Környezetfüggő akusztikai modellek létrehozása Kullback-Leibler–divergencia alapú klaszterezéssel

Grósz Tamás, Gosztolya Gábor, Tóth László

MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos krt. 103.
{ groszt, ggabor, tothl } @ inf.u-szeged.hu

Kivonat Az elmúlt néhány év során a beszédfelismerésben a rejtett Markov modellek Gauss keverékmodelljeit (Gaussian Mixture Models, GMM) háttérbe szorították a mély neuronhálók (Deep Neural Networks, DNN). Ugyanakkor a neuronhálókra épülő felismerők számos olyan tanítási algoritmust megörököltek (változatlan formában vagy apróbb változtatásokkal), melyeket eredetileg HMM/GMM rendszerekhez fejlesztettek ki; ezek optimalitása az új környezetben egyáltalán nem garantált. Ilyen tanítási lépés a környezetfüggő fonémaállapot-halmaz meghatározása is, amire az általánosan elfogadott megoldás egy döntésifa-alapú algoritmus. Ez az eljárás arra törekszik, hogy az előálló állapotokhoz tartozó példák Gauss-görbékkel optimálisan modellezhetőek legyenek. Jelen cikkünkben egy alternatív eljárást vizsgálunk meg, mely a döntési fát egy Kullback-Leibler–divergencia alapú döntési kritériumra támaszkodva építi fel. Feltételezésünk szerint ez a kritérium alkalmasabb a neuronháló kimeneteinek leírására, mint a gaussos modellezés. A módszert korábban már sikeresen alkalmazták egy KL-HMM rendszerben, most pedig megmutatjuk, hogy egy HMM/DNN hibrid rendszerben is működőképes. Alkalmazásával 4%-os relatív hibacsökkenést értünk el egy nagyszótáras szófelismerési feladaton. [1]

Kulcsszavak: beszédfelismerés, környezetfüggő fonémaállapotok, mély neuronhálók, Kullback-Leibler divergencia

1. Bevezetés

Az utóbbi pár évben a hagyományos Gauss keverékmodelleket (Gaussian Mixture Models, GMMs) alkalmazó beszédfelismerő rendszerek helyét átvették a mély neuronhálókra (Deep Neural Networks, DNN) épülő HMM/DNN hibridek. A rejtett Markov-modellek (Hidden Markov Models, HMM) megjelenése

[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.

óta azonban elég sok eljárást fejlesztettek ki, melyeket a DNN-t használó keretrendszerek is átvettek, holott ezek az algoritmusok DNN-ek használata esetén nem feltétlenül működnek optimálisan. Talán a legismertebb ilyen a „flat start" indítás (melyben a hangfelvételből és annak átiratából iterálva találjuk meg az egyes beszédhangok helyét), valamint a környezetfüggő (context-dependent, CD) fonémamodellek kialakítása.

Habár a HMM/ANN hibrid modellekben sokáig csak környezetfüggetlen modelleket alkalmaztak (azaz egy-egy beszédhangot önmagában, az azt megelőző és követő fonémák ignorálásával modelleztek), mostanra nyilvánvalóvá vált, hogy a nagypontosságú beszédfelismeréshez hibrid modellek esetében is célszerű környezetfüggő (trifón) beszédhangmodelleket alkalmazni. Az összes trifónt külön modellezni azonban nem hatékony, érdemes ehelyett az egymáshoz valamilyen szempontból hasonlóakat összevontan kezelni.

Erre a feladatra már megjelenése óta Young [1] és Odell [2] döntésifa-alapú klaszterezési módszerét szokás alkalmazni. Ez az eljárás a faépítés során egyetlen normális eloszlással modellezi az egy állapothalmazhoz tartozó összes példát, és arra törekedve osztja ketté a halmazt, hogy a két nem átfedő részhalmaz külön-külön optimálisan legyen modellezhető. Ez egy igen gyors eljárás, azonban, bár kapcsolódása egy HMM/GMM alapon nyugvó rendszerhez nyilvánvaló, egy neuronháló-alapú beszédfelismerő rendszer esetén optimalitása több okból is megkérdőjelezhető.

Az egyik ilyen ok, hogy a GMM-alapú eljárások feltételezik, hogy a jellemzők kovarianciamátrixa diagonális, azaz dekorrelált jellemzőkészletet (pl. MFCC) várnak el. Ugyanakkor a HMM/DNN hibrid rendszerek általában jobban teljesítenek egyszerűbb jellemzőkön (pl. Mel szűrősorok). Mivel a hagyományos HMM/GMM rendszerek ilyen jellemzővektorokra nem taníthatóak hatékonyan, először egy HMM/GMM rendszert kell tanítanunk hagyományos jellemzőkön, ennek segítségével elkészíteni a környezetfüggő állapotok összevont halmazait és a fonémák keretszintű illesztését, majd eldobni a már leszámolt jellemzővektorokat. Ehelyett logikusabbnak tűnik az állapotok összevonását egy neuronháló kimenete alapján végezni (Senior et al. [3]). Ennek finomított változata az utolsó rejtett réteg (Bacchiani et al. [4]) értékeit használni, esetleg ezen réteg kimeneteit normális eloszlású valószínűségi eloszlásokká konvertálni (Zhang et al. [5]).

Bár a felsorolt kutatók a neuronháló kimenetét igyekeztek az eljáráshoz igazítani, maga az állapotok összevonására szolgáló algoritmus minden esetben változatlan maradt, csupán annak bemenete változott meg. Ugyanakkor jogosnak tűnő ellenvetés, hogy az eljárásnak olyan környezetfüggő állapotokat kellene különválasztania, melyek külön kezelése az adott beszédfelismerő rendszerben alkalmazott eljárás (GMM vs. DNN) számára kedvezőbb. Mivel egy GMM és egy DNN tanítása során alapvetően más jellegű döntési függvényt optimalizálunk, annak vizsgálata, hogy egy normális eloszlással hogyan tudjuk modellezni az egyes állapotokhoz tartozó példákat, akár teljesen független is lehet attól, hogy egy mély neuronháló hogyan tudja modellezni az adott osztályt. Akkor viszont a normális eloszláson alapuló döntési kritérium helyett érdemesebb lenne valamilyen másfajta építési kritériumot alkalmazni.

A közelmúltban Imseng et al. a döntésifa-alapú eljárás olyan változatát dolgozta ki, mely közvetlenül a neuronhálók kimenetét használja [6]. A korábban felsorolt művekkel ellentétben, melyek a neuronháló-kimeneteket feltételes osztályvalószínűséggé konvertálták és normális eloszlással modellezték, ez az eljárás kihasználja, hogy egy neuronháló kimenetvektora diszkrét valószínűségi eloszlás. Ezek különbözőségének mérésére kézenfekvő választás a Kullback-Leibler–divergencia [7], így az állapothalmazokat meghatározó eljárás döntési kritériumában érdemesebb ezt használni a normális eloszlásra épülő, hagyományos döntési függvény helyett. Imseng et al. sikerrel alkalmazta ezt az algoritmust Kullback-Leibler–divergenciára épülő rendszerükben (KL-HMM) [8].

Jelen cikkünkben ezt az eljárást egy HMM/DNN hibrid beszédfelismerő rendszerben értékeljük ki. A teszteket egy 28 órányi magyar nyelvű híradófelvételt tartalmazó adatbázison [9] végezzük; viszonyítási alapnak egy HMM/DNN hibrid rendszert veszünk, melynek környezetfüggő fonémamodell-halmazait a bevett GMM-alapú eljárással állítjuk elő.

2. Döntésifa-alapú modellösszevonás

A döntésifa-alapú fonémamodell-összevonási algoritmus húsz évvel ezelőtti bevezetése óta [1] a nagyszótáras beszédfelismerő rendszerek tanításának elhagyhatatlan részévé vált. Alapötlete, hogy egy (környezetfüggetlen) állapot összes előfordulását összevonja egy $\mathcal{S}$ halmazba, majd ezen halmaz lépésenkénti kettéosztásával egy döntési fát épít. Az algoritmus minden lépésben kiválaszt egyet az előre definiált kérdések közül annak alapján, hogy az így előálló két nem átfedő részhalmaz elemei a lehető legjobban különbözzenek egymástól. Ezt a különbözőséget egy valószínűség-alapú döntési kritérium méri. Ez az eljárás annyira sikeresnek bizonyult, hogy kisebb javításokat (pl. a kérdések automatikus előállítását [10]) leszámítva azóta is változatlan formában használják.

2.1. Valószínűség-alapú döntési kritérium

Odell [2] megfogalmazott egy maximum likelihood-alapú döntési kritériumot, és adott is egy hatékony algoritmust a kiszámítására, a szétválasztási kritériumot a következő képlettel becsülve:

$$L(\mathcal{S}) \simeq -\frac{1}{2}\big(\log[(2\pi)^K|\Sigma(\mathcal{S})|] + K\big)\sum_{s\in\mathcal{S}} N(s), \tag{1}$$

ahol $s \in \mathcal{S}$ jelöli az egyes állapotokat, $\Sigma(\mathcal{S})$ az $\mathcal{S}$-ba tartozó példák szórása, míg $N(s)$ az s állapothoz tartozó példák száma a tanítóhalmazban. Így azt a q kérdést kell választanunk a példák kettéválasztására, melyre a $\Delta L(q|\mathcal{S})$ valószínűségkülönbség maximális, ahol

$$\Delta L(q|\mathcal{S}) = \big(L(\mathcal{S}_y(q)) + L(\mathcal{S}_n(q))\big) + L(\mathcal{S}), \tag{2}$$

és $\mathcal{S}_y(q)$ és $\mathcal{S}_n(q)$ az $\mathcal{S}$ halmaz két nem átfedő részhalmaza a q kérdésre adott válasznak megfelelően. Látható, hogy a valószínűség-értékek nem függnek a tanítópéldáktól, csupán azok szórásától és az egyes állapotokhoz tartozó tanítópéldák (keretek) számától. Ez a feltevés tökéletesen illeszkedik egy GMM-alapú beszédfelismerő rendszerhez, ugyanakkor egy HMM/DNN hibridben valamely más döntési kritérium használata a mély neuronhálókhoz jobban illeszkedő állapothalmazhoz is vezethet.

2.2. Kullback-Leibler–divergencia alapú döntési kritérium

Ezt a kritériumot Imseng et al. vezette be [11], és sikeresen alkalmazták KL-HMM rendszerükben. A következőkben [6] és [8] alapján röviden ismertetjük az eljárást.

Habár a Kullback-Leibler–divergencia nem távolságfüggvény (például nem szimmetrikus), a szimmetrikus KL-divergenciára épülő költségfüggvény kiszámítására nincs zárt formula. Emiatt az aszimmetrikus KL-divergenciát fogjuk alkalmazni, mely két K-dimenziós posterior-vektorra (z_t és y_s) a következő alakot veszi fel [7]:

$$D_{KL}(y_s||z_t) = \sum_{k=1}^{K} y_s(k) \log \frac{y_s(k)}{z_t(k)}. \tag{3}$$

A KL-divergencia mindig nemnegatív, és pontosan akkor nulla, ha a két eloszlásvektor megegyezik. Így a faépítés során a likelihood maximalizálása helyett minimalizáljuk a KL-divergenciát:

$$D_{KL}(\mathcal{S}) = \sum_{s \in S} \sum_{f \in F(s)} \sum_{k=1}^{K} y_{\mathcal{S}}(k) \log \frac{y_{\mathcal{S}}}{z_f(k)}, \tag{4}$$

ahol $\mathcal{S}$ állapotok egy halmaza, és $F(s)$ az s állapothoz tartozó tanítóminták halmaza. Az $\mathcal{S}$ halmazhoz tartozó $y_{\mathcal{S}}$ posterior valószínűségi vektor $\mathcal{S}$ elemeinek mértani közepeként számítható, azaz

$$y_{\mathcal{S}}(k) = \frac{\left(\prod_{s \in \mathcal{S}} \prod_{f \in F(s)} z_f(k)\right)^{\frac{1}{N(S)}}}{\sum_{k=1}^{K} \tilde{y}_{\mathcal{S}}(k)}. \tag{5}$$

Néhány behelyettesítő és egyszerűsítő lépés után a következőt kapjuk [6]:

$$D_{KL}(\mathcal{S}) = -\sum_{s \in \mathcal{S}} N(s) \log \sum_{k=1}^{K} \tilde{y}_{\mathcal{S}}(k), \tag{6}$$

tehát a $\mathcal{S}$ állapothalmazhoz tartozó KL-divergenciát kiszámíthatjuk az egyes állapotok y_s és $N(s)$ értékei alapján.

Egy $\mathcal{S}$ állapothalmaz kettéosztása során kézenfekvő azt a kérdést választani, amely maximalizálja a KL-divergencia különbségét ($\Delta D_{KL}(q|\mathcal{S})$):

$$\Delta D_{KL}(q|\mathcal{S}) = D_{KL}(\mathcal{S}) - \big(D_{KL}(\mathcal{S}_y(q)) + D_{KL}(\mathcal{S}_n(q))\big). \tag{7}$$

3. KL-alapú állapotösszevonás HMM/DNN hibrid rendszerekben

Viszonyítási alapként a hagyományos tanítási utat követtük: első lépésben környezetfüggő HMM/GMM fonémamodelleket tanítottunk, majd ezeket felhasználva kényszerített illesztéssel állítottuk elő a tanító címkéket a DNN számára. Ez a módszer MFCC jellemzőkészletet használ, megvalósításához a HTK [12] programcsomagot használtuk. A HMM/GMM környezetfüggő fonémamodellek tanítása során a hagyományos normáliseloszlás-alapú állapotösszevonást alkalmaztuk, majd miután megkaptuk a klaszterezett állapotokat, felhasználásukkal egy mély neuronhálót tanítottunk. Az így tanított DNN-t használtuk a dekódolás során akusztikus modellként, a HTK módosított Hdecode rutinja segítségével.

A KL-alapú klaszterező algoritmus bemenetként környezetfüggetlen állapotok posterior valószínűségeit várja. Ezen értékek előállításához egy környezetfüggetlen *segéd neuronhálót* használtunk (a keretszintű címkézést a fönti HMM/GMM rendszer szolgáltatta). Ezután alkalmaztuk a KL-alapú klaszterező algoritmust a segédháló kimenetére, és a környezetfüggő mély neuronhálót az így kapott összevont állapotokat címkeként használva tanítottuk be. A viszonyítási alapként szolgáló módszerhez hasonlóan itt is a klaszterezés után tanított mély hálót használtuk a felismerés során.

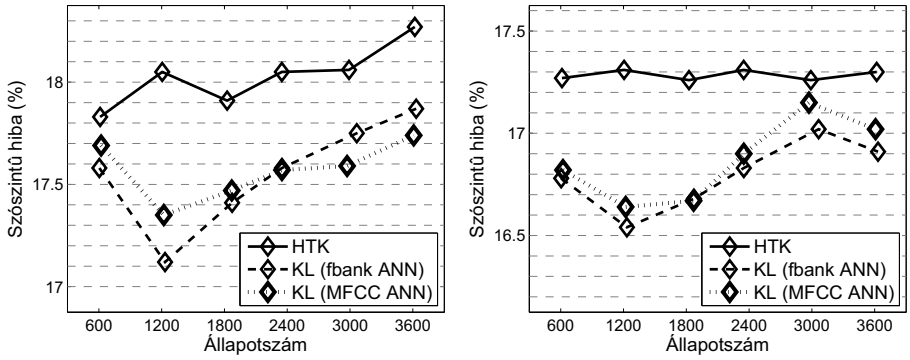
4. A kísérletek technikai jellemzői

A hibrid rendszerünk DNN komponenseként egy mély rectifier hálót [13] alkalmaztunk, amelynek fő előnye, hogy körülményes előtanítási módszerek nélkül, hagyományos backpropagation algoritmussal is hatékonyan tanítható [14]. Saját implementációnkat használtuk, amellyel a TIMIT adatbázison az általunk ismert legjobb eredményt, 16, 7%-os fonémaszintű hibát tudtunk elérni [15].

Az akusztikus modellezésre használt mély rectifier hálónk 5 rejtett rétegből állt, mindegyikben 1000 neuronnal, míg a kimeneti rétegben a softmax aktivációs függvényt alkalmaztuk. Bemenetként az ún. FBANK jellemzőkészletet használtuk [12], amely 40 mel szűrősor energiáiból, illetve azok első- és másodrendű deriváltjaiból állt.

Kísérleteinket híradófelvételeken végeztük [9]. Az adatbázis összesen 28 órányi hangzóanyagot tartalmaz, melyet a szokásos felosztásban használtunk: 22 órányi anyag volt a betanítási rész, 2 órányi a fejlesztési halmaz, a maradék 4 órányi hanganyag pedig a tesztelésre szolgáló blokk. Az adatbázisban összesen 13 467 különböző trifón fordult elő, ami összesen 40 401 kiindulási fonémaállapotot eredményezett.

A segéd-neuronhálók inputjaként először, a HMM/GMM rendszerrel megegyezően, MFCC jellemzőkészletet használtunk, majd kipróbáltuk az FBANK jellemzőkészletet is. A viszonyítási alapként szolgáló módszer esetében eltérő jellemzőkészletet kellett használnunk a klaszterezéshez és az akusztikus modell tanításához (MFCC vs. FBANK), mivel az FBANK jellemzőkön tanított GMM-ek használhatatlan eredményt adtak volna. A KL-klaszterezés hátránya, hogy



1. ábra. Az elért szószintű hibaarányok az állapotok számának függvényében a fejlesztési (balra) és a teszthalmazon (jobbra)

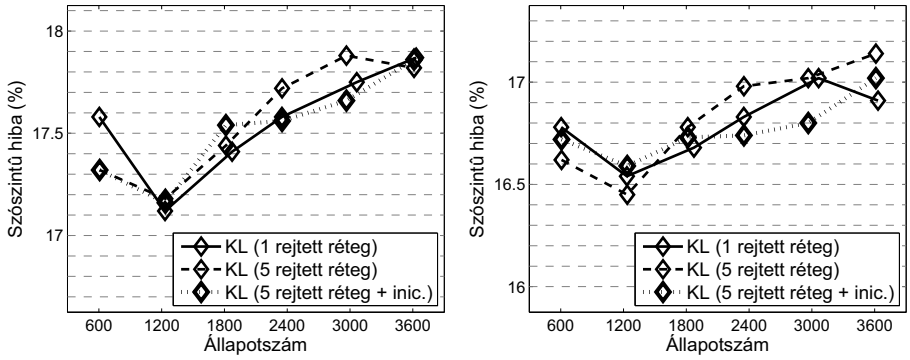
két neuronhálót kell tanítanunk; ennek csökkentése céljából kísérletet tettünk a segéd-neuronháló „újrahasznosíthatóságára" a második háló tanítása során. A klaszterezési eljárások küszöbértékeit úgy választottuk meg, hogy végül körülbelül 600, 1200, 1800, 2400, 3000 és 3600 összevont állapotot kapjunk.

5. Eredmények

Ahogy az az 1. ábrán megfigyelhető, a KL-divergencia–alapú klaszterezési algoritmus mindkét halmazon következetesen és szignifikánsan jobban teljesített, mint a hagyományos GMM/HMM alapú eljárás. A hagyományos módszer optimuma 600 környezetfüggő állapot körül van, habár a szószintű pontosságok minden kipróbált állapotszám esetén nagyon hasonlóan alakulnak. A KL-alapú algoritmus optimuma 1200 összevont állapotnál van: itt mintegy 4%-os relatív hibaarány-csökkenést hoz az alkalmazása a standard eljárás legjobb eredményéhez viszonyítva. A segéd-neuronháló két kipróbált változata közül a mel szűrősorokat használó bizonyult valamivel jobbnak (ezt a jellemzőkészletet használtuk a mély neuronháló tanításánál is), bár a különbség nem jelentős.

A KL-divergenciát használó klaszterezési eljárás alapvetően a segéd-neuronháló kimenete alapján dönt, így annak pontossága triviális módon meghatározza az állapothalmaz minőségét; ugyanakkor ennek mértéke egyáltalán nem nyilvánvaló, és mivel utána ezt a hálót eldobjuk, nem biztos, hogy megéri nagy pontosságú (és nagyméretű) segédhálót használni. Ennek kiderítésére további kísérleteket végeztünk: az eddigi egyrétegű háló helyett próbát tettünk egy mély (5 rejtett rétegű) neuronháló alkalmazásával is.

Egy másik lehetőség a segédháló felhasználása a végső DNN súlyainak inicializálásához, mely egyrészt csökkentheti a tanítási időt, másrészt pontosabb akusztikus modellhez vezethet. Természetesen ez csak akkor megvalósítható, ha mindkét neuronháló azonos számú neuront használ a rejtett rétegeiben, továbbá azonos jellemzőkészleten dolgozik; ugyanakkor korábban azt tapasztaltuk, hogy szűrősorok használatával nem kapunk rosszabb eredményeket, mint MFCC-vel,



2. ábra. Az elért szószintű hibaarányok az állapotok számának függvényében a fejlesztési (balra) és a teszthalmazon (jobbra)

így ez nem nagy megkötés. Emiatt a továbbiakban minden segédhálót FBANK szűrősorokra tanítottuk. Ezt az inicializálási stratégiát kipróbáltuk az egy és az öt rejtett réteggel rendelkező segédhálók alkalmazása során is.

Az eredmények (ld. 2. ábra és 1. táblázat) alapján annak, hogy a segédháló egy vagy öt rejtett réteget tartalmaz, nincs különösebb jelentősége. Hasonlóképpen, bár az akusztikus mély neuronháló inicializálása a segédháló megfelelő súlyainak felhasználásával 2-3 iterációval csökkentette a tanítás időigényét, a betanított háló pontossága enyhén romlott.

1. táblázat. A különböző állapotklaszterezési eljárások használatával elért szószintű hibaarányok

Klaszterezési eljárás	Szószintű hiba (%)	
	Fejl. halmaz	Teszthalmaz
KL (MFCC ANN)	17.35%	16.64%
KL (fbank ANN)	17.12%	16.54%
KL (fbank ANN) + ANN inic.	17.38%	16.79%
KL (fbank ANN, 5 rejtett réteg)	17.18%	16.45%
KL (fbank ANN, 5 rejtett réteg) + ANN inic.	17.16%	16.59%
GMM/HMM	17.83%	17.26%

Az eredményeket összegezve kijelenthetjük, hogy a Kullback-Leibler–divergenciára épülő döntési kritérium használata a környezetfüggő állapothalmazok kialakítása során szignifikánsan csökkentette a felismerés szószintű hibáját. Következtetésünk megerősítése érdekében a közeljövőben tervezzük, hogy a módszert más adatbázisokon is kiértékeljük.

6. Konklúzió

Jelen cikkben egy olyan eljárás hatékonyságát vizsgáltuk meg, amely a környezetfüggő fonémamodellek halmazát egy Kullback-Leibler–divergenciára épülő kritérium használatával határozza meg. Azt feltételeztük, hogy ez a kritérium alkalmasabb a neuronháló kimeneteinek leírására, mint a gaussos modellezés. Az algoritmus környezetfüggetlen állapotok valószínűség-eloszlását várja bemenetként; erre a célra egy segéd neuronhálót tanítottunk. A módszert egy nagyszótáras beszédfelismerési feladaton teszteltük, és használatával szignifikánsan tudtuk csökkenteni a szószintű hibát a hagyományos normális eloszlásra alapuló döntési kritériumhoz képest, több mint 4%-os relatív hibaarány-csökkenést elérve.

Hivatkozások

1. Young, S.J., Odell, J.J., Woodland, P.C.: Tree-based state tying for high accuracy acoustic modelling. In: Proceedings of HLT. (1994) 307–312
2. Odell, J.: The Use of Context in Large Vocabulary Speech Recognition. PhD thesis, University of Cambridge (1995)
3. Senior, A., Heigold, G., Bacchiani, M., Liao, H.: GMM-free DNN training. In: Proceedings of ICASSP. (2014)
4. Bacchiani, M., Rybach, D.: Context dependent state tying for speech recognition using deep neural network acoustic models. In: Proceedings of ICASSP. (2014) 230–234
5. Zhang, C., Woodland, P.: Standalone training of context-dependent Deep Neural Network acoustic models. In: Proceedings of ICASSP. (2014) 5597–5601
6. Imseng, D., Dines, J.: Decision tree clustering for KL-HMM. Technical Report Idiap-Com-01-2012, Idiap Research Institute (2012)
7. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Statist. **22**(1) (1951) 79–86
8. Imseng, D., Dines, J., Motlicek, P., Garner, P., Bourlard, H.: Comparing different acoustic modeling techniques for multilingual boosting. In: Proceedings of Interspeech. (2012)
9. Grósz, T., Kovács, G., Tóth, L.: Új eredmények a mély neuronhálós magyar nyelvü beszédfelismerésben. In: Proceedings of MSZNY. (2014) 3–13
10. Beulen, K., Ney, H.: Automatic question generation for decision tree based state tying. In: Proceedings of ICASSP. (1998) 805–808
11. Razavi, M., Rasipuram, R., Magimai-Doss, M.: On modeling context-dependent clustered states: Comparing HMM/GMM, hybrid HMM/ANN and KL-HMM approaches. In: Proceedings of ICASSP. (2014)
12. 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, UK (2006)
13. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier networks. In: Proceedings of AISTATS. (2011) 315–323
14. Tóth, L.: Phone recognition with deep sparse rectifier neural networks. In: Proceedings of ICASSP. (2013) 6985–6989
15. Tóth, L.: Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition. In: Proceedings of ICASSP. (2014) 190–194