Nevetések automatikus felismerése mély neurális hálók használatával

Gosztolya Gábor[1,2], Beke András[3], Neuberger Tilda[3]

[1] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103., e-mail: ggabor@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
[3] MTA Nyelvtudományi Intézet
Budapest, Benczúr u. 33., e-mail: {beke.andras, neuberger.tilda}@nytud.mta.hu

Kivonat A nonverbális kommunikáció fontos szerepet játszik a beszéd megértésében. A beszédstílus függvényében a nonverbális jelzések típusa és előfordulása is változik. A spontán beszédben például az egyik leggyakoribb nonverbális jelzés a nevetés, amelynek számtalan kommunikációs funkciója van. A nevetések funkcióinak elemzése mellett megindultak a kutatások a nevetések automatikus felismerésére pusztán az akusztikai jelből [1,2,3,4,5,6]. Az utóbbi években a beszédfelismerés területén, a keretszintű fonémaosztályozás feladatában uralkodóvá vált a mély neurális hálók (DNN-ek) használata, melyek háttérbe szorították a korábban domináns GMM-eket [7,8,9]. Jelen kutatásban mély neurális hálókat alkalmazunk a nevetés keretszintű felismerésére. Kísérleteinket három jellemzőkészlettel folytatjuk: a GMM-ek esetében hagyományosnak számító MFCC és PLP jellemzők mellett alkalmazzuk az FBANK jellemzőkészletet, amely 40 Mel szűrősor energiáiból, illetve azok első- és másodrendű deriváltjaiból áll. Vizsgáljuk továbbá, hogy az egyes frekvenciasávok milyen mértékben segítenek a mély neurális hálónak a nevetést tartalmazó keretek azonosításában. Ezért a dolgozat második részében kísérletileg rangsoroljuk, hogy az egyes sávok mennyire járulnak hozzá a mély neurális háló pontosságának eléréséhez.[1]

Kulcsszavak: nevetés, akusztikus modellezés, mély neurális hálók

1. Bevezetés

A nonverbális kommunikációnak nagy jelentősége van a beszédészlelés és a beszédmegértés során. Az üzenet átadásában fontos szerepet játszhatnak mind a vizuális elemek (pl. gesztusok, szemkontaktus), mind pedig a nem verbális hangjelenségek, mint amilyen a nevetés, a torokköszörülés vagy a hallható ki-, illetve

[1] Jelen kutatási eredmények megjelenését a „Telemedicina-fókuszú kutatások orvosi, matematikai és informatikai tudományterületeken” című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatja. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg. A kutatás továbbá a 108762 számú OTKA támogatásával jött létre.

belégzés. A természetes kommunikáció során a hallgatók egyidejűleg dekódolják a különböző modalitásokból érkező (vizuális és auditív) információkat, a multimodális percepció során a feldolgozási műveletek nem csupán a hallottakra, hanem a látottakra is kiterjednek.

A nevetések a spontán beszéd relatíve gyakori kísérőjelenségei, amelyek számos funkcióval rendelkezhetnek. Gyakoriságát tekintve spontán társalgásokban 1-3 előfordulást adatoltak percenként [10,11], de természetesen a nevetés előfordulását sok tényező befolyásolja, így például a beszédtéma, a kontextus, a társalgó partnerek ismeretsége, valamint a hierarchiaviszonyok. A nevetés funkcióját tekintve általában a beszélő érzelmi állapotának jelölője, az öröm fajspecifikus jelzője, a humor velejárója. Társas jelzés, társadalmilag kialakított, könnyen dekódolható jelenség. A kutatók számára azért lehet fontos a nevetések különböző szempontú vizsgálata, mert általa többet tudhatunk meg az emberi viselkedésről, a szociális interakció szerveződéséről. A nevetés a társalgásban számtalan funkciót tölthet be; gyakran kontextualizációs utasításként vagy értelmezési keretet pontosító jelzésként működik, utólag pontosíthatja, illetve eleve kijelölheti a társalgási stílust és a hozzátartozó értelmezési keretet [12,13].

A kutatások többsége elkülönít nevetéstípusokat. Günther [14] a diskurzus szerkezeti szerveződése szerint az alábbi típusokat különbözteti meg: csatlakozó/barátkozó, kontextualizáló, ellenkező/kihívó, reflexív, heterogén nevetés, valamint a be nem sorolható esetek (Günther [14]: 153-161; idézi Hámori [13]: 116). A perceptuális benyomás szerint Campbell és munkatársai [4] négy fonetikai típusra osztották a nevetéseket: zöngés nevetés, kuncogás, levegős és nazális nevetés. A produkciós oldalról vizsgálva az akusztikum alapján a fő megkülönböztető jegy, hogy a nevetés zöngével vagy a nélkül valósult meg (így megkülönböztethetők pl. zöngés énekszerű, zöngétlen horkantásszerzű vagy kevert típusok, vö. [15]). A nevetés produkcióját tekintve a társalgó partnerek részvételének szempontjából megkülönböztethetők az alábbi típusok: önálló nevetés, együttnevetés, nevetés a társalgó partner beszéde alatt (háttércsatorna-jelzésként), nevetős beszéd, kevert típus [11,16]. A korábbi akusztikai vizsgálataink alapján megállapítható volt, hogy ezek a típusok eltérő időtartamban, valamint eltérő harmonikus-zaj aránnyal (HNR) jelennek meg.

A nevetések akusztikai jellemzőit számos tanulmány elemezte a nemzetközi szakirodalomban [17,18,15,19], a magyar nyelvre vonatkozóan azonban alig akad ilyen jellegű munka [11,16,20]. A vizsgálatok szerint a nevetések akusztikai jellemzői (F0, formánsok, amplitúdó, zöngeminőség) a beszédhez hasonlatosak, azok hehezetes CV /hV/ szótagok sorozataként realizálódnak, bár a beszédhez képest hosszabb zöngétlen résszel valósulnak meg. A szöveges részektől való elkülönítésükben (a nevetések detektálásában) nagy szerepet játszik a zöngétlen-zöngés rész aránya. Két, amerikai angol beszélő (egy nő és egy férfi) nevetéseit vizsgálva kimutatták, hogy egy szótagnyi nevetés átlagos időtartama 204, illetve 224 ms, és a nevetések átlagosan 6,7, illetve 1,2 szótagból állnak [17]. A további eredmények szerint a nevetés produkciójában másodpercenként átlagosan 4,7 szótag adatolható, ami nagy hasonlóságot mutat az (angol, francia és svéd) olvasott mondatok másodpercenkénti szótagszámával. Német és olasz anyanyelvű beszélők esetében

azt találták, hogy a nevetések átlagos időtartama 798 ms nőknél és 601 ms férfiaknál, valamint hogy az alaphangmagasság átlagosan 472 Hz nőknél és 424 Hz férfiaknál, tehát a beszéd és a nevetések megkülönbözteté-sében az alapfrekvencia értékének is jelentős szerepe van [18]. Bachorowski és munkatársai [15] összehasonlítottak szakirodalmi adatokat a nevetések átlagos alapfrekvencia-értékeire vonatkozólag, amelyek nők esetében 160 és 502 Hz, férfiak esetében 126 és 424 Hz közötti értékkel jelentek meg a különböző kutatásokban. A magyar nevetéseket a BEA adatbázis [21] hanganyagaiban vizsgálták, és azt találták, hogy az átlagos időtartamuk 911 ms (átlagos eltérés: 605 ms), átlagos F0-értékük 207 Hz (átl. elt.: 49 Hz) férfiaknál, 247 Hz (átl. elt.: 40 Hz) nőknél. A nevetések számos akusztikai paraméterben (jitter, shimmer, jel-zaj viszony, F0-átlag) szignifikánsan különbséget mutattak a beszédszegmentumokhoz (jelen esetben szavakhoz) képest, ami megkönnyítheti az elkülönítésüket a szöveges részektől.

2. Nevetések felismerése

A nevetések automatikus osztályozása megközelítőleg egy évtizedes múltra tekint vissza. A nevetések leggyakrabban önállóan fordulnak elő, egy részük azonban nem önállóan, hanem a beszéddel egyidejűleg jelenik meg a társalgásokban [22,23], de gyakoriak az együttnevetések is, vagyis amikor két beszélő szimultán nevetése hangzik el. Kennedy és Ellis [2] tanulmányukban az átfedő nevetéseket is elemezték, a detektálásukhoz SVM-et használtak osztályozó algoritmusként (az MFCC-t, a spektrális ingadozást és a két asztali mikrofon jele közötti időbeni eltéréseket vizsgálva). A rendszerükkel 87%-os helyes osztályozási eredményt értek el. A Chist Era JOKER projekt keretein belül Tahon és Devillers a pozitív és a negatív hangulatú nevetések automatikus detektálását tűzték ki célul [24]. A nevetéseket az alaphangmagassággal, a formánsaival, intenzitásukkal és spektrális jellemzőikkel reprezentálták. A kutatásukban a WEKA szoftver SMO osztályozót használták RBF kernelfüggvénnyel. A tanító korpuszban a pozitív nevetések száma 140 db volt, míg a negatív nevetéseké 117 db. A tesztadatbázis 48 pozitív nevetést tartalmazott, és 27 negatívat. Az eredmények szerint a pozitív nevetések F-értéke $64,5\%$, míg a negatívaké $28,5\%$ volt.

Számos kutatás tűzte ki céljául a nevetések automatikus felismerését. Truong és van Leeuwen [1,3] Gauss-keverék modellel és perceptuális lineáris predikciós (PLP) akusztikai előfeldolgozási eljárással mutattak ki $87,6\%$-os helyes osztályozási eredményt. A PLP együtthatóinak száma befolyással lehet a nevetések osztályozási eredményére. Petridis és Pantic [25] kutatásukban kimutatták, hogy ha 13 együtthatót tartalmazó PLP-t használnak feed-forwald neurális hálózattal kombinálva, akkor az F-érték 64%, míg egy másik munkájukban 7 együtthatót tartalmazó PLP-vel 68% volt [26]. Reuderink és munkatársai [27] a PLP-RASTA jellemzőt használták a nevetések detektálására, illetve GMM és HMM osztályozót. Az eredmények azt mutatták, hogy a GMM osztályozó kicsivel jobb eredményt adott (átlagos AUC-ROC 0,825), mint a HMM (átlagos AUC-ROC 0,822). A PLP jellemző mellett igen népszerű az MFCC akusztikai reprezentáció. Ahogy a PLP esetén, úgy az MFCC együtthatóinak számának megválasztása sem

egyértelmű. Kennedy és Ellis [2] csak az első 6 MFCC együtthatót használták, és hasonló eredményt érték el, mint a 13 együtthatót tartalmazó MFCC-vel. Mindez arra utalhat, hogy a nevetések megkülönböztető akusztikai jellemzőit az alsóbb frekvenciasávokban érdemes keresni. A spektrális akusztikai jellemzők mellett vizsgálták a prozódia szerepét a nevetések detektálásában. Az eredmények azt mutatták, hogy a prozódia jól reprezentálja a nevetések dinamikus jellemzőit [1,28].

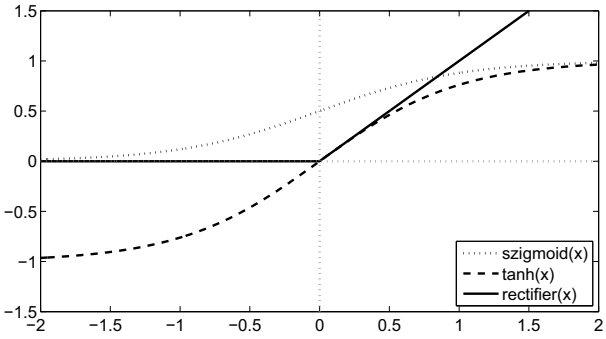
Számos kutatás nem egy-egy akusztikai jellemzőt alkalmazott, hanem több jellemző kombinációját. Knox és Mirghafori [6] neurális hálózatok alkalmazásával (MFCC és alaphangmagasság kombinált paraméterekkel) 90% fölötti teljesítményt értek el a nevetések detektálásában. Hasonlóan, a magyar spontán beszédből származó nevetések esetében is 90% fölötti eredményt mutatott a nevetések és a beszédszegmensek osztályozása GMM-SVM kevert módszerrel MFCC, PLP és akusztikai paramétereken tesztelve [16]. Egy másik vizsgálatunkban [29] különböző osztályozási technikákat és ezek kombinációit (GMM-ANN, GMM-SVM) alkalmaztuk a nevetések és a szavak osztályozásához. A legjobb eredményt (EER: $2,5\%$) az MFCC és az akusztikai paraméterek használatával a GMM és SVM kombinált osztályozóval értük el.

Az akusztikai jellemzők mellet fontos kérdés, hogy milyen osztályozó eljárást érdemes alkalmazni a nevetések detektálásában. A kutatások többsége vagy SVM-et [27], vagy GMM-et [1,27,30], vagy ANN-t [6,25,26,28] alkalmazott. Az utóbbi években a beszédfelismerés területén, a keretszintű fonémaosztályozás feladatában uralkodóvá vált a mély neurális hálók (DNN-ek) használata, melyek háttérbe szorították a korábban domináns GMM-eket. Korábbi kísérleteink során mi is kiemelkedő eredményeket értünk el a használatukkal mind magyar [7,8], mind angol [9] nyelven. Jelen dolgozatban mély neurális hálókat alkalmazunk a nevetés felismerésére. A korábbi megközelítéseinkkel ellentétben, ahol az egész szegmensre illesztettünk GMM-et, majd a modell paraméterei alapján végeztük el a döntést, most kizárólag keretszintű felismerést végzünk; a szegmensszintű döntést a keretszintű valószínűségek szorzata alapján hozzuk meg. Korábbi módszerünk kézenfekvőnek tűnő adaptálása már csak azért sem járható út, mert a GMM-mel szemben egy DNN nehezen interpretálható, és nagyságrendekkel több paraméterrel (súllyal) is rendelkezik.

3. Mély neurális hálók

A beszédfelismerés területén, a lokális valószínűség-eloszlások modellezésére hagyományosan GMM-eket volt szokás használni, azonban ezeket a mély neurális hálók az utóbbi pár évben szinte teljesen kiszorították. Ennek oka, hogy a mély neurális hálók jóval nagyobb pontosságot képesek elérni ebben a feladatban (l. pl. [9,31]), miközben elérhetővé váltak azok a hardverek, melyekkel a mély neurális hálók viszonylag gyorsan betaníthatók és kiértékelhetők.

A hagyományos hálózatok esetében egy vagy maximum két rejtett réteget szoktunk csak használni, és a neuronok számának növelésével próbáljuk javítani az osztályozási pontosságot. Az utóbbi idők kísérleti eredményei azonban amel-



1. ábra. A szigmoid, tanh és rectifier aktivációs függvények

lett szólnak, hogy – adott neuronszám mellett – több réteg hatékonyabb reprezentációt tesz lehetővé [32]. Az ilyen sok rejtett réteges, „mély" architektúrának azonban nem triviális a betanítása. A hagyományos neuronhálók tanítására általában az ún. backpropagation algoritmust szokás használni, ez azonban kettőnél több rejtett réteg esetében egyre kevésbé hatékony. Ennek egyik oka, hogy egyre mélyebbre hatolva a gradiensek egyre kisebbek, egyre inkább eltűnnek („vanishing gradient"), ezért az alsóbb rétegek nem fognak kellőképpen tanulni [32].

A mély neurális hálók tanítására először Hinton et al. javasolt egy módszert [33]. Ebben az eljárásban a tanítás két lépésben történik: egy felügyelet nélküli előtanítást egy felügyelt finomhangolási lépés követ. A felügyelt tanításhoz használhatjuk a backpropagation algoritmust, az előtanításhoz azonban egy új módszer, a DBN előtanítás szükséges. Ez az inicializálási lépés, habár megnöveli a végül betanított neurális háló pontosságát, elég körülményessé és időigényessé teszi a tanítási folyamatot.

Ennek egy alternatívájaként javasolta Seide et al. a diszkriminatív előtanítást [31]. Ebben az első lépésben hagyományos módon egy egyetlen rejtett réteget tartalmazó neurális hálót tanítanak be. Ezután minden lépésben eldobják a kimeneti réteget a hozzá tartozó súlyokkal együtt, majd a hálót kiegészítik egy újabb rejtett réteggel és egy kimeneti réteggel. Az új kapcsolatokat véletlen súlyokkal inicializálják, és ezt a hálót tanítják a hagyományos módon (általában backpropagation eljárással). Ezeket a lépéseket ismétlik mindaddig, míg a tervezett számú rejtett réteget el nem érik.

A harmadik megoldás, mellyel a mély neurális hálók taníthatóvá válnak, nem egy új tanítási módszer, hanem a rejtett rétegek neuronjainak aktivációs függvényének lecserélése. A hagyományos szigmoid (vagy ennek skálázott változata, a tanh) függvény helyett az ún. *lineáris rectifier* függvényt alkalmazva (l. 1. ábra) az összes rejtett réteg taníthatóvá válik szimplán backpropagation módszerrel [34], ugyanakkor szükség van a súlyok regularizációjára (pl. L1 vagy L2 norma használatával).

Korábbi kísérleteink során mindhárom megközelítést teszteltük mind angol, mind magyar nyelvű szöveg fonémafelismerése során [7,35]. Tapasztalataink szerint a legjobb eredményt a rectifier aktivációs függvény használata hozta, mi-

közben ez bizonyult a leggyorsabbnak is, ezért kísérleteinkben ezt a fajta hálót fogjuk alkalmazni.

4. Kísérletek és eredmények

A jelen kutatásban a nevetést tartalmazó és azt nem tartalmazó keretek elkülönítésére használtunk mély neurális hálót; a kimeneti réteg két neuronja felelt meg ennek a két osztálynak. Kísérleteinket magyar spontán beszédből vett nevetés- és beszédrészleteken végeztük, melyek a BEszélt nyelvi Adatbázisból (BEA, [21,36]) lettek kiválogatva. A BEA különböző típusú spontán beszédet tartalmaz: narratíva, véleménykifejtés, három fős társalgás. A jelen kutatáshoz 75 adatközlő felvételét választottuk ki, átlagosan 16 percet egy beszélőtől. A hanganyagokban megjelenő összes nevetést manuálisan címkéztük fel a Praat programban. Ugyanezektől a személyektől az elhangzott nevetések számával nagyságrendileg megegyező mennyiségű beszédszegmenst választottunk ki. Összesen 331 nevetés- és 320 beszédszegmenst használtunk, melyekből 463-ra tanítottunk, és 188 szegmens került a teszthalmazba. A tanító adatbázis 240 nevetést és 223 beszédszegmenst tartalmazott, míg a tesztadatbázis 91 nevetést és 97 beszédszegmenst. Mivel az adatmennyiség nem túl nagy, külön fejlesztési halmaz definiálása helyett tízszeres keresztvalidációt (cross-validation, CV) alkalmaztunk.

Saját neurálisháló-implementációnkat használtuk, mellyel korábban sok különböző területen értünk el jó eredményeket (pl. [9,37,38,39]). A neurális hálókat keretszinten tanítottuk. A GMM-ek esetében bevettnek számító MFCC és PLP jellemzők mellett kipróbáltuk az FBANK jellemzőkészletet, amely 40 Mel szűrősor energiáiból, illetve azok első- és másodrendű deriváltjaiból áll [40]. Alkalmaztuk azt a fonémaosztályozás esetén bevett megoldást is, hogy a szomszédos keretek jellemzővektorait is felhasználtuk az egyes keretek osztályozása során. Az alkalmazott neurális hálók előzetes tesztek eredményei alapján három rejtett réteggel rendelkeztek, melyek mindegyikében 256 rectifier függvényt alkalmazó neuron volt, míg a kimeneti rétegben softmax függvényt használtunk. A súlyokat L2 regularizációval tartottuk kordában. A keretszintű valószínűségbecsléseket szegmensenként és osztályonként (nevetés, ill. beszéd) összeszoroztuk, és a szegmenst abba az osztályba soroltuk, amelyre ez az érték magasabb volt.

Mivel a neurális háló tanítása sztochasztikus folyamat (köszönhetően a súlyok véletlen inicializálásának), a tízszeres keresztvalidáció minden esetére öt-öt hálót tanítottunk. Ezekből öt hálót értékeltünk ki a teszthalmazra, majd az egyes kiértékelésekre kapott eredményeket összegeztük.

4.1. Eredmények

Az egyes jellemzőkészletekkel, valamint a szomszédos keretek számával elért keretszintű pontosságértékeket a 1. táblázat, míg a szegmensszintűeket a 2. táblázat tartalmazza. Egy-egy jellemzőkészleten belül a legjobb értékeket (0, 2% tűréssel) kiemeltük.

1. táblázat. A különböző jellemzőkészletek és szomszédszámok használatával elért keretszintű pontosságértékek

Jellemző-készlet	N	Keresztvalidáció				Teszthalmaz			
		Pr.	Re.	F_1	Acc.	Pr.	Re.	F_1	Acc.
MFCC	1	72,3%	79,9%	75,9%	84,3%	49,8%	**89,3%**	63,9%	72,1%
	5	76,4%	81,6%	79,1%	86,6%	54,2%	**89,1%**	67,4%	76,1%
	9	78,8%	**82,3%**	**80,5%**	**87,7%**	64,1%	85,8%	**73,4%**	**82,8%**
	13	**79,3%**	**82,1%**	**80,7%**	**87,8%**	63,7%	84,1%	72,5%	82,3%
	17	**79,1%**	81,5%	80,3%	**87,6%**	**64,9%**	82,3%	72,6%	**82,8%**
PLP	1	76,2%	78,8%	77,5%	85,8%	52,4%	84,1%	64,6%	74,5%
	5	80,7%	81,3%	81,0%	**88,2%**	63,1%	**87,1%**	73,2%	82,3%
	9	80,6%	**81,9%**	**81,3%**	**88,3%**	**65,1%**	86,7%	**74,4%**	**83,4%**
	13	**81,5%**	79,8%	80,6%	**88,1%**	**65,3%**	85,4%	74,1%	**83,4%**
	17	**81,3%**	78,1%	79,7%	87,7%	62,8%	85,3%	72,3%	81,9%
FBANK	1	**99,1%**	**99,8%**	**99,5%**	**99,7%**	**97,6%**	**99,3%**	**98,4%**	**99,1%**
	5	98,7%	**99,6%**	99,2%	**99,5%**	95,4%	**99,3%**	97,3%	98,5%
	9	98,0%	99,5%	98,7%	99,2%	95,2%	**99,2%**	97,1%	98,4%
	13	97,7%	99,2%	98,4%	99,0%	94,5%	98,9%	96,6%	98,1%
	17	97,4%	99,0%	98,2%	98,9%	93,0%	98,8%	95,8%	97,6%

Látható, hogy míg MFCC jellemzőkészlet esetén 9–17 szomszédos kereten együtt tanítva kapjuk a legjobb keretszintű értékeket, PLP esetén ez 9–13 keret, ennél nagyobb szomszédságot használva már romlanak az eredmények, FBANK esetén pedig szomszédok nélkül tanítva kapjuk meg az optimumot. Az MFCC és a PLP használatával kapott értékek esetén megfigyelhető egy eltolódás a pontosság (precision) és a fedés között a keresztvalidációval, illetve a teszthalmazon kapott értékek között; ennek oka valószínűleg az, hogy a két halmazon belül a két osztály példáinak aránya nem ugyanaz.

A szegmensszintű pontosságok lényegesen magasabbak a keretszintűeknél, ami érthető, hiszen egy szegmens pontos besorolásához elég, ha a benne található keretek többségét jól osztályozzuk. MFCC esetén ezúttal is 9–13 szomszédos kereten tanítani tűnik az optimális választásnak, míg PLP használatával ez 5–9. Ezekben az esetekben is megfigyelhető az eltolódás a fedés és a pontosság (precision) között a keresztvalidációs és a teszthalmazon mért értékek között. Az FBANK jellemzőkészletnél 1, illetve 5 szomszédos keretre is tökéletes osztályozást kapunk, több szomszédot használva ez keresztvalidáció során egy picit romlik, de mindig 99% fölött marad.

Az FBANK jellemzőkészlet kiugró eredménye valószínűleg annak köszönhető, hogy az eredeti beszédhanghoz közelebb álló jellemzőket tartalmaz. Valószínűleg a nevetés felismeréséhez néhány kitüntetett frekvenciasáv vizsgálata fontos, és ezek sokkal jobban detektálhatóak ebben a szűrősorokból álló jellemzőkészletben, mint akár az MFCC-ben, akár a PLP-ben. Ehhez azonban neurális háló használatára van szükségünk, mivel GMM-et csak dekorrelált jellemzőkészletre lehet tanítani (a tesztelt jellemzők közül ilyen a MFCC és a PLP).

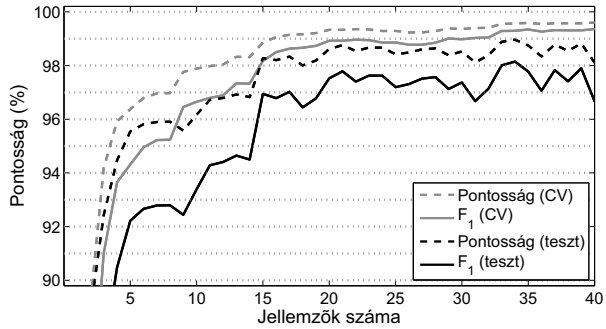
2. táblázat. A különböző jellemzőkészletek és szomszédszámok használatával elért szegmensszintű pontosságértékek

Jellemző-készlet	N	Keresztvalidáció				Teszthalmaz			
		Pr.	Re.	F_1	Acc.	Pr.	Re.	F_1	Acc.
MFCC	1	**98,8%**	85,2%	91,5%	92,4%	86,1%	94,9%	90,3%	89,5%
	5	**98,6%**	88,3%	93,1%	93,7%	90,4%	**96,7%**	93,4%	93,0%
	9	98,0%	89,1%	**93,3%**	**93,9%**	**97,5%**	95,5%	**96,5%**	**96,4%**
	13	97,4%	**90,0%**	**93,5%**	**94,0%**	96,2%	93,4%	94,8%	94,7%
	17	96,6%	88,1%	92,1%	92,7%	96,7%	89,3%	92,8%	92,9%
PLP	1	98,2%	85,1%	91,2%	92,1%	86,4%	94,2%	90,1%	89,4%
	5	**99,1%**	**88,3%**	**93,4%**	**94,0%**	93,8%	96,3%	95,0%	94,8%
	9	98,7%	**88,3%**	**93,2%**	**93,8%**	**94,4%**	**97,3%**	**95,8%**	**95,6%**
	13	98,7%	85,0%	91,3%	92,2%	**94,5%**	95,9%	95,2%	95,0%
	17	98,2%	81,9%	89,3%	90,5%	93,0%	93,2%	93,1%	92,9%
FBANK	1	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**
	5	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**
	9	**100,0%**	99,7%	**99,9%**	**99,9%**	**100,0%**	**100,0%**	**100,0%**	**100,0%**
	13	99,6%	99,6%	99,6%	99,6%	**100,0%**	**100,0%**	**100,0%**	**100,0%**
	17	99,4%	99,3%	99,3%	99,4%	**100,0%**	**100,0%**	**100,0%**	**100,0%**

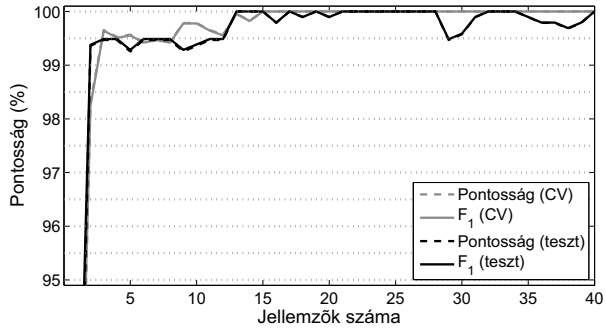
4.2. A felhasznált jellemzők vizsgálata

Tesztjeink során az FBANK jellemzőkészlettel meglepően jó eredményeket kaptunk. Ez a jellemzőkészlet az eredeti hanghoz közelebb álló, jobban interpretálható, mint akár az MFCC, akár a PLP. Az is nyilvánvaló, hogy az egyes frekvenciasávok nem ugyanolyan mértékben segítenek a mély neurális hálónak a nevetést tartalmazó keretek azonosításában. Ezért a következő részben kísérletileg rangsoroljuk, hogy az egyes sávok mennyire járulnak hozzá a mély neurális háló pontosságának eléréséhez.

A jellemzőkészletben 123 attribútum van; nyilvánvaló, hogy az összes lehetséges részhalmaz letesztelése aránytalanul sok időt emésztene fel. Ezért első lépésben sorba rendeztük a jellemzőket; a sok lehetséges mód közül mi ezt is a (már betanított) neurális hálók segítségével tettük meg. Egy neurális háló bemeneti rétegének neuronjai a jellemzőknek felelnek meg (amennyiben nem használjuk a szomszédos keretek jellemzőit), így az egyes neuronokból kimenő súlyok összessége (valamilyen mértékben) tükrözi az egyes jellemzők fontosságát. Mivel a súlyok valós számok, érdemes az egyes jellemzőkhöz tartozó súlyok négyzetösszegét venni: minél nagyobb ez az érték, annál fontosabbnak ítéli az adott neurális háló a jellemzőt. Ezután a jellemzőket ezen érték alapján csökkenő sorrendbe rendeztük, és mindig az első N db attribútum használatával tanítottunk neurális hálókat. (A kísérleti körülmények megegyeztek a 4. fejezetben bemutatottakkal.) Az elért pontosságértékek a 2. és 3. ábrán láthatóak. A keretszintű pontosságértékek már 15 jellemző használatával eljutnak $97-99\%$ közé, míg a tökéletes szegmensszintű osztályozáshoz 13 jellemző is elegendő. Ez igazolta azt a hipotézisünket, hogy csak néhány frekvenciasáv vizsgálata alapján is nagy pontossággal eldönt-



2. ábra. A mély neurális hálókkal elért keretszintű pontosságértékek a használt jellemzők számának függvényében

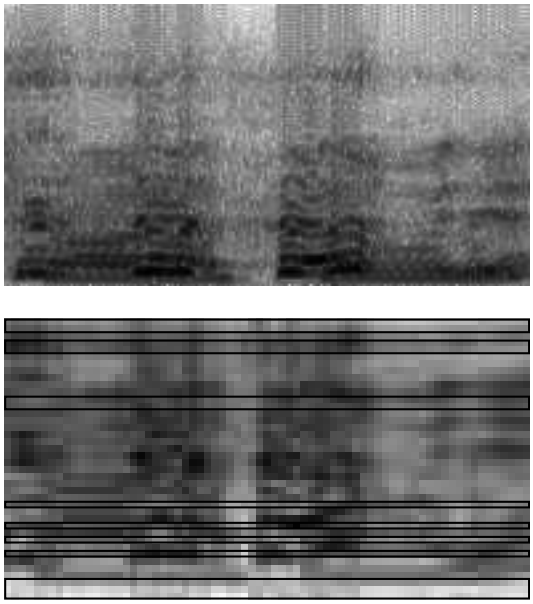


3. ábra. A mély neurális hálókkal elért szegmensszintű pontosságértékek a használt jellemzők számának függvényében.

hető, hogy az adott keret vagy szegmens nevetést tartalmaz-e. A kiválasztott frekvenciasáv-jellemzők a 4. ábrán láthatók. Az egyszerűség kedvéért nem ábrázoltuk az energiát, amely meglepő módon nem szerepelt a kiválasztott jellemzők között; illetve eltekintettünk az első- és másodrendű deriváltak megjelenítésétől is. A tizenöt jellemző közül egyébként tíz volt frekvenciasáv-szűrő, kettő első-, három pedig másodrendű derivált.

5. Összegzés

A jelen kutatásban mély neurális hálózatokat használtunk a nevetés keretszintű automatikus felismeréséhez. A kutatás során vizsgáltuk, hogy mely akusztikai jellemzők alkalmasabbak a nevetés azonosítására. Az akusztikai jellemzők közül a nevetésfelismerés szakirodalmában használt MFCC-t és PLP-t vettük górcső alá, illetve a mély neurális hálózatokhoz kiválóan alkalmas FBANK jellemzőt. Vizsgáltuk továbbá azt, hogy a keretszintű felismeréskor hány szomszédos kereten érdemes tanítani. A jellemzőkészlet mellett vizsgáltuk, hogy mely frekvenciasávok vesznek részt a mély neurális hálózatokkal történő nevetésfelismerésben.



4. ábra. Egy nevetésszegmens spektrogramja (fent), valamint az ebből kinyert BANK jellemzőkészlet és azon a 15 kiválasztott jellemzőhöz tartozó frekvenciatartományok (lent)

Ennek elemzéséhez a mély neurális hálózat kimeneti súlyait használtuk fel, illetve annak alapján rangsoroltuk az egyes frekvenciasávokat. Az eredmények azt mutatták, hogy a legjobb eredményt akkor kaptuk, ha az akusztikai jellemzők közül az FBANK reprezentációt használtuk a mély neurális hálózatok tanításához. A szomszédos keretek számának tekintetében az MFCC és a PLP használatakor érdemes nagyszámú szomszédos kereten tanítani, ugyanakkor az FBANK esetében szinte nincs is szükség szomszédos keretekre. A jellemzőválogatás során azt találtuk, hogy már 15 jellemző felhasználásakor is igen magas pontossági mutatót kapunk. Mindez azt bizonyítja, hogy az egyes frekvenciasávok nem egyenlő módon vesznek részt a nevetés azonosításában.

Összességében elmondható, hogy a mély neurális hálózatok jól alkalmazhatók a nevetések automatikus osztályozásában a korábbi kísérletekben használt GMM-SVM-mel összehasonlítva is. Az FBANK jellemzőkinyerés esetén a korábbi, ugyanezen korpuszra vonatkozó eredményeket felül is teljesíti. A kutatásunk egy újabb bizonyíték arra, hogy milyen kiválóan alkalmazható a mély neurális hálózat szinte bármely beszédtechnológiai kihívásban.

Hivatkozások

1. Truong, K.P., van Leeuwen, D.A.: Automatic detection of laughter. In: Interspeech, Lisszabon, Portugália (2005) 485–488

2. Kennedy, L.S., Ellis, D.P.W.: Laughter detection in meetings. In: Proceedings of the NIST Meeting Recognition Workshop at ICASSP, Montreal, Kanada (2004) 118–121
3. Truong, K.P., van Leeuwen, D.A.: Automatic discrimination between laughter and speech. Speech Communication **49**(2) (2007) 144–158
4. Campbell, N., Kashioka, H., Ohara, R.: No laughing matter. In: Interspeech, Lisszabon, Portugália (2005) 465–468
5. Nick, C.: On the use of nonverbal speech sounds in human communication. In Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M., eds.: Verbal and nonverbal communication behaviours. Springer-Verlag, Berlin, Heidelberg (2004) 117–128
6. Knox, M.T., Mirghafori, N.: Automatic laughter detection using neural networks. In: Interspeech. (2007) 2973–2976
7. Grósz, T., Kovács, Gy., Tóth, L.: Új eredmények a mély neuronhálós magyar nyelvű beszédfelismerésben. In: MSZNY, Szeged, Magyarország (2014) 3–13
8. Grósz, T., Gosztolya, G., Tóth, L.: Környezetfüggő akusztikai modellek létrehozása Kullback-Leibler–divergencia alapú klaszterezéssel. In: MSZNY, Szeged, Magyarország (2015) 174–181
9. Tóth, L.: Phone recognition with hierarchical Convolutional Deep Maxout Networks. EURASIP Journal on Audio, Speech, and Music Processing **2015**(25) (2015) 707–710
10. Holmes, J., Marra, M.: Having a laugh at work: How humour contributes to workplace culture. Journal of Pragmatics **34**(12) (2002) 1683–1710
11. Neuberger, T.: Nonverbális hangjelenségek a spontán beszédben. In Gósy, M., ed.: Beszéd, adatbázis, kutatások. Akadémiai Kiadó, Budapest (2012) 215–235
12. Glenn, P.: Laughter in interaction. Cambridge University Press, Cambridge, UK (2003)
13. Hámori, A.: Nevetés a társalgásban. In Laczkó, K., Tátrai, S., eds.: Elmélet és módszer. ELTE Eötvös József Collegium, Budapest (2014) 105–129
14. Günther, U.: What's in a laugh? Humour, jokes, and laughter in the conversational corpus of the BNC. PhD thesis, Universität Freiburg (2002)
15. Bachorowski, J.A., Smoski, M.J., Owren, M.J.: The acoustic features of human laughter. Journal of the Acoustical Society of America **110**(3) (2001) 1581–1597
16. Neuberger, T., Beke, A.: Automatic laughter detection in spontaneous speech using GMM-SVM method. In: TSD, Pilsen, Csehország (2013) 113–120
17. Bickley, C., Hunnicutt, S.: Acoustic analysis of laughter. In: ICSLP, Banff, Kanada (1992) 927–930
18. Rothgänger, H., Hauser, G., Cappellini, A.C., Guidotti, A.: Analysis of laughter and speech sounds in Italian and German students. Naturwissenschaften **85**(8) (1998) 394–402
19. Trouvain, J.: Segmenting phonetic units in laughter. In: ICPhS, Barcelona, Spanyolország (2003) 2793–2796
20. Bóna, J.: Nonverbális hangjelenségek fiatalok és idősek spontán beszédében. Beszédkutatás **23**(8) (2015) 106–119
21. Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T.E., Beke, A., Neuberger, T., Nikléczy, P.: BEA: Beszélt nyelvi adatbázis. In Gósy, M., ed.: Beszéd, adatbázis, kutatások. Akadémiai Kiadó, Budapest (2012) 9–24
22. Provine, R.R.: Laughter. American Scientist **84**(1) (1993) 38–45
23. Nwokah, E.E., Davies, P., Islam, A., Hsu, H.C., Fogel, A.: Vocal affect in three-year-olds: a quantitative acoustic analysis of child laughter. Journal of the Acoustical Society of America **94**(6) (1993) 3076–3090

24. Tahon, M., Devillers, L.: Laughter detection for on-line human-robot interaction. Cough **85**(65) (2015) 1–77
25. Petridis, S., Pantic, M.: Audiovisual discrimination between laughter and speech. In: ICASSP. (2008) 5117–5120
26. Petridis, S., Pantic, M.: Fusion of audio and visual cues for laughter detection. In: CIVR. (2008) 329–337
27. Reuderink, B., Poel, M., Truong, K., Poppe, R., Pantic, M.: Decision-level fusion for audio-visual laughter detection. In: MLMI, Utrecht, Hollandia (2008) 137–148
28. Petridis, S., Pantic, M.: Is this joke really funny? Judging the mirth by audiovisual laughter analysis. In: ICME. (2009) 1444–1447
29. Neuberger, T., Beke, A., Gósy, M.: Acoustic analysis and automatic detection of laughter in Hungarian spontaneous speech. In: ISSP, Köln, Németország (2014) 285–287
30. Ito, A., Wang, X., Suzuki, M., Makino, S.: Smile and laughter recognition using speech processing and face recognition from conversation video. In: CW. (2005) 437–444
31. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: ASRU. (2011) 24–29
32. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. (2010) 249–256
33. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation **18**(7) (2006) 1527–1554
34. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier networks. In: AISTATS. (2011) 315–323
35. Grósz, T., Tóth, L.: A comparison of Deep Neural Network training methods for Large Vocabulary Speech Recognition. In: TSD, Pilsen, Csehország (2013) 36–43
36. Neuberger, T., Gyarmathy, D., Gráczi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative hungarian language. In: TSD2014. (2014) 424–431
37. Gosztolya, G.: On evaluation metrics for social signal detection. In: Interspeech, Drezda, Németország (2015) 2504–2508
38. Grósz, T., Busa-Fekete, R., Gosztolya, G., Tóth, L.: Assessing the degree of nativeness and Parkinson's condition using Gaussian Processes and Deep Rectifier Neural Networks. In: Interspeech, Drezda, Németország (2015) 1339–1343
39. 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
40. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book. Cambridge University Engineering Department, Cambridge, UK (2006)