárást. A beszélőszegmentáló algoritmus kétlépcsős. Az első lépés egy BIC-alapú szegmentáló, amely MFCC-ket használ 2,5–3,5 kHz-es tartományban. A második lépésben pedig a téves riasztások kompenzálására egy Kullback–Leibler-divergencia alapú szegmentálót hoztunk létre, amelynek szintén az $MFCC_{2,5–3,5}$ volt a bemeneti jellemzője. A beszédfordulók detektálása után az egyes beszédszegmenseket beszélőkhöz rendeltünk, ami beszélőklaszterezési feladat. A klaszterezés bemeneti jellemzője a GMM-UBM-PCA (Gauss-keverék modell, általános háttérmodell és

főkomponens-analízis) *i*-vektorok. A klaszterezést pedig BIC-alapú nemellenőrzött módszerrel végeztünk el.

Eredmények

Az átfedő beszédrészek automatikus detektációjával átlagosan 2,5%-os relatív javulást tudtunk elérni, vagyis a DER értékét 31,21%-ról le tudtuk csökkenteni 28,71%-ra (4. táblázat). Ez a javulás szignifikáns (Wilcoxon-próba Monte Carlo szimulációval kiegészítve: $Z = -3{,}06$; $p = 0{,}002$).

4. táblázat: A DER értéke az átfedőbeszéd-detektálóval és anélkül

Felvétel sorszáma	DER		ΔDER	Az átfedő beszéd és a társalgás hosszának aránya
	Átfedő beszéd			
	tartalmaz	nem tartalmaz		
bea071n037	14,98%	12,36%	−2,62%	21,52%
bea072n038	27,83%	24,85%	−2,98%	38,62%
bea073n039	35,64%	33,70%	−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%	34,60%	−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	31,80%	30,54%	−1,26%	40,66%
Átlagos	**31,21%**	**28,71%**	**−2,5%**	**30,94%**

Elemeztük, hogy a teszteléskor használt társalgásokban milyen arányban fordulnak elő egyszerre beszélések (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.

Következtetések

Az egyszerre beszélések magas, 12%-os előfordulása a korpuszban indokolja, hogy a beszélődetektálásban foglalkozzunk ezen jelenség automatikus osztályozásának lehetőségével. Jóllehet az egyszerre beszélések automatikus osztályozása igen fontos feladat a beszélődetektálásban, mégis csak néhány tanulmány foglalkozik ezzel a kérdéssel (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 átfedőbeszéd-detektálásra. Yella és Valente (2012) munkájukban azt a jelenséget igyekeztek mo-

dellezni, 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 Zelenák és Hernando (2011) hasonló F-score-t tudtak elérni az átfedőbeszéd-detektálásra, amely 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 Ben-Harush és munkatársai (2009) az időtartományban adott entrópia jellemzők becslésével próbálta meg detektálni az egyszerre beszéléseket, ez a munka azonban csak kétbeszélős társalgásokra vonatkozik.

Yella és Bourlard (2013) Shriberg és munkatársainak 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 létrehoztak egy olyan algoritmust, 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 habár az egyszerre beszélések detektálásának eredménye jóval elmarad a kívánttól, a beszélődetektálóba való integrációja 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 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 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. Az EER értéke ekkor 44,33% volt. Kimutattuk, hogy a mély neuronhálók alkalmasak a jelen problémában a jellemzők kialakítására 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 egyszerrebeszélés-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 vagy 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 előzetes feltételezésüknek megfelelően, ha az egyszerre beszéléseket detektáló algoritmust integráltuk a beszélődetektálóba, akkor annak DER-értékét csökkenteni tudtuk, vagyis a beszélődetektáló eredményei javultak. Az átfedő beszéd automatikus detektációjával a DER értékét 31,21%-ról 28,71%-ra tudtuk csökkenteni, így átlagosan 2,5%-os relatív javulást lehetett elérni.

Összességében tehát elmondható, hogy bár az egyszerre beszélések detektálása az általunk kialakított módszerrel még mindig az elméletileg lehetséges értéknél alacsonyabb eredménnyel működik, mégis alkalmas arra, hogy a beszélődetektálóba integrálva növelje annak eredményességét.

Irodalom

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

Beattie, Geoffrey 1983. *Talk: An analysis of speech and non-verbal behaviour in conversation*. Open University Press, Milton Keynes.

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

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

Boakye, Kofi A. 2008. *Audio segmentation for meetings speech processing*. PhD dissertation, University of California, Berkeley.

Boakye, Kofi A. – Vinyals, Oriol – Friedland, Gerald 2011. Improved overlapped speech handling for speaker diarization. In: *Proceeding of Interspeech 2011*. Firenze, Italy. 941–944.

Boakye, Kofi A. – Trueba-Hornero, Beatriz – Vinyals, Oriol – Friedland, Gerlad 2008. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *Proceeding of ICASSP*. 4353–4356.

Boronkai Dóra 2008a. Konverzációelemzés és anyanyelvtanítás I. *Anyanyelv-pedagógia* 2008/2. http://www.anyanyelv-pedagogia.hu/cikkek.php?id= 60.

Boronkai Dóra 2008b. Konverzációelemzés és anyanyelvtanítás II. *Anyanyelv-pedagógia* 2008/3–4. http://www.anyanyelv-pedagogia.hu/cikkek.php?id=115.

Bourlard, Hervé – Morgan, Nelson 1993. Continuous speech recognition by connectionist statistical methods. *IEEE Transactions on Neural Networks* 4/6. 893–909.

Burges, Christopher J. C. 1998. A Tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2. 121–167.

Cetin, Ozgür – Shriberg, Elizabeth 2006. Analysis of overlaps in meetings by dialog factors. Hot spots, speakers, and collection site: Insights for automatic speech recognition. In: *Proceedings of ICSLP*, Pittsburgh. 293–296.

Chang, Chih-Chung – Lin, Chih-Jen 2012. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2/3. 27. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cho, KyungHyun 2014. *Advances in deep learning (draft)*. PhD dissertation. Aalto University, Aalto.

Cho, KyungHyun. The RBM code for Matlab. http://users.ics.tkk.fi/kcho/.

Dahl, George. E. – Ranzato, Marc'Aurelio R. – Mohamed, Abdel-rahman – Hinton, Geoffrey 2010. Phone recognition with the mean-covariance restricted Boltzmann machine. In: *NIPS 2010*. 469–477.

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

Ellis, Dan 2005. Reproducing the feature outputs of common programs using Matlab and melfcc.m. http://labrosa.ee.columbia.edu/matlab/rastamat/mfccs.html

Gósy, Mária 2012. BEA - A multifunctional Hungarian spoken language database. *The Phonetician* 105–106. 50–61.

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.

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

Heldner, Mattias – Edlund, Jens 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38. 555–568.

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, Taipei.

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

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

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

Martin, Alvin F. – Doddington, George R. – Kamm, Terri – Ordowski, Mark – Przybocki, Mark A. 1997. The DET curve in assessment of detection task performance. In: *Proceedings of Eurospeech*. Rhodes, Greece. 1899–1903.

Moattar, Hossein – Homayounpour, Mohammed M. 2006. Speech overlap detection using spectral features and its application in speech indexing. *Information and Communication Technologies* 2/1. 1270–1274.

Mohamed, Abdel-rahman – Hinton, Geoffrey – Penn, Gerald 2012. Understanding how deep belief networks perform acoustic modelling. In: *Proceedings of ICASSP*. 4273–4276.

Mowlaee, Pejman – Christensen, Mads G. – Tan, Zheng-Hua – Jensen, Søren 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)*. 538–541.

Neuberger, Tilda – Beke, András 2013. Automatic laughter detection in spontaneous speech using GMM-SVM method. In: *Proceedings of TSD2013*, Springer Berlin Heidelberg. 113–120.

Ogden, Richard 2004. Non-modal voice quality and turn-taking in Finnish. In Couper-Kuhlen, Elisabeth − Ford, Cecilia E. (eds.): *Sound patterns in interaction*. Benjamins, Amsterdam. 29−62.

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

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

Saeidi, Rahim – Mowlaee, Pejman – Kinnunen, Tom – Tan, Zheng-Hua – Christensen, Mads G. – Jensen, Søren H. – Franti, Pasi 2010. Improving monaural speaker identification by double-talk detection. In: *Eleventh Annual Conference of the International Speech Communication Association*. 1069–1072.

Sarikaya, Ruhi – Pellom, Bryan L. – Hansen, John H. L. 1998. Wavelet packet transform features with application to speaker identification. In: *Proceedings of IEEE Nordic Signal processing Symposium*. Visgo, Denmark. 81–84.

Shriberg, Elizabeth – Stolcke, Andreas – Baron, Don 2001. Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation. In: *Proceedings of EUROSPEECH*. Aalborg, Denmark. 1359–1362.

Suykens, Johan A. K. – Van Gestel, Tony – De Brabanter, Jos – De Moor, Bart – Vandewalle, Joos 2002. Least squares support vector machines. *World Scientific* 4. Singapore.

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

Tóth, László –Grósz, Tamás 2013. A comparison of deep neural network training methods for large vocabulary speech recognition. In: *Proceedings of TSD2013*. Springer Berlin Heidelberg. 36–43.

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

Vipperla, Ravichander – Geiger, Jürgen T. – Bozonnet, Simon – Wang, Dong – Evans, Nicholas – Schuller, Björn – Rigoll, Gerhard 2012. Speech overlap detection and attribution using convolutive non-negative sparse coding. In: *Proceedings of ICASSP-12*. 4181–4184.

Wells, Bill – Peppé, Sue 1996. Ending up in Ulster: Prosody and turn-taking in English dialects. In Couper-Kuhlen, Elisabeth – Selting, Margret (eds.): *Prosody in*

Conversation: Interactional studies. Cambridge University Press, Cambridge–New York–Melbourne. 101–130.

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

Yella, Sree Harsha – Bourlard, Hervé 2013. Improved overlap speech diarization of meeting recordings using long-term conversational features. In: *Proceedings of ICASSP*. 7746–7750.

Yella, Sree Harsha – Valente, Fabio 2012. Speaker diarization of overlapping speech based on silence distribution in meetings recordings. In: *Proceedings of Interpseech 2012*. Portland, USA.

Zelenák, Martin – Hernando, Javier 2011. The detection of overlapping speech with prosodic features for speaker diarization. In: *Proceedings of Interspeech 2011*. 32–35.

Zelenák, Martin – Segura, Carlos – Hernando, Javier 2010. Overlap detection for speaker diarization by fusing spectral and spatial features. In: *Proceedings of Interspeech 2010*, Makuhari, Japan. 2302–2305.

A kutatás a 108762-es számú OTKA-pályázat keretében készült.