Neurális hálók tanítása valószínűségi mintavételezéssel nevetések felismerésére

Gosztolya Gábor[1,2], Grósz Tamás[2], Tóth László[1],
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 Intézet
Szeged, Árpád tér 1.
[3] MTA Nyelvtudományi Intézet
Budapest, Benczúr u. 33., e-mail: beke.andras@nytud.mta.hu

Kivonat Mikor a feladat spontán beszédben nevetések előfordulásait megtalálni, kézenfekvő megközelítés a beszédfelismerés feladatkörében gyakran használt technikákat alkalmazni. Például becsülhetjük a nevetés valószínűségét lokálisan, a keretek szintjén, mely valószínűségbecsléseket szolgáltathatja például egy mély neurális háló. Ugyanakkor a hangfelvételeknek csak kis része (néhány százaléka) felel meg nevetésnek; a többit beszéd, csend, háttérzajok, stb. teszik ki. Ez azt eredményezi, hogy a mély neurális hálót olyan adatokon tanítjuk, melyeknél az osztályelőfordulás szélsőségesen kiegyensúlyozatlan. Jelen cikkünkben a valószínűségi mintavételezés (*probabilistic sampling*) nevű eljárást alkalmaztuk a mély neurális hálók tanítása során, mellyel 7%-os relatív hibacsökkentést tudtunk elérni a keretszintű F_1 pontosságértékeket tekintve.

Kulcsszavak: nevetésdetektálás, mély neurális hálók, tanítópélda-mintavételezés

1. Bevezetés

Az emberiséget mindig is érdekelte viselkedésének alapvető megértése, előrejelezhetőségének lehetősége. Az elmúlt évtizedekben köszönhetően a technikai fejlődésnek (főként az agyi képalkotó eljárásoknak, a hang- és videórögzítésnek, valamint ezek gyors feldolgozhatóságának) egyre mélyebb ismereteink vannak az emberi viselkedésről. A beszédtudomány fókuszában főként annak vizsgálata áll, hogy hogyan viselkedünk a társas kommunikáció során. Ezen viselkedés feltérképezésnek az egyik kulcseleme a non-verbális kommunikáció vizsgálata a társalgás során. Egyes feltételezések szerint a non-verbális kommunikáció közel kétharmadát teszi ki a teljes kommunikációnak [1], és használata kevésbé kontrollált, így vizsgálatával alapvető viselkedési mintázatokat lehet kimutatni. A

Grósz Tamást az Emberi Erőforrások Minisztériuma ÚNKP-16-3 kódszámú Új Nemzeti Kiválóság Programja támogatta.

non-verbális kommunikáció során folyamatos nem-lexikális elemek küldése és fogadása történik az egyes emberek között. Modalitásukat tekintve ezek különfélék lehetnek, mint a testtartás, a szemmozgás vagy a non-verbális vokális elemek. Elsőleges szerepük a magatartás és az érzelmek kifejezésében van [2]. Mindemellett fontos szerepet töltenek be a dialógusok szerveződésében [3], illetve sok szempontból tükrözik személyiségünket [4].

A non-verbális jelek további két csoportra oszthatók: vizuális és vokális [5,6]. A vokális non-verbális jelek közé tartoznak a paralingvisztikai jelek (pl. zöngeminőség, hangerő), illetve a non-verbális vokalizációk (pl. nevetés, sóhajtás, kitöltött szünetek) [7,8]. Jelen munka a nevetések automatikus felismerésére koncentrál, mivel maga a nevetés, mint non-verbális vokalizációs elem, az egyik kulcseleme a társalgás során mutatott viselkedés feltérképezésének, illetve modellezésének.

Korábbi munkáinkban [9,10] beszédszegmensek osztályozásával (nevetés vagy szöveg/csend) foglalkoztunk, és az irodalomban is számos ilyen munkával találkozhatunk (pl. [11,12]). Egy másik elterjedt megközelítésben mind a modelltanítás, mind a kiértékelés kizárólag a keretek szintjén történik (pl. [13,14,15]). A valós alkalmazásokhoz azonban közelebb áll az a megközelítés, melyben spontán beszédben akarjuk meghatározni azokat a szegmenseket, melyek nevetést tartalmaznak. Kézenfekvő, ha ekkor a beszédfelismerés területéről veszünk át eszközöket, például a keretszintű valószínűségbecsléseket egy rejtett Markov modell (Hidden Markov model, HMM) segítségével kombináljuk. Magukat a valószínűségbecsléseket előállíthatjuk Gauss keverékmodellekkel (Gaussian Mixture Models, GMM), de neurális hálókkal (Artificial Neural Networks, ANN) vagy mély neurális hálókkal (Deep Neural Networks, DNN) is.

Akusztikus modellünket tehát keretszintű jellemzővektorokon tanítjuk, melyek a két kézenfekvő osztály (nevetés illetve nem-nevetés, beleértve a beszédet, csendet, torokköszörülést, különböző háttérzajokat stb.) valamelyikébe tartoznak. Egy lényeges különbség azonban a beszédfelismerés feladatához képest, hogy a két osztályhoz tartozó példák száma nagyon kiegyensúlyozatlan: tipikusan a keretek 4-6%-a tartalmaz nevetést [7,8,16]. Ez egy diszkriminatív osztályozó (például egy mély neurális háló) tanítása során azt jelenti, hogy az az egyik osztályból lényegesen több példát lát, így azt jobban képes megtanulni, míg a másik osztály súlyosan alulreprezentált. Ez kezelhető a gyakoribb osztályokba tartozó példák egy részének elhagyásával, azonban ez nyilvánvalóan csökkenti az adott osztály variabilitását. A másik megközelítés, hogy (amennyiben nem tudunk további, a ritkább osztályokba tartozó példákat szerezni vagy generálni) egyes tanítópéldákat gyakrabban használunk a tanítás során.

Jelen cikkünkben egy, az utóbbi kategóriába tartozó tanítási eljárást alkalmazunk nevetésfelismerésre tanított mély neurális hálók esetében. Először bemutatjuk az alkalmazott módszert (*valószínűségi mintavételezés*, [17]), majd elemezzük, hogy alkalmazása hogyan befolyásolja a neurális háló által generált valószínűségbecslések rejtett Markov modellben való alkalmazását. Ezután leírjuk a kísérleti környezetet (a felhasznált adatbázist, a pontosságmetrikákat és a DNN paramétereit), végül bemutatjuk és elemezzük az eredményeket.

2. Valószínűségi mintavételezés

Mint az osztályozó módszerek általában, a neurális hálók is érzékenyek arra, ha az egyes osztályokhoz nem egyenletesen állnak rendelkezésre tanítópéldák. Ilyen esetekben hajlamosak pontatlan valószínűségbecsléseket adni az alulreprezentált osztályokhoz tartozó példákra. Ennek kezelésére talán a legegyszerűbb megközelítés, ha a gyakoribb osztályokhoz tartozó tanítópéldák számát redukáljuk; ekkor azonban nyilvánvalóan információt is veszítünk, mely az osztályozási pontosság csökkenéséhez is vezethet. Egy másik megközelítés, ha inkább gyakrabban használjuk a ritkábban előforduló osztályok tanítópéldáit. Egy matematikailag jól meghatározott ilyen tanítási stratégia a valószínűségi mintavételezés (*probabilistic sampling*, [17,18]). Ennek során a következő tanítópéldát egy kétlépéses eljárásban választjuk ki: először a példa *osztályát* határozzuk meg valamely valószínűségi eloszlást követve, majd választunk egy tanítópéldát az adott osztályból. Az osztályok kiválasztásának valószínűségére az alábbi képlet szolgál:

$$P(c_k) = \lambda \frac{1}{K} + (1 - \lambda) Prior(c_k), \tag{1}$$

ahol $Prior(c_k)$ a k. osztály (c_k) előzetes (prior) valószínűsége, K az osztályok száma, míg $0 \leq \lambda \leq 1$ egy paraméter. $\lambda = 0$ esetén ez a képlet az eredeti osztályeloszlást adja, míg $\lambda = 1$ az egyenletes eloszláshoz vezet, melyet követve a tanítás során minden osztályból közelítőleg ugyanannyi példát használunk fel. Köztes λ értékeket használva lineárisan képezünk átmenetet a két eloszlás között.

Beszédfelismerés során ritkán használnak tanítópélda-mintavételezést, melynek véleményünk szerint több oka is van. Egyrészt a tanító adatbázisok gépi tanulási szempontból igen nagynak számítanak, így egy DNN kellően pontos modellt képes építeni az egyes fonémaállapotokról (melyek az osztályoknak felelnek meg). Egy további ok szerintünk, hogy az egyes osztályokhoz tartozó példák eloszlása relatíve egyenletes. (Ezt tovább erősíti a kontextusfüggő állapotmodellezés [19,20,21] alkalmazása, melynek egyik célja épp annak garantálása, hogy minden osztályhoz kellő számú tanítópélda álljon rendelkezésre.) Érdemes kitérnünk García-Moral és tsai [22] igen részletes tanulmányára, melyben tanítópéldákat hagytak el a gyakoribb osztályokból. Habár ezzel lényegesen fel tudták gyorsítani a neurális háló tanítását, beszédfelismerő rendszerük pontossága valamelyest csökkent. Tóth és Kocsor 2005-ben alkalmazták a fönt ismertetett valószínűségi mintavételezési módszert egy kisszótáras, izolált szavas felismerő akusztikus modelljének (sekély neurális háló) tanítására. García-Moral cikkével ellentétben ők ezzel növelni is tudták a felismerés pontosságát.

Ezek a tanulmányok beszédfelismerési kontextusban mintavételezték a tanítópéldákat az akusztikus modell tanítása során, mely feladatban az osztályok eloszlásának különbsége minimális. Ugyanakkor nevetés és a hasonló nemverbális hangjelenségek (pl. kitöltött szünetek) felismerése esetén az osztályok megoszlása sokkal kiegyensúlyozatlanabb, hiszen a felvételeknek csak egy töredéke (nevetések esetén pl. tipikusan 4-6%-a) felel meg a keresett jelenségnek. Ebben az esetben joggal várhatjuk, hogy valamely mintavételezési eljárás alkalmazása a tanítás során jelentősen javítja a detektálás hatékonyságát.

1. táblázat. A BEA adatbázis felhasznált részének néhány jellemzője

		Halmaz			Összes
		Tanító	Fejlesztési	Teszt	felvétel
Felvételek összhossza (p:mp)		100:07	20:32	26:57	147:36
Nevetések	összhossza (p:mp)	7:53	1:55	2:14	12:01
	aránya	7,8%	9,3%	8,3%	8,1%
	gyakorisága (1/p)	5,21	5,07	5,53	5,25
	átlagos hossza (ms)	903	1106	901	930

2.1. Valószínűségi mintavételezés alkalmazása rejtett Markov modellben

Egy szokásos rejtett Markov modell minden keretszintű x_t megfigyelésvektorhoz és minden c_k állapothoz $p(x_t|c_k)$ valószínűség-becsléseket vár bemenetként. Mivel a neurális hálók a $P(c_k|x_t)$ értékeket becslik, a várt $p(x_t|c_k)$ értékeket a Bayes-tétel alkalmazásával kaphatjuk meg. Így egy HMM/ANN vagy HMM/DNN hibrid modell használatakor a neurális háló keretszintű kimeneteit el kell osztanunk a megfelelő osztály a priori valószínűségével ($P(c_k)$). Ezzel a kívánt $p(x_t|c_k)$ becsléseket kapjuk egy konstans szorzótól eltekintve, amely konstans szorzót azonban (a Viterbi keresés során alkalmazott maximalizálás miatt) figyelmen kívül hagyhatjuk.

Ugyanakkor Tóth és Kocsor [18] megmutatták, hogy amennyiben neurális hálóinkat $\lambda = 1$ paraméterrel tanítjuk (azaz egyenletes osztályeloszlást használunk), azok a $p(x_t|c_k)$ értékeket fogják becsülni (ismét egy konstansszorzótól eltekintve, amelyet megint figyelmen kívül hagyhatunk). Eszerint tehát $\lambda = 1$ paraméterérték használata esetén a háló által szolgáltatott valószínűségbecsléseket már nem kell tovább transzformálnunk, hanem azokat közvetlenül használhatjuk egy rejtett Markov modellben.

Elviekben tehát vagy $\lambda = 0$ paraméterezést kellene használnunk, és osztanunk az osztályok prior valószínűségeivel ($P(c_k)$), vagy $\lambda = 1$-et, és nem alkalmazni a Bayes-formulát. A gyakorlatban azonban a valószínűségbecslések nem pontosak, így jobb eredményeket kaphatunk köztes λ paraméterértékek használatával. Tóth és Kocsor cikkében [18] szintén köztes λ értékek adódtak optimálisnak. Mivel ebben az esetben nem egyértelmű, hogy érdemes-e alkalmaznunk a Bayes-formulát, mi mind a két stratégiát ki fogjuk próbálni.

3. Kísérletek

3.1. Adatbázis

Kísérleteinket a BEA adatbázis [23] egy részhalmazán végeztük. A BEA a legnagyobb magyar szabadon elérhető beszédadatbázis, melynek teljes felvételhossza 260 óra összesen 280 beszélőtől, hangszigetelt stúdiókörülmények között rögzítve.

Az adatbázis egy lényeges tulajdonsága, hogy spontán beszédet tartalmaz, mely fontos kritériuma annak, hogy nevetést tartalmazzon. Kísérleteinket 62 felvételen végeztük; 42-n tanítottuk az akusztikus neurális hálókat, 10-et fejlesztési halmazként, 10-et pedig tesztként használtunk. A tanító rész összesen 100, a fejlesztési halmaz 21, míg a teszthalmaz összesen 27 perc hosszú volt.

Az 1. táblázat tartalmazza a kísérletekhez használt felvételek néhány nevetés-specifikus jellemzőjét. Látható, hogy habár a fejlesztési és a teszthalmazt véletlenszerűen választottuk és csupán tíz-tíz felvételből állnak, elég jól reprezentálják a teljes hanganyagot. Ezen az adathalmazon a nevetésnek annotált részek aránya az irodalomban jellemzően említett 4-6%-nál valamivel nagyobbnak, 8% körülinek adódott, mely azonban még így is csak a felvételek töredéke. Az átlagos nevetéshossz majdnem egy másodperc, amely meglepően magas, azonban más cikkekben (pl. [16]) is hasonló értékekkel találkozhatunk.

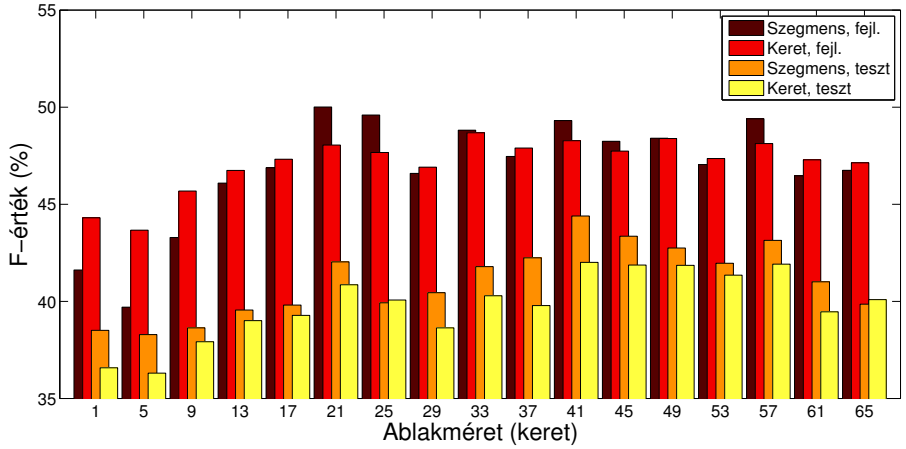
3.2. Kiértékelés

A nevetésdetektálás feladatánál nincs olyan egyértelműen elterjedt kiértékelési metrika, mint amilyen a szószintű hiba a beszédfelismerés területén. A legegyszerűbb megoldás, ha keretszintű pontosságot vizsgálunk; az osztályozási pontosság azonban közismerten nem rangsorolja jól a modelleket, ha az osztályeloszlás nagyon kiegyensúlyozatlan. Ennek egy finomításaként értékelhető az a gyakran használt megközelítés (ld. pl. [13,14,15]), melyben a keresett jelenséghez tartozó, keretszintű osztályvalószínűségekre meghatározzuk a ROC görbét, valamint a görbe alatti területet (Area Under Curve, AUC). Ennél életszerűbb kritériumnak gondoljuk ugyanakkor, hogy a valószínűségbecslésekből egy rejtett Markov modell segítségével szegmensszintű (kezdet- és végponttal rendelkező) előfordulás-hipotéziseket alkossunk, és a modellt ezek alapján értékeljük.

Tekintve, hogy a nevetésfelismerés egy standard információ-visszakeresési (Information Retrieval, IR) feladat, szokásos IR metrikákat számoltunk a modellek pontosságának mérésére: pontosságot (*precision*), fedést (*recall*) és F-értéket (*F-measure* vagy F_1). Ezeket csak a nevetés osztályra számítottuk ki, azonban két megközelítést is alkalmaztunk. Az egyikben nevetés*szegmenseket* vizsgáltunk (egy annotált szegmenst akkor tekintettünk megtaláltnak, ha egy hipotézis szegmens metszette a referencia annotációt és a két szegmens közepe maximum 0,5 másodpercre esett egymástól [24]). A másikban a rejtett Markov modell kimenetét keretszintre konvertáltuk, és a három metrikát a keretekre számítottuk ki [16].

3.3. A neurális háló és paraméterei

Saját neurálisháló-implementációnkat használtuk, mellyel korábban sok különböző feladaton értünk el jó eredményeket (pl. [25,26,27,28]). A neurális hálókat keretszinten tanítottuk, az FBANK jellemzőkészletet használva, amely 40 Mel szűrősor energiáiból, illetve azok első- és másodrendű deriváltjaiból áll [29]. Alkalmaztuk azt a fonémaosztályozás esetén bevett megoldást is, hogy a szomszédos



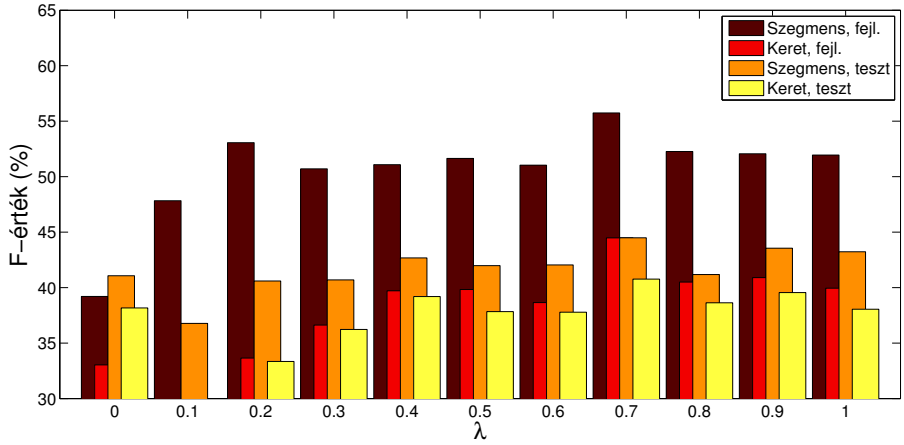
1. ábra. Átlagos F_1-értékek a tanításra használt mozgó ablak méretének függvényében

keretek jellemzővektorait is felhasználtuk az egyes keretek osztályozása során. Az alkalmazott neurális hálók előzetes tesztek eredményei alapján öt rejtett réteggel rendelkeztek, melyek mindegyikében 256 rectifier függvényt alkalmazó neuron volt, míg a kimeneti rétegben softmax függvényt használtunk. A súlyokat L2 regularizációval tartottuk kordában.

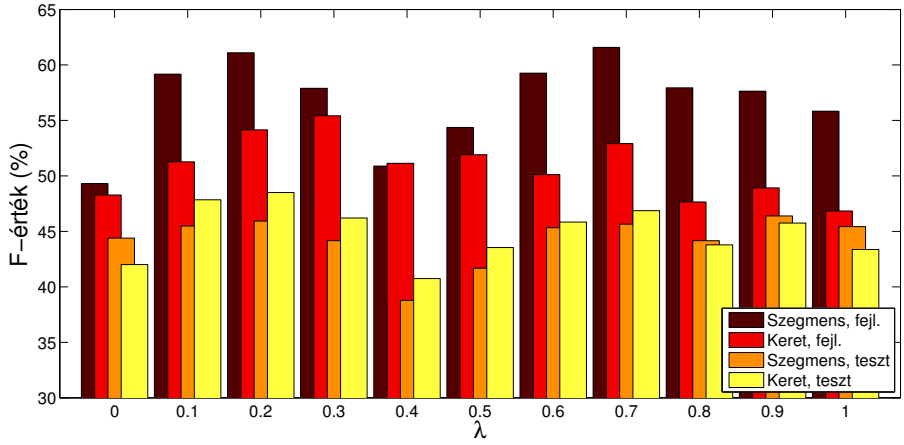
Mivel a neurális háló tanítása sztochasztikus folyamat (köszönhetően a súlyok véletlen inicializálásának), minden tesztelt λ paraméterváltozatra öt-öt hálót tanítottunk, és a kapott pontosságértékeket kiátlagoltuk. Salamin és tsai. [16] dolgozatát követve keretszintű nyelvi modellt számítottunk a tanítási halmazon; ennek súlyát minden neurális hálóra külön-külön, a fejlesztési halmazon határoztuk meg. Külön nyelvimodell-súlyt állapítottunk meg annak függvényében is, hogy a pontosságot szegmens- vagy keretszinten mértük-e.

Viszonyítási alapként teljes mintavételezéssel tanított mély neurális hálók szolgáltak. A tanítást a kereteken vett csúszó ablakokon végeztük, melyek optimális méretét előzetes tesztekkel határoztuk meg. Ennek során a csúszó ablak $1, 5, \ldots, 65$ keret széles volt, a DNN által szolgáltatott keretszintű valószínűségbecsléseket pedig a Bayes-formulával korrigáltuk. Az eredmények az 1. ábrán láthatóak; a fejlesztési halmazon mért szegmens- és keretalapú F_1-értékek alapján az eredmények alapján a mozgó ablak méretét a továbbiakban 41 keretnek választottuk.

A valószínűségi mintavételezés λ paraméterét a $0 < \lambda \leq 1$ intervallumban teszteltük, $0,1$-es lépésközt használva, minden λ értékre öt hálót tanítva. Az optimális λ értéket a fejlesztési halmazon határoztuk meg. Teszteltük, hogy a posterior valószínűségeket érdemes-e a Bayes-tétel alkalmazásával transzformálnunk, vagy inkább az eredeti értékeket érdemes használnunk. Fontos észrevétel, hogy ehhez nem volt szükséges új hálókat tanítanunk.



2. ábra. Átlagos F_1-értékek a valószínűségi mintavételezés λ paraméterének függvényében, a Bayes-tétel alkalmazása nélkül



3. ábra. Átlagos F_1-értékek a valószínűségi mintavételezés λ paraméterének függvényében, a Bayes-tétel alkalmazása után

3.4. Eredmények

A 2. és 3. ábrán láthatóak az átlagos F_1 értékek a λ paraméter függvényében. Ahogyan az várható volt, az eredeti posterior értékek használata esetén (ld. 2. ábra) a magasabb λ értékek ($\lambda \geq 0,7$), míg a Bayes-formulával korrigált valószínűségbecslések esetén (ld. 3. ábra) inkább az alacsonyabb λ értékek mellett mért pontosságok adódtak valamivel magasabbnak. Látható, hogy mindkét esetben köztes ($0 < \lambda < 1$) λ értékek adódtak optimálisnak. Ugyanakkor az eredeti posteriorok használatával nem sikerült elérni a referencia-értékeket (amelyek a Bayes-képlet alkalmazásával, viszont teljes mintavételezés mellett születtek).

2. táblázat. A valószínűségi mintavételezési eljárással kapott optimális átlagos F_1-értékek

Kiértékelés szintje	Halmaz	Priorokkal osztás	Opt. λ	Pontosság Prec.	Rec.	F_1	Relatív hibacsökk.
Szegmens	Fejlesztési	nem	0,7	53,51%	58,46%	55,74%	12,68%
		igen	0,7	59,36%	64,03%	61,58%	24,21%
		igen	—	41,11%	62,50%	49,31%	—
	Teszt	nem	0,7	43,85%	45,37%	44,49%	0,16%
		igen	0,7	45,96%	45,37%	45,65%	2,25%
		igen	—	39,42%	51,55%	44,40%	—
Keret	Fejlesztési	nem	0,7	61,45%	34,96%	44,48%	-7,33%
		igen	0,3	46,49%	68,60%	55,42%	13,82%
		igen	—	38,02%	66,53%	48,27%	—
	Teszt	nem	0,7	51,60%	33,81%	40,77%	-2,14%
		igen	0,3	36,09%	64,22%	46,20%	7,23%
		igen	—	30,94%	66,14%	42,01%	—

Az 2. táblázat foglalja össze a legjobb pontosságértékeket a fejlesztési-, és az azonos meta-paraméterekkel született pontosságértékeket a teszthalmazon. A táblázatban az átlagos F_1 érték mellett a pontosságot és a fedést is feltüntettük. Látható, hogy az F_1 értékeken szegmensszinten lényegesen sikerült javítani a fejlesztési halmazon, azonban a teszthalmazra ennek csak egy töredékét sikerült átvinni. A keretek szintjén enyhe csökkenést tapasztalhatunk, mikor a mély neurális háló valószínűségbecsléseit közvetlenül alkalmaztuk a rejtett Markov modellben; a Bayes-tétel alkalmazását követően azonban az F_1-értékek a teszthalmazon is jelentősen javultak: a teszthalmazon 42%-os viszonyítási értékről 46% fölé nőttek, mely 7%-os relatív hibacsökkentést jelent.

A referencia esetekben a fedés jóval magasabb volt, mint a pontosság, ami sok fals pozitív találatra utal. Valószínűségi mintavételezést használva szegmensszinten a két érték szinte tökéletesen kiegyensúlyozott, keretszinten azonban eltérések tapasztalhatóak. Ez arra utal, hogy a rejtett Markov modell ugyan elég jó pontossággal megtalálja a nevetés-előfordulásokat, a szegmensek határait illetően azonban bizonytalan.

A mély hálók kimeneteit változatlan formában használva keretszinten magas pontosságot és alacsony fedést, míg a Bayes-tétel után relatíve alacsony pontosságot és magas fedést láthatunk. Ez elég logikus: a mély háló vélhetően alapvetően alacsony valószínűségértékeket becsült a nevetés osztályra, melyeket közvetlenül használva a rejtett Markov modellben csak az egyértelműen nevetést tartalmazó keretek lettek azonosítva. Az osztályok a priori valószínűségeivel osztva a hálók kimenetét azonban változik a helyzet: mivel a nevetés osztálynak alacsony az a priori valószínűsége, a beszédet és csendet jelentő osztálynak pedig elég magas, a Bayes-tétel alkalmazásával a nevetésre adott becsléseinket

nagymértékben megnöveljük, míg a másik osztályét csak alig. Így vélhetően a nevetést tartalmazó szegmensek környezetében található kereteket is nevetésnek azonosítjuk, mely a szegmensszintű pontosságértékeket nem változtatja meg, keretszinten azonban csökkenti a fals negatív és növeli a fals pozitív találatok arányát.

3.5. Konklúzió

Jelen dolgozatban spontán beszédben kerestük nevetések előfordulását egy rejtett Markov modell/mély neurális háló keretrendszerben. Mivel a nevetés a hanganyagnak csak mintegy 8%-át tette ki, a tanítópéldák osztályeloszlása egyenetlen volt, így mély neurális hálónk tanítása szuboptimális volt. Kísérletileg megmutattuk, hogy a tanítás javítható a tanítópéldák újra-mintavételezésével. A valószínűségi mintavételezés nevű eljárás használatával a keretszintű hibát 7%-kal tudtuk csökkenteni egy magyar nyelvű, spontán beszédet tartalmazó adatbázison.

Hivatkozások

1. Hogan, K.: Can't Get Through: Eight Barriers to Communication. Pelican Publishing (2003)
2. Halberstadt, A.G.: Family socialization of emotional expression and nonverbal communication styles and skills. Journal of personality and social psychology **51**(4) (1986) 827
3. Johannesen, R.L.: The emerging concept of communication as dialogue. (1971)
4. Isbister, K., Nass, C.: Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. International journal of human-computer studies **53**(2) (2000) 251–267
5. Glenn, P.: Laughter in interaction. Cambridge University Press, Cambridge, UK (2003)
6. Hámori, A.: Nevetés a társalgásban. In Laczkó, K., Tátrai, S., eds.: Elmélet és módszer. ELTE Eötvös József Collegium, Budapest, Hungary (2014) 105–129
7. Holmes, J., Marra, M.: Having a laugh at work: How humour contributes to workplace culture. Journal of Pragmatics **34**(12) (2002) 1683–1710
8. Neuberger, T.: Nonverbális hangjelenségek a spontán beszédben. In Gósy, M., ed.: Beszéd, adatbázis, kutatások. Akadémiai Kiadó, Budapest (2012) 215–235
9. Gosztolya, G., Beke, A., Neuberger, T.: Nevetések automatikus felismerése mély neurális hálók használatával. In: MSZNY, Szeged (2016) 122–133
10. Gosztolya, G., Beke, A., Tóth, L., Neuberger, T.: Laughter classification using deep rectifier neural networks with a minimal feature subset. Archives of Acoustics **41**(4) (2016) 669–682
11. Knox, M.T., Mirghafori, N.: Automatic laughter detection using neural networks. In: Proceedings of Interspeech, Antwerp, Belgium (2007) 2973–2976
12. Neuberger, T., Beke, A.: Automatic laughter detection in spontaneous speech using GMM–SVM method. In: TSD. (2013) 113–120
13. Gupta, R., Audhkhasi, K., Lee, S., Narayanan, S.S.: Detecting paralinguistic events in audio stream using context in features and probabilistic decisions. Computer, Speech and Language **36**(1) (2016) 72–92

14. Kaya, H., Ercetin, A., Salah, A., Gürgen, S.: Random forests for laughter detection. In: WASSS. (2013)
15. Brueckner, R., Schuller, B.: Social signal classification using deep BLSTM recurrent neural networks. In: Proceedings of ICASSP. (2014) 4856–4860
16. Salamin, H., Polychroniou, A., Vinciarelli, A.: Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In: Proceedings of SMC. (2013) 4282–4287
17. Lawrence, S., Burns, I., Back, A., Tsoi, A., Giles, C.: Chapter 14: Neural network classification and prior class probabilities. In: Neural Networks: Tricks of the Trade. Springer (1998) 299–313
18. Tóth, L., Kocsor, A.: Training HMM/ANN hybrid speech recognizers by probabilistic sampling. In: Proceedings of ICANN. (2005) 597–603
19. Young, S.J., Odell, J.J., Woodland, P.C.: Tree-based state tying for high accuracy acoustic modelling. In: HLT. (1994) 307–312
20. Wang, W., Tang, H., Livescu, K.: Triphone state-tying via deep canonical correlation analysis. In: Interspeech, San Francisco, USA (Sep 2016) 3444–3448
21. Gosztolya, G., Grósz, T., Tóth, L., Imseng, D.: Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying. In: ICASSP, Brisbane, Ausztrália (2015) 4570–4574
22. García-Moral, A.I., Solera-Urena, R., Peláez-Moreno, C., de María, F.D.: Data balancing for efficient training of hybrid ANN/HMM Automatic Speech Recognition systems. IEEE Trans. ASLP **19**(3) (2011) 468–481
23. Gósy, M.: Bea a multifunctional hungarian spoken language database. The Phonetician **105**(106) (2012) 50–61
24. : NIST Spoken Term Detection 2006 Evaluation Plan. http://www.nist.gov/speech/tests/std/docs/std06-evalplan-v10.pdf. (2006)
25. Tóth, L.: Phone recognition with hierarchical Convolutional Deep Maxout Networks. EURASIP Journal on Audio, Speech, and Music Processing **2015**(25) (2015) 707–710
26. Gosztolya, G.: On evaluation metrics for social signal detection. In: Interspeech, Drezda, Németország (2015) 2504–2508
27. 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
28. 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
29. Young, S., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge, Anglia (2006)