Többsávos, zajtűrő beszédfelismerés mély neuronhálóval

Kovács György[1], Tóth László[2]

[1] KU Leuven, Department of Electrical Engineering
Leuven, Kasteelpark Arenberg 10, e-mail: gkovacs@esat.kuleuven.be
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103., e-mail: tothl@inf.u-szeged.hu

Kivonat Miközben az automatikus beszédfelismerés terén jelentős előrelépések történtek az elmúlt években, a beszédfelismerő rendszerek eredményessége spontán vagy zajjal szennyezett beszéd esetén továbbra sem kielégítő. Ezen probléma kiküszöbölésére számos módszert javasoltak, melyek közül több jól kiegészíti egymást. Jelen cikkünkben három ilyen módszer, az ARMA spektrogram, a neuronháló-tanítással egyidejűleg optimalizált spektro-temporális jellemzőkinyerés, és a többsávos feldolgozás kombinációját vizsgáljuk az Aurora-4 beszédfelismerési feladaton.[1]

Kulcsszavak: zajtűrő beszédfelismerés, mély neuronháló, többsávos feldolgozás, ARMA, Aurora-4

1. Bevezetés

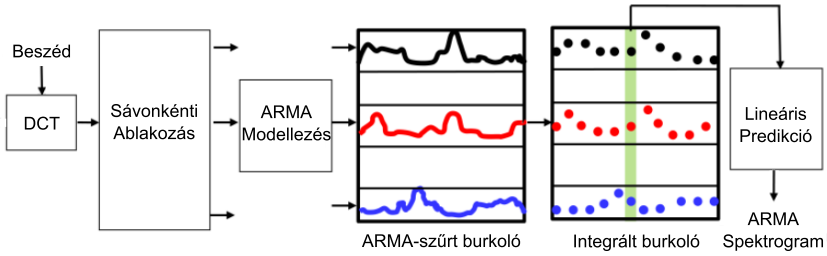
A beszédfelismerésben számos módszer létezik zajjal szennyezett beszéd pontosabb felismerésére. Jobb felismerési eredményeket érhetünk el például azáltal, ha a beszédnek olyan reprezentációját állítjuk elő, amely a hagyományos melskála szerinti sávszűrőknél (filter-bank) kevésbé érzékeny a beszédjelbe keveredő zajra. Ilyenre példa az itt használt ARMA spektrogram [1]. A felismerési pontosságot úgy is javíthatjuk, ha a spektrogramból zajtűrő jellemzőket próbálunk kinyerni. Ez motiválta a spektro-temporális jellemzőkinyerési módszerek, pl. Gábor-szűrők [2] bevezetését. Ezen módszerek egyik előnye, hogy a jellemzők előállításánál a spektrum korlátozott frekvenciatartományára támaszkodnak, így egy sávhatárolt zaj nem rontja el az összes jellemzőt. Hasonló megfontolásra épít a több-adatfolyamos (multi-stream) beszédfeldolgozás [3], amely a beszédjel különböző reprezentációit egymástól független taszkokban dolgozza fel, majd ezen taszkok eredményeit kombinálja. A többsávos (multi-band) beszédfeldolgozás a több adatfolyamos beszédfeldolgozás speciális esete, ahol az adatfolyamokat a beszédjel különböző frekvenciatartományaiból kinyert jellemzők jelentik. Jelen kísérleteinkben ezen megközelítéseket (zajtűrő reprezentáció használata, spektro-temporális jellemzők kinyerése, és többsávos feldolgozás) kombináljuk.

[1] A szerzők köszönetüket fejezik ki Sriram Ganapathy-nak az ARMA spektrogramokért, valamint Deepak Baby-nek az Aurora-4 használatában nyújtott segítségért.

2. Zajtűrő beszédfelismerési technikák

2.1. ARMA spektrogram

Ezen spektrogram előállítása az autoregresszív mozgóátlag (AutoRegressive Moving Average - ARMA) modell használatával történik. Ez a hagyományos AR modellezés [4] általánosítása. Az ARMA modellt a diszkrét koszinusz transzformáció (DCT) részsáv komponenseire alkalmazzák, a temporális burkológörbe becslésére. Az így kapott görbék rövid szakaszokon történő integrálásának eredménye egy spektrális reprezentáció. Az ARMA spektrogram ebből a spektrális reprezentációból lineáris predikción alapuló spektrális simítással áll elő. Ezt a folyamatot szemlélteti az 1. ábra. A cikkben felhasznált spektrogramokat Sriram Ganapathy (IBM T.J. Watson Research Center) bocsátotta rendelkezésünkre.



1. ábra. Az ARMA spektrogram előállítása (részletes leírásért, lásd [1]).

2.2. Spektro-temporális jellemzőkinyerés

A spektro-temporális jellemzőkinyerés során egy-egy jellemző előállításához a spektrális reprezentáció időben és frekvenciában korlátozott tartományát használjuk. Ennek egyik oka a megfigyelés, hogy az agykérgi neuronok a hangjel frekvencia- és időbeli (spektro-temporális) változásaira egyidejűleg reagálnak [5]. Így használatukkal az automatikus beszédfelismerésnél zajtűrőbb emberi hallás működéséhez közelíthetjük rendszerünket (ahogy azt tesszük a mel-skála használatával is). A spektro-temporális jellemzők másik előnye, hogy egy-egy jellemző a beszédjelnek csak egy korlátozott frekvenciatartományára támaszkodik, így egy sávhatárolt zaj jelenléte a beszédjelben nem érinti az összes jellemzőt.

2.3. Többsávos beszédfelismerés

A spektro-temporális feldolgozáshoz hasonlóan a többsávos beszédfelismerés esetében is az első lépés egymástól független akusztikus jellemzők kinyerése a különböző frekvenciasávokból. Ezért bizonyos esetekben a spektro-temporális beszédfelismerésre jellemzőrekombináció (feature recombination) néven, mint a többsávos feldolgozás egy speciális esetére hivatkoznak [6]. Ez is mutatja, hogy a többsávos beszédfelismerés módszerek széles skáláját fedi le. Az általunk követett módszer [7] esetében a különböző frekvenciatartományokból származó jellemzőkön különálló neuronhálókat tanítunk, majd ezen neuronhálók kimenetét egy kombinációs neuronháló bemeneteként használva végzünk újabb tanítást.

3. Mély neuronhálók

A mély neuronhálók jelentős előrelépést hoztak a beszédfelismerésben. Ezek olyan neuronhálók, melyek a hagyományosnál több rejtett réteget tartalmaznak. Ez nehézségekkel jár a háló tanítása során, melyek kiküszöbölésének érdekében változathatunk a tanítási algoritmuson, vagy a felhasznált neuronok típusán [8]. Cikkünkben ez utóbbi megoldás mellett döntöttünk, és neuronhálónkat rectifier ($max(0, x)$ függvényt megvalósító) aktivációs függvényt alkalmazó neuronokból (ún. ReLU-kból) építettük fel, miközben a tanítást továbbra is a hagyományos hibavisszaterjesztési (backpropagation) algoritmussal végezzük.

3.1. Együttes neuronháló-tanítás és jellemzőkinyerés

A spektro-temporális jellemzőkinyerést végrehajtó szűrők megvalósíthatók speciális, lineáris aktivációs függvényt alkalmazó neuronok formájában [9]. Könnyen belátható, hogy a lineáris neuronok alkalmasak erre, ha összehasonlítjuk a lineáris neuronok kimenetét és a szűrés eredményét meghatározó egyenleteket

$$\begin{aligned} o &= \left(\sum_{i=1}^{L} x_i \cdot w_i + b \right), \\ o &= \sum_{f=0}^{N} \sum_{t=0}^{M} A(f,t) F(f,t), \end{aligned} \tag{1}$$

ahol x a neuron bemeneti vektora, w a neuron súlyvektora, amelyek a megfelelő indexeléssel (továbbá feltételezve, hogy $L = N \cdot M$) megfeleltethetők A és F változóknak, ahol A az ablak amire az F szűrőt alkalmazzuk. Ekkor a b biast nullának választva, a két egyenlet egymás megfelelője. Erre az alapötletre építve mutattuk meg korábban, hogy a szűrők paramétereinek optimalizálása és a neuronháló betanítása egyetlen közös optimalizálási lépésként is elvégezhető [9]. Jelen cikkben is ezt a módszert alkalmazzuk.

3.2. Konvolúció

Az időbeli konvolúció egy megoldást ad arra, hogy a spektro-temporális ablakok méreténél hosszabb időintervallumból tudjunk információt kinyerni az ablakok megnövelése (és így további neuronháló-súlyok) bevezetése nélkül. Esetünkben ennek megvalósítása oly módon történik, hogy a lokális jellemzőket a neuronháló több, egymást követő időpontban nyeri ki, az egymást követő ablakokra azonos súlyokat alkalmazva, majd az eredmények közül csak minden negyediket ad át a következő rétegnek. Ezzel a konvolúció mindhárom követelménye, a lokális ablakok használata (local windows), a súlyok megosztása (weight sharing), és eredmények csoportjainak egyetlen értékkel történő reprezentációja (pooling) teljesül [10]. Arra vonatkozóan, hogy mely ablakokból származó eredményeket adjuk tovább, és mely ablakok eredményét nem vesszük a későbbiekben figyelembe, korábban több kísérletet végeztünk [11]. Az általunk használt neuronháló architektúra részletesebb leírása is megtalálható ebben a munkában.

4. Kísérleti beállítások

4.1. Előfeldolgozás

A bemenő beszédjelből először egy 39 csatornás ARMA spektrogramot számoltunk. A kapott ARMA spektrogramokat bemondásonként normalizáltuk oly módon, hogy átlaguk nulla, varianciájuk pedig egy legyen. Továbbá Ganapathy nyomán [1] derivatív jellemzőket számoltunk a spektrogramból, minden csatornára a szomszédos csatornák alapján ($band(K+1) - band(K-1)$). Így minden időpillanathoz egy 78 (39·2) komponensű jellemzővektort rendeltünk.

4.2. Aurora-4

Az Aurora-4 a Wall Street Journal beszédadatbázis zajjal szennyezett változata [12]. Elérhető (az általunk használt) 16 kiloherzes, valamint 8 kiloherzes mintavételezéssel. Az adatbázis két 7138 mondatból álló tanító halmazt, és 14, egyenként 330 mondatból álló teszthalmazt tartalmaz. A tanító halmaz első (tiszta) változata a mondatok zaj nélküli változatát tartalmazza Sennheiser mikrofonnal rögzítve, míg a második változatban az egyes mondatok különböző zajokkal szennyezettek, illetve rögzítésük eltérő mikrofonnal történt. Mivel cikkünkben azt kívántuk megvizsgálni, hogy a beszédfelismerő rendszer hogy teljesít általa nem ismert zajtípusok (illetve átviteli karakterisztika) esetén, ezért csak a tiszta tanító halmazt használtuk. A tanító halmaz kilencven százalékán tanítottuk a neuronhálók súlyait, míg a fennmaradó részt megállási feltételként (és validációs halmazként) használtuk.

A kiértékelést mind a 14 teszthalmaz felhasználásával végeztük. Ezen teszthalmazok ugyanazt a 330 mondatot tartalmazzák különböző verziókban: az első hét teszthalmazban lévő hangfájlok rögzítése a Sennheiser mikrofonnal történt, míg a második hét teszthalmazban ettől eltérő mikrofonnal rögzített felvételeket találunk. Mindkét (különböző mikrofonnal felvett) hetes csoport belső felosztása azonos: az első halmaz zajjal nem szennyezett beszédet tartalmaz, míg a következő hatban hat különböző zajjal (autó, csevegés, étterem, utca, repülőtér, vonat) szennyezett beszéd található.

4.3. Kaldi

Beszédfelismerési kísérleteinkben a Kaldi [13] Aurora-4 receptjéből indultunk ki, melyet a saját neuronhálónkkkal kombináltunk. A Kaldi recept egy HMM/GMM-et tanít be, majd kényszerített illesztést végez, ami minden adatvektorhoz az 1997 kontextusfüggő trifón állapot valamelyikét rendeli. A kombinációs neuronhálókat úgy tanítottuk be, hogy valószínűségeket biztosítsanak minden kerethez. A neuronháló kiértékelése után ezeket a valószínűségeket a Kaldi dekóderének inputjaként használtuk. A dekódolás során emellett egy trigram modellt, valamint egy 5000 szót tartalmazó szótárat használtunk.

4.4. Korábbi módszereink adaptálása

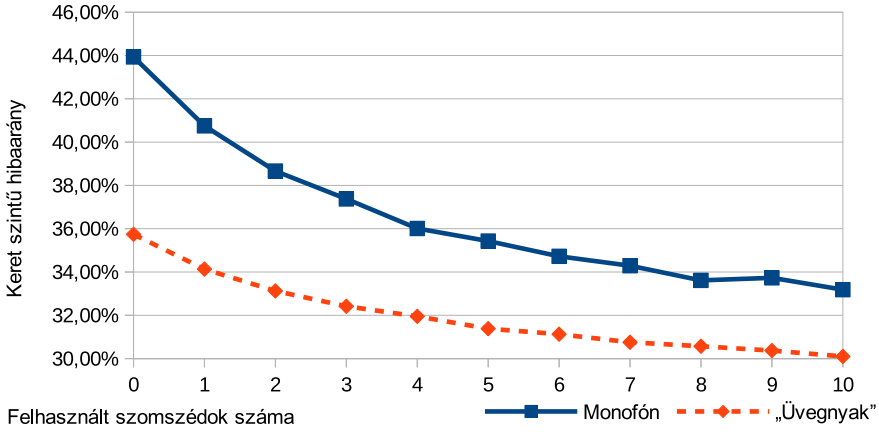
A korábbi munkáinkban felhasznált módszereket [7,11] jelen munkában adaptálnunk kellett az új spektrogramhoz, valamint adatbázishoz. A felhasznált architektúrát bemutató korábbi munkánkban [11] egy tükrözéssel 30 sávosra bővített 26 sávos spektrogramot használtunk a háló bemeneteként, így az ott leírt ablakméret és átfedés használatával hat, a szűrők alkalmazását megvalósító rétegre volt szükségünk a neuronhálóban. Jelen munkában ezen rétegek száma az ablakméret és átfedés megtartásával (de a tükrözés elhagyásával) 16-ra bővült (8-8 az eredeti spektrogramon, és annak derivatív változatán). Mivel az adatbázison végzett korábbi kísérleteinkben a szűrők irányított inicializálása nem javította a felismerést, jelen kísérleteinkben a szűrőket véletlen számokkal inicializáltuk. Adaptálnunk kellett továbbá a többsávos beszédfeldolgozás módszerét is: korábbi cikkünkben [7] a spektrogramot (a tükrözést nem számítva) 5 sávra bontottuk az ablakozásnak megfelelően. Jelen esetben mind a spektrogramból, mind a derivatív változatból egy-egy időszakaszon 8-8 ponton nyerünk ki ablakokat, így legjobban úgy tudtuk közelíteni a korábban használt paramétereket, hogy mind a kettőből 2-2 ablakot foglaltunk egy sávba, bemenetünket 4 sávra osztva.

További eltérés korábbi, a TIMIT beszédadatbázison végzett kísérleteinkhez képest, hogy az Aurora-4 esetén 42 monofón címkével ellátott címkézéssel rendelkeztünk, így a különböző sávokon tanított neuronhálók 42 osztálycímkére tanultak. (A kísérletek során ezen hálók kombinációjára „monofón" néven hivatkozunk). Mivel azonban a kombinációs hálók tanítása jelen kísérletek során 1997 állapotra történt, problémát jelenthet, hogy az eltérő szinteken lévő neuronhálók eltérő célfüggvényekkel tanulnak. Így további kísérleteket végeztünk, melyek során a különböző sávokon tanított neuronhálókat szintén 1997 állapotra tanítjuk. Ily módon azonban ha továbbra is ezen neuronhálók kimeneteit használnánk a kombinációs háló bemeneteként, jelentős dimenziószám-növekedéssel kéne szembenéznünk (168 helyett közel 8000 jellemző), ezért az alsó rétegben tanított neuronhálókba a kimeneti réteg elé egy második (50 lineáris aktivációs függvényt alkalmazó neuronból álló) „üvegnyak" (bottleneck) réteget illesztettünk, és ezen rétegek kimenetét használtuk a kombinációs háló bemeneteként. (A kísérletek során ezen hálók kombinációjára „üvegnyak" néven hivatkozunk).

4.5. Neuronháló paraméterek

Kísérleteinkben kétfajta neuronhálót használtunk: egyet, mely a különböző sávokon tanult (illetve összehasonlítási alapként ezeknek egy olyan változatát, mely az összes sávon tanult), valamint az előbbi hálók kimenetét bemenetként használó kombinációs hálókat. Ez utóbbiak ReLu hálók két, egyenként ezer neuronból álló rejtett réteggel, és 1997 kimeneti neuronnal. A különböző sávokon tanított hálók architektúrja a korábban leírtaknak megfelelő[2]. Az összehasonlítási alapként, az összes sávon egyszerre tanított (a kísérletek során jellemzőrekombinációsként

[2] A 4.4 szekcióban leírt módosításoktól eltekintve a hálók megegyeznek a korábbi cikkünkben [11] utolsóként leírt hálóval.



2. ábra. Keretszintű hibaarányok a validációs halmazon, a kombinációs háló által használt szomszédos keretek számának függvényében.

hivatkozott módszerhez tartozó) háló annyiban tér el ettől, hogy 4 helyett 16 szűrő réteggel rendelkezik, második bottleneck rétege 200 neuront tartalmaz, és minden köztes réteg kétszer annyi neuronból áll, mint az egyes sávokon külön tanított társaié (így a két változat paraméterszáma közel azonos).

5. Kísérletek és eredmények

Első lépésben a többsávos módszerek felismerési eredményeit hasonlítottuk össze, a kombinációs hálóban felhasznált szomszédok szempontjából. Ehhez betanítottuk a sávokat feldolgozó hálókat, valamint ezek kimenetén 3-3 kombinációs hálót minden konfiguráció esetére. A validációs halmazon kapott keretszintű hibaarányok leolvashatók a 2. ábráról. Megfigyelhető, hogy az „üvegnyak” módszerrel kapott eredmények konzisztensen felülmúlják a monofón módszer eredményeit. Bár a keretszintű eredmények a tizedik szomszéd felhasználásával is javulnak, mivel a görbe ellaposodik, valamint figyelembe véve azt a megfigyelést, hogy a kontextus kiterjesztése a keretszintű eredményeken akkor is javíthat, amikor a szószintű eredményekre már negatív hatással van [14], a teszteket a négy szomszédos keretet felhasználó változaton végeztük. (A monofón verzión végzett előzetes teszteredményeink igazolták a feltételezésünket, hogy négynél több szomszéd alkalmazása nem javítja tovább a szófelismerési pontosságot).

Hogy a szószintű felismerés hatékonyságát is elbírálhassuk, a kiválasztott, négy szomszédot használó hálókat kiértékeltük a teszthalmazokon szószintű hibákra. Ezt tettük a jellemzőrekombinációs háló kimenetén tanított kombinációs hálóval is (amely az összehasonlíthatóság miatt szintén négy szomszédot használt). Az összehasonlítás teljességéért hozzáadtuk a táblázathoz Sriram Ganapathy [1] eredményeit is (aki szintén az ARMA spektrogramot használva, de

1. táblázat. Szófelismerési-hibaszázalékok az Aurora-4 teszthalmazain.

Mikrofon	Zaj	Jellemzőrekombináció	Monofón	Üvegnyak	Ganapathy[1]
Azonos mikrofon	Tiszta	3,9%	3,8%	3,7%	**3,0%**
	Autó	6,4%	6,1%	5,8%	**5,0%**
	Csevegés	13,6%	13,9%	**12,6%**	13,0%
	Étterem	17,7%	18,8%	**17,2%**	17,3%
	Utca	**14,0%**	15,5%	**13,7%**	**13,6%**
	Repülőtér	14,0%	14,5%	**13,5%**	**13,7%**
	Vonat	**14,6%**	17,1%	**14,5%**	**14,5%**
	Átlag	12,0%	12,8%	**11,6%**	**11,4%**
Eltérő mikrofon	Tiszta	14,3%	**11,9%**	**11,6%**	**11,7%**
	Autó	21,7%	18,6%	**17,7%**	18,4%
	Csevegés	30,1%	29,8%	**27,5%**	29,6%
	Étterem	30,6%	31,9%	**29,6%**	31,1%
	Utca	29,8%	30,9%	**27,8%**	28,3%
	Repülőtér	29,4%	29,4%	**27,3%**	29,5%
	Vonat	30,3%	31,0%	**27,3%**	29,1%
	Átlag	26,6%	26,2%	**24,1%**	25,4%
Átlag		19,3%	19,5%	**17,8%**	18,5%

eltérő jellemzőkkel, és eltérő architektúrájú neuronhálóval dolgozott), melyek az általunk ismert legjobb olyan publikált eredmények az Aurora-4 adatbázison, melyeket a tiszta tanító halmaz felhasználásával értek el. A kapott eredmények kiolvashatók az 1. táblázatból.

Az „üvegnyak" és monofón oszlopokat összehasonlítva azt látjuk, hogy a validációs halmazon mutatott különbség a szószintű felismerésben is megjelenik. Nem csupán az átlagolt felismerési eredményben múlja felül az „üvegnyak" a monofón módszert, de (három kivétellel) minden teszthalmazon szignifikánsan jobb annál. Érdekesebb megfigyeléseket tehetünk a jellemzőkombináció összehasonlításával az „üvegnyak", illetve monofón módszerekkel. Itt egy kettősséget fedezhetünk fel, a Sennheiser mikrofonnal készült teszthalmazok és a többi teszthalmaz között. Míg az első esetén az „üvegnyak" módszer eredménye alig múlja felül a jellemzőkombinációs módszerét (miközben a monofón módszer alatta is marad), a második esetben a monofón módszer is egy kevéssel túlteljesíti a jellemzőkombinációs változatot, míg az „üvegnyak" módszer előnye jóval jelentősebb (9,3 százalékos hibaarány-csökkenés, szemben a korábbi 3,7 százalékkal). Úgy tűnik tehát, hogy miközben a többsávos módszerrel javult a felismerés eredményessége additív zaj jelenlétében (ez akár szignifikáns különbséget is jelenthet), igazán akkor profitáltunk belőle, amikor a beszédfelismerés a tanítóhalmaztól eltérő mikrofonnal rögzített beszéden történik. Ezt láthatjuk akkor is, ha az „üvegnyak" módszer eredményeit Ganapathy eredményeivel vetjük össze: miközben a Sennheiser mikrofonnal rögzített teszthalmazok esetén nem látunk érdemben jobb felismerési eredményeket az „üvegnyak" módszernél (sőt, néhány esetben rosszabbul is teljesít), a többi teszthalmazon jelentősen jobb eredményeket kaphatunk a használatával.

6. Konklúzió és jövőbeni tervek

Kísérleteink azt mutatták, hogy korábban bemutatott módszereink nem csak sikeresen kombinálhatók, de új környezetben is alkalmazhatók. Az is világossá vált, mennyire fontos a módszerek megfelelő adaptációja (az „üvegnyak" bevezetése előtt nem tudtunk javulást elérni a felismerési eredményekben). Ismételten megtapasztaltuk továbbá, hogy mennyire körülményes lehet az eltérő sávokat feldolgozó neuronhálók eredményeinek egyesítése. Így a jövőben megpróbálunk olyan módszert találni, amely a jelenleg két lépéses módszert egyszerűsítheti.

Hivatkozások

1. Ganapathy, S.: Robust speech processing using ARMA spectrogram models. In: Proceedings of ICASSP. (2015) 5029–5033
2. Kleinschmidt, M., Gelbart, D.: Improving word accuracy with Gabor feature extraction. In: Proceedings of ICSLP. (2002) 25–28
3. Hermansky, H., Timbrewala, S., Pavel, M.: Towards ASR on partially corrupted speech. In: Proceedings of ICSLP. (1996) 464–465
4. Atal, B.S., L, H.S.: Speech analysis and synthesis by linear prediction of the speech wave. The Journal of the Acoustical Society of America **50**(2) (1971) 637–655
5. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America **118**(2) (2005) 887–906
6. Okawa, S., Bocchieri, E., Potamianos, A.: Multi-band speech recognition in noisy environments. In: Proceedings of ICASSP. (1998) 644–644
7. Kovács, Gy., Tóth, L., Grósz, T.: Robust multi-band ASR using deep neural nets and spectro-temporal features. In: Proceedings of SPECOM. (2014) 386–393
8. Grósz, T., Kovács, Gy., Tóth, L.: Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben. In: MSZNY. (2014) 3–13
9. Kovács, Gy., Tóth, L.: The joint optimization of spectro-temporal features and neural net classifiers. In: Proceedings of TSD, Springer (2013) 552–559
10. Veselý, K., Karafiát, M., Grézl, F.: Convolutive bottleneck network features for LVCSR. In: Proceedings of ASRU. (2011) 42 – 47
11. Kovács, Gy., Tóth, L.: Joint optimization of spectro-temporal features and deep neural nets for robust automatic speech recognition. Acta Cybernetica **22**(1) (2015) 117–134
12. Hirsch, H.G., Pearce, D.: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW). (2000) 29–32
13. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Veselý, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. (2011)
14. Peddinti, V., Chen, G., Povey, D., Khudanpur, S.: Reverberation robust acoustic modeling using with time delay neural networks. In: Proceedings of Interspeech. (2015) 3214 – 3218