Szegedi Tudományegyetem

Ph.D értekezés tézisei

Tanítási módszerek mély neuronhálós akusztikus modellekhez beszédfelismerésben

Szerző:
Grósz Tamás

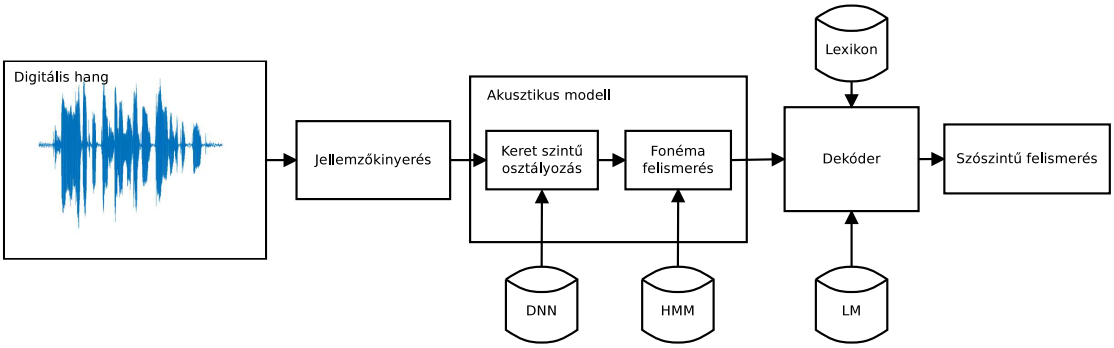
Témavezető:
Dr. Tóth László

Informatika Doktori Iskola
MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Informatikai Intézet
Szegedi Tudományegyetem

Szeged, 2018

1. Bevezetés

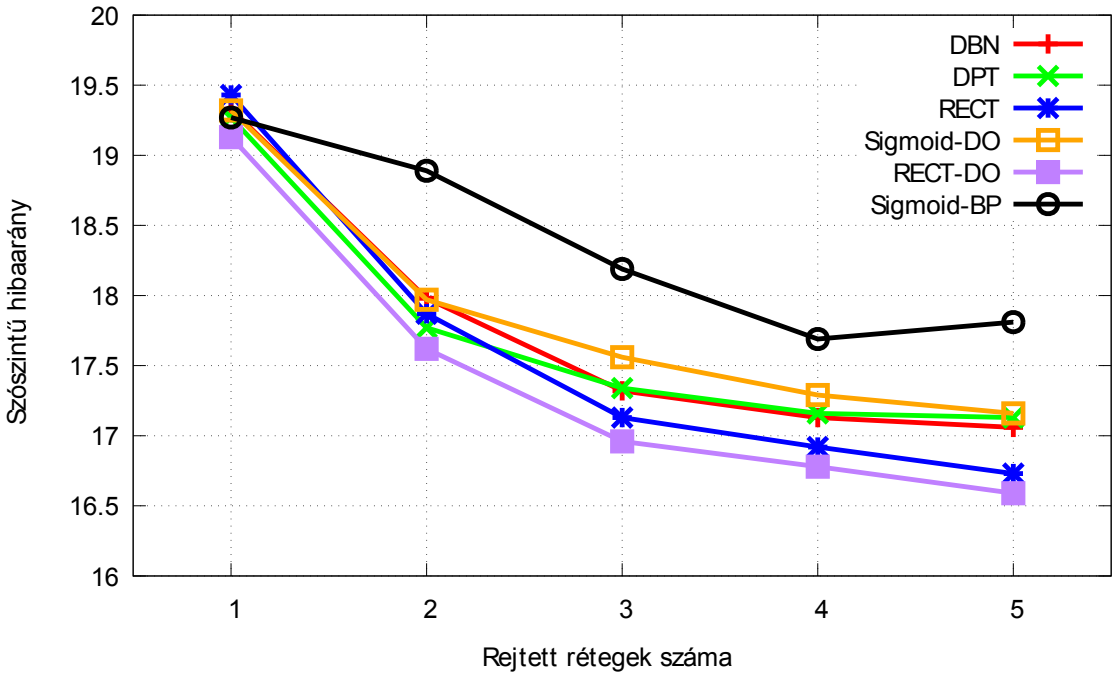
Az automatikus beszédfelismerés a beszédtechnológia egyik fontos területe, melynek célja, hogy a beszédből automatikus módon az elhangzott szósorozatot kinyerje. Évtizedeken át a beszédfelismerők rejtett Markov Modelleket (HMM) használtak Gauss-keverékmodellekkel (GMM), ez a HMM/GMM modell képviselte a legmodernebb technológiát egészen a mély neurális hálózatok (DNN) megjelenéséig. Napjainkban a mély hálók sikeresen leváltották a GMM komponenst a hagyományos HMM/GMM modellben, az így létrejött HMM/DNN hibrid modellt szemlélteti a 1. ábra [1]. A mély neuronhálók, ahogy a név is jelzi, egy fontos tulajdonságban térnek el a hagyományos neuronhálóktól, mégpedig a rejtett rétegek számában. Ezen új mély struktúrák tanítása viszont problémás a hagyományos módszerekkel, mivel azok nincsenek felkészítve a sok rejtett réteg által okozott nehézségek kezelésére. Pontosan ezért a mély hálók hatékony tanításához egyéb módosítások is szükségesek, mint például a rejtett neuronok aktivációs függvényének vagy a tanító algoritmusnak a lecserélése.



1. ábra: Egy modern HMM/DNN hibrid beszédfelismerő felépítése.

A manapság széleskörűen használt HMM/DNN hibrid modell számos algoritmust örökölt a korábbi HMM/GMM módszertől, azonban ezen algoritmusok optimalitása egyáltalán nem garantált az új modell esetén. A disszertáció fő fókuszában ezen algoritmusok módosítása áll, célunk, hogy a módosítás után azok jobban illeszkedjenek az új hibrid rendszerhez. Fő célunk egy teljesen GMM-mentes rendszer kidolgozása, amely képes HMM/DNN alapú beszédfelismerők tanítására GMM használata nélkül. A célunk eléréséhez két problémára kellett új megoldást kidolgoznunk, az első a beszédfelismerők betanítása részletes szegmentálási annotáció nélkül (angol szakzsargonban „flat start" tanítás), a második fontos feladat pedig a környezetfüggő (CD) beszédhang-modellek automatikus kialakítása.

A kidolgozott módszereket több angol és magyar nyelvű adatbázison is kiértékeltük, de hogy éreztessük a haladás mértékét, a disszertáció minden fejezetében közlünk eredményeket a Szeged magyar nyelvű Híradós adatbázis [2] felhasználásával is.



2. ábra: Szószintű hibaarányok a rejtett rétegek számának függvényében.

2. Mély neuronhálós tanítási módszerek összehasonlítása nagyszótáras beszédfelismerésben

A disszertációban első lépésként négy mély neuronhálós tanítási módszert hasonlítottunk össze. Az első módszer a Hinton és társai által kidolgozott eredeti előtanító algoritmus (DBN) [3], a második módszer pedig az úgynevezett diszkriminatív előtanítás (DPT), amelyet Seide és társai publikáltak [4]. Ezen két algoritmusban közös, hogy két fontos fázisból állnak; az előtanítás során inicializálják a neuronhálót, majd a második lépésben finomhangolják azt. A mély egyenirányított hálók (DRN) [5], a harmadik módszer, amit megvizsgáltunk, jelentősen eltérnek a korábbiaktól, hiszen ebben az esetben nem a tanítási algoritmus módosul, hanem a rejtett neuronok aktivációs függvénye. A szokásos Sigmoid függvény helyett az egyenirányított függvényt (RECT) használjuk, amelynek képlete: $max(0, x)$. A negyedik módszerként egy regularizációs technikát választottunk, az úgynevezett Dropout (DO) [6] algoritmust, melynek lényege, hogy tanítás során véletlenszerűen kikapcsolunk neuronokat a hálózatban. Fontos megjegyezni, hogy ez a módszer nem egy önálló algoritmus, hanem csak más módszerekkel (bármelyik korábbival) kombinálva használható.

Kísérleteinkben ezen módszereket hasonlítottuk össze a Szeged Híradós korpuszon. A 2. ábrán a kapott szószintű hibaarányok láthatóak. Megfigyelhető, hogy mind a négy módszer elég hasonló eredményeket tudott elérni, de a legjobbnak az egyenirányított hálózatok bizonyultak, tekintve, hogy ezek érték el a legjobb felismerési pontosságokat és betanításuk is kevesebb időt igényelt. Ezen eredményekre alapozva, a dolgozatom további részében csak egyenirányított neuronhálókat alkalmaztam.

Adatbázis	Módszer		Dev.	Teszt
TIMIT [7]	egyállapotú (39)	CTC + DRN	**26.69**%	**28.60**%
		MMI + DRN	27.70%	30.94%
		Kézi címkézés	27.26%	29.35%
		Kényszerített ill.	27.10%	28.92%
	egyállapotú (61)	CTC + DRN	**26.07**%	**27.34**%
		MMI + DRN	25.16%	27.89%
		Kézi címkézés	26.42%	27.94%
		Kényszerített ill.	25.92%	27.55%
	háromállapotú (183)	CTC + DRN	23.20%	24.41%
		MMI + DRN	**20.32**%	**22.76**%
		Kézi címkézés	22.75%	24.7%
		Kényszerített ill.	22.78%	24.48%
Hangoskönyv [8]	egyállapotú (52)	CTC + DRN	17.85%	16.55%
		MMI + DRN	**16.95**%	**16.12**%
		Kényszerített ill.	17.76%	16.98%
	háromállapotú (156)	CTC + DRN	12.58%	11.67%
		MMI + DRN	**10.08**%	**9.67**%
		Kényszerített ill.	12.53%	11.96%
Híradós [2]	egyállapotú (52)	CTC + DRN	25.96%	**25.58**%
		MMI + DRN	35.66%	65.26%
		Kényszerített ill.	**25.82**%	25.64%
	háromállapotú (156)	CTC + DRN	21.62%	21.23%
		MMI + DRN	**20.74**%	**20.42**%
		Kényszerített ill.	22.13%	21.74%

1. táblázat: Fonémaszintű felismerési hibák különböző módszerekkel.

3. Mély egyenirányított neurális hálók tanítása szekvenciatanuló módszerekkel

Miután kiválasztottuk a legjobb mély tanulásos módszert, a flat start nevű feladatra fordítottuk figyelmünket. Ezen feladat megoldása az első lépés minden beszédfelismerő rendszer létrehozása során, lényege, hogy meghatározzuk a kontextusfüggetlen címkék időbeli illesztését. A hosszútávú célunk egy teljesen mély hálókon alapuló módszer kidolgozása volt, ezért ebben a fejezetben két szekvenciatanuló módszert hasonlítottunk össze, amelyek alkalmasnak tűntek a kezdeti kontextusfüggetlen modellek tanítására. A konnekciós temporális osztályozás (CTC) algoritmust [9] vettettük össze a maximális kölcsönös információn (MMI) alapulóval [10]. Mindkét vizsgált módszert mély

egyenirányított hálók tanítására használtuk. Az alap MMI algoritmushoz több módosítást is javasoltunk, melyek lehetővé tették, hogy véletlenszerűen inicializált hálók tanítására használjuk ezt a módszert időben illesztett címkék nélkül. A legfontosabb módosítások, melyek lehetővé tették az MMI használatát flat start tanításra:

1. A keretek tanítási célértékét az előre-hátra algoritmussal határozzuk meg.
2. Beszédhang-szintű átiratokkal és környezetfüggetlen beszédhang-modellekkel dolgozunk.
3. Nem használunk a priori valószínűségeket és nyelvi modellt.
4. Az MMI hibafüggvényében a hányados értékét csak a legvalószínűbb felismerési útvonal felhasználásával közelítjük.
5. A tanítás hibáját a validációs halmazon mérjük, és ha ez a hiba növekedne, akkor visszatérünk az iteráció előtti paraméterekhez, viszont csökkentjük a tanulási rátát.

A kísérleteink során különböző fonémafelismerési feladatokon hasonlítottuk össze a két módszert, az eredményeket a 1. táblázatban láthatjuk. Mindegyik adatbázis esetén azt találtuk, hogy a szekvenciatanuló módszerek jobban működtek mint a hagyományos rendszerek, amelyeket egy HMM/GMM által generált időben illesztett címkék segítségével tanítottunk. Az eredményekből az is egyértelműen kiderült, hogy az MMI módszer jobb eredményeket képes elérni mint a CTC algoritmus. A CTC algoritmus egy további hátránya, hogy a használatával betanított hálók nem alkalmasak a címkék kényszerített illesztésére. Mindezeket figyelembe véve megállapítható, hogy az MMI módszer a legjobb választás a flat start lépés megoldására.

4. GMM-mentes mély neuronhálós beszédfelismerők

A flat start tanítás után az állapotkapcsolási algoritmust is adaptáltuk, célunk a GMM függőségek eltávolítása volt. A környezetfüggő állapotokat általában a standard algoritmussal állítják elő, annak ellenére, hogy az algoritmus speciálisan a Gauss-görbék illeszkedését használja ki, így optimalitása egy mély hálós rendszerben megkérdőjelezhető. Az utóbbi időben azonban több olyan állapotklaszterező algoritmust is publikáltak, amelyek megkísérlik a korábbi eljárást a mély neuronhálós modellezéshez igazítani.

Néhány új módszer csak a klaszterező algoritmus bemenetén változtat, azaz a klaszterezést a DNN kimenetén futtatják le, magát az algoritmust pedig egyáltalán nem módosítják [11, 12, 13, 14]. Más szerzők a bemenet kicserélésén túl a klaszterező eljárás döntési kritériumát is módosítják oly módon, hogy az jobban illeszkedjen a neuronhálós eloszlás-modellezéshez [15, 16].

Módszer	Dev.	Teszt
Iteratív CE	28.63%	20.47%
MMI	15.78%	10.07%
MMI+CE	15.43%	9.64%

2. táblázat: A különböző flat start módszerekkel tanított CI hálók szószintű hibája a WSJ adatbázison.

Vizsgálataink során három különböző módszert hasonlítottunk össze a saját Kullback-Leibler-divergencián alapuló módszerünkkel [15], ugyanazon a nagy szótáras beszédfelismerési feladaton. Kombinálva ezen módszereket a korábban bemutatott MMI-alapú flat start módszerrel megmutattuk, hogy lehetséges HMM/DNN beszédfelismerőket tanítani GMM használata nélkül is.

Flat start stratégia	Klaszterezési módszer	Dev.	Teszt
Iteratív CE	MFCC + Likelihood	11.02%	8.20%
	DNN + Likelihood	11.48%	7.64%
	DNN (rejtett) + Likelihood	11.05%	7.81%
	Kullback-Leibler	10.47%	**7.27%**
	Entropy	**10.24%**	**7.27%**
MMI	MFCC + Likelihood	8.58%	6.13%
	DNN + Likelihood	8.7%	6.47%
	DNN (rejtett) + Likelihood	8.85%	6.04%
	Kullback-Leibler	**8.06%**	**5.72%**
	Entrópia	**8.03%**	5.92%
MMI + CE	MFCC + Likelihood	8.79%	5.97%
	DNN + Likelihood	9.14%	6.45%
	DNN (rejtett) + Likelihood	9.43%	6.77%
	Kullback-Leibler	8.5%	**6.15%**
	Entrópia	**8.09%**	6.20%

3. táblázat: A WSJ korpuszon elért szószintű hibaarányok különböző flat start és klaszterezési módszerekkel.

Az algoritmusaink tesztelésére a jól ismert és széles körben használt Wall Street Journal (WSJ) angol nyelvű korpuszt [17] használtuk, amely 81 órányi olvasott beszédanyagot tartalmaz. Első vizsgálataink során a különböző flat start módszereket hasonlítottuk össze, a 2. táblázatban láthatóak az eredmények. Jól látható, hogy az MMI-alapú módszer sokkal jobban teljesített mint az iteratív algoritmus. A 3. táblázatban a klaszterezési módszereket vetettük össze, megállapítható, hogy azok a módszerek teljesítettek a legjobban, amelyek a döntési kritériumot is lecserélték.

Végül megvizsgáltuk, hogy a legjobb magyar nyelvű beszédfelismerő rendszerünk milyen jellegű hibákat vétett. Ehhez kigyűjtöttük a teszthalmaz egy részén előforduló szószintű hibákat, majd manuálisan kategorizáltuk és elemeztük azokat. A konklúziónk, hogy szükség lenne egy új magyar nyelvre kidolgozott hibametrikára, hiszen a manapság használt szószintű hibaarány (WER) néhány hibát, amely a megértést nem befolyásolja, sokkal súlyosabban bünteti mint az indokolt lenne.

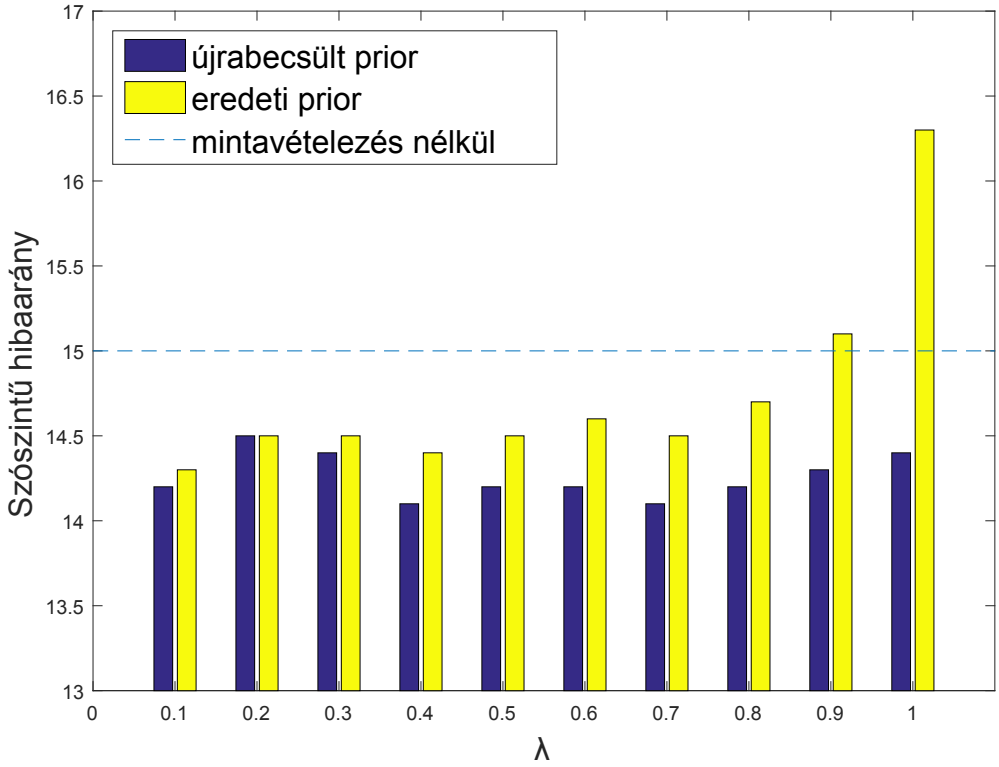
5. Kontextusfüggő mély neuronhálós akusztikus modellek tanítása valószínűségi mintavételezéssel

A manapság használatban lévő beszédfelismerőkben a DNN komponensek feladata, hogy állapot-kapcsolt trifónok posterior valószínűségét becsüljék. A problémát az jelenti, hogy a címkék eloszlása nem egyenletes, így a gyakorlatban az egyes osztályokhoz tartozó tanítópéldák száma jelentősen eltér. A tanítóadat egyenlőtlen eloszlása problémát jelent a legtöbb gépi tanuló algoritmusnak, ez alól a mély hálók sem kivételek.

A probléma megoldására a valószínűségi mintavételezés módszerét használtuk [18], amelynek előnye, hogy egyszerre alkalmazza az alul- és a felül-mintavételezést. Az adatbázis újramintavételezéséhez egy új osztályeloszlást definiál a módszer, ez az új eloszlás az eredeti és az egyenletes eloszlás lineáris kombinációjaként áll elő, a két eloszlás súlyát a λ paraméterrel tudjuk megadni. A korábbi tanulmányokhoz [18, 19] képest mi a priori valószínűségek újraszámlálására is javasoltunk egy módszert, erre azért volt szükség, mert az adat újramintavételezése révén jelentősen eltért a tanító és a teszt adatbázis egymástól.

A 3. ábrán láthatóak a valószínűségi mintavételezéssel elért szószintű hibaarányok a TED-LIUM adatbázison. Megfigyelhető, hogy az eredeti prior valószínűségeket használva nagy λ értékek esetén (közel az egyenletes eloszláshoz) romlik a pontosság, optimálisnak a viszonylag kicsi λ értékek bizonyultak, itt a 0.4. A valószínűségi mintavételezés segítségével 5% és 6% szószintű hibaarány redukciót sikerült elérnünk két nagy méretű adatbázison (TED-LIUM [20] és AMI [21]). Megmutattuk azt is, hogy ezzel a módszerrel a korábbi fejezetben bemutatott GMM-mentes rendszer is jobb eredményeket képes elérni. A kísérleti eredményeink alátámasztották azon sejtésünket is, hogy a priori valószínűségek újrabecslése kritikus az újramintavételezés miatt a tanító és teszt adat között fellépő különbség kezelése szempontjából. Ezek az újrabecsült priorok robusztusabbá tették a módszerünket, hatásukra a felismerési pontosságok csak csekély mértékben változtak, ahogy az egyenletes eloszlás felé mozgattuk az osztályok eloszlását a mintavételezés során.

Szintén sikeresen alkalmaztunk valószínűségi mintavételezéssel tanított DRN hálókat különböző paralingvisztikai feladatokon, amelyek a Computational Paralinguistics ChallengE (ComParE) versenysorozat részeként lettek meghirdetve. A paralingvisztika



3. ábra: Valószínűségi mintavételezéssel elért szószintű hibaarányok a TED-LIUM adatbázison.

fő célja, hogy a beszédből az elhangzott szakon túl egyéb, a beszélőre vonatkozó jellemzőket nyerjen ki. 2014-ben egy olyan rendszert tanítottunk, amely a beszélő fizikai és kognitív terhelését detektálta [22]. Később egy olyan módszeren dolgoztunk, amely azt próbálta azonosítani, hogy a beszélő megpróbálta-e megtéveszteni a hallgatóságot [23]. Tavaly, a Cold nevű versenyfeladatra kidolgozott módszerünkkel megnyertük a versenyt, a módszerünk feladata annak felismerése volt, hogy a beszélő meg van-e fázva [24].

6. Konklúzió és jövőbeli kutatási irányok

A dolgozatban bemutattuk, hogy a standard HMM/GMM rendszerhez kidolgozott módszerek hogyan adaptálhatóak az új HMM/DNN hibrid modellhez. Ehhez kidolgoztunk új, tisztán DNN alapú módszereket a kezdeti tanítási fázis (flat start) és az állapotkapcsolási lépés megoldására. Ezek összekapcsolásával sikeresen létrehoztunk egy új tanítási módszert, amely során nincs szükség GMM-ek használatára. Végül megmutattuk, hogy a végső tanítási lépés javítható egy egyszerű újramintavételező algoritmussal. A kísérleteink során felhasznált magyar nyelvű Szeged Híradós korpuszon egy hagyományos HMM/GMM 20.07%-os szószintű hibaarányt képes elérni. Az új hibrid módszer esetében, ami még változatlanul használja a megörökölt algoritmusokat,

	[2]	[25]	[26]	[15]	[27]	[28]	[22]	[23]	[24]
I.	•								
II/1.		•							
II/2.		•	•						
III/1.				•					
III/2.			•	•	•				
IV.						•	•	•	•

4. táblázat: A tézispontok és a szerző publikációinak viszonya.

a szószintű hibaarány 16.59%-ra csökkent, míg a legjobb GMM-mentes módszerünk még ennél is jobb eredményt (15.79%) ért el.

Természetesen rengeteg további kísérletet lehetne még elvégezni, ezeket sajnos idő hiányában a jövőbeli munkáink közé soroljuk. A következőkben felsorolunk néhány lehetséges jövőbeli kutatási irányt.

- Az elmúlt pár évben megjelent egy új típusú neuronháló, a konvolúciós neuronháló (CNN), amely jelentős sikereket ért el képfeldolgozásban és beszédfelismerésben. A kidolgozott módszereinket célszerű lenne kipróbálni ilyen típusú hálókkal is.

- A 3. fejezet kibővítése céljából más szekvenciatanuló algoritmusokat, például a minimális fonéma hibát (MPE) és a minimális Bayes kockázat (sMBR) módszert is tervezzünk megvizsgálni.

- Érdekes kérdés, hogy vajon hogyan alakulna a GMM-mentes modelljeink pontossága, amennyiben a mostaninál több klaszter létrejöttét is engednénk. A hipotézisünk, hogy több kontextus-függő állapot esetén jobb eredményeket tudnának elérni a hálók, természetesen ennek az ára a megnövekedett tanítási és kiértékelési idők lennének.

- Szintén megérné megvizsgálni, hogy a mintavételezéssel tanított hálók hogyan viselkednének, egy végső szekvencia-diszkriminatív tanítási lépés végrehajtása után.

7. Az eredmények tézisszerű összefoglalása

Az alábbiakban tézispontokba rendezve összegezzük a szerző kutatási eredményeit. A 4. táblázat összegzi a kutatásokból származó publikációk és az egyes tézispontok viszonyát.

I. A szerző kísérleti úton összehasonlított négy mély tanulásos módszert: két előtanításos algoritmust, az egyenirányított aktivációs függvényt és a Dropout nevű regularizációs technikát. A kiértékeléseket egy magyar nyelvű adatbázison is elvégeztük, az itt közölt eredmények, legjobb tudomásunk szerint, a legelső mély neuronhálós eredmények magyar nyelvű beszédfelismerésben. Az eredmények alapján megállapíthatjuk, hogy a HMM/DNN hibrid szignifikánsan jobban teljesít mint a hagyományos HMM/GMM. A végső konklúziója a kísérleteknek az volt, hogy mind a négy módszer elég hasonló eredményeket tudott elérni, de az egyenirányított hálók konzisztensen jobbnak bizonyultak a többi módszernél.

II/1. A szerző megmutatta, hogy a CTC algoritmust, amit eredetileg rekurrens neuronhálók tanítására készítettek, fel lehet használni előrecsatolt hálók tanítására is. A kísérletek célja annak megállapítása volt, hogy ez a módszer alkalmas-e a flat start tanítási lépés elvégzésére, ezért mély egyenirányított neuronhálók lettek tanítva CTC algoritmussal, különböző adatbázisokon. Az eredmények azt mutatták, hogy a CTC módszer alkalmas véletlenszerűen inicializált neuronhálók flat start tanítására időben illesztett címkék nélkül.

II/2. A CTC algoritmus versenytársaként megvizsgálásra került az MMI algoritmus is. A szerző több módosítást is javasolt, hogy ezt a módszert alkalmassá tegye a flat start tanításra. Az összehasonlítás során egyértelművé vált, hogy az MMI sokkal jobb megoldás mint a CTC algoritmus véletlenszerűen inicializált neuronhálók tanítására időben illesztett címkék nélkül.

III/1. A szerző kidolgozott egy új, mély neuronhálós állapotkapcsolási algoritmust, a standard algoritmus döntési kritériumának lecserélésével. Tekintve, hogy a módszer bemenetként DNN által predikált posterior valószínűségi vektorokat kap, ezért döntési kritériumnak a KL-divergencia tűnt logikus választásnak. Ezt a kísérleti eredmények is alátámasztották, az új algoritmus lényegesen jobban teljesített, mint az eredeti módszer.

III/2. Az MMI-alapú flat start módszer és a KL-divergenciát alkalmazó állapot klaszterezési algoritmus kombinálásával a szerző egy teljesen GMM-mentes eljárást hozott létre. Ezt az új eljárást más, közelmúltban javasolt módszerrel hasonlította össze. A kísérletek során kiderült, hogy az új GMM-mentes módszerek jobb eredményeket képesek elérni mint azok, amelyek felhasználnak GMM-eket tanításuk során.

IV. A szerző megvizsgálta a valószínűségi mintavételező algoritmust és alkalmazta azt kontextusfüggő DNN tanításra. A hipotézise az volt, hogy a tanítóadat újramintavételezésével a prior valószínűségek újrabecslése szükségessé válik. Kísérleti úton igazolta ezt a sejtést és megmutatta, hogy újramintavételezéssel és a

priorok helyes beállításával szignifikánsan javítható a mély hálók pontossága. A mintavételező algoritmust paralingvisztikus feladatokon is sikeresen alkalmazta.

Hivatkozások

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, *et al.*, „Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] T. Grósz and L. Tóth, „A comparison of Deep Neural Network training methods for Large Vocabulary Speech Recognition," in *Proceedings of TSD*, pp. 36–43, 2013.

[3] G. E. Hinton, S. Osindero, and Y.-W. Teh, „A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[4] F. Seide, G. Li, X. Chen, and D. Yu, „Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proceedings of ASRU*, pp. 24–29, 2011.

[5] X. Glorot, A. Bordes, and Y. Bengio, „Deep sparse rectifier networks," in *Proceedings of AISTATS*, pp. 315–323, 2011.

[6] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, „Improving neural networks by preventing co-adaptation of feature detectors," in *CoRR*, vol. 1207.0580, 2012.

[7] S. S. Lamel L., Kassel R., „Speech database development: Design and analysis of the acoustic-phonetic corpus," in *DARPA Speech Recognition Workshop*, pp. 121–124, 1986.

[8] L. Tóth, B. Tarján, G. Sárosi, and P. Mihajlik, „Speech recognition experiments with audiobooks," *Acta Cybernetica*, pp. 695–713, 2010.

[9] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385 of *Studies in Computational Intelligence*. Springer, 2012.

[10] X. He and L. Deng, *Discriminative Learning for Speech Recognition*. San Rafael, CA, USA: Morgan & Claypool, 2008.

[11] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, „GMM-free DNN training," in *Proceedings of ICASSP*, 2014.

[12] C. Zhang and P. Woodland, „Standalone training of context-dependent Deep Neural Network acoustic models," in *Proceedings of ICASSP*, pp. 5597–5601, 2014.

[13] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, „GMM-free DNN acoustic model training," in *Proceedings of ICASSP*, pp. 5639–5643, 2014.

[14] M. Bacchiani and D. Rybach, „Context dependent state tying for speech recognition using deep neural network acoustic models," in *Proceedings of ICASSP*, pp. 230–234, 2014.

[15] G. Gosztolya, T. Grósz, L. Tóth, and D. Imseng, „Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying," in *Proceedings of ICASSP*, pp. 4570–4574, 2015.

[16] L. Zhu, K. Kilgour, S. Stüker, and A. Waibel, „Gaussian free cluster tree construction using Deep Neural Network," in *Proceedings of Interspeech*, pp. 3254–3258, Sep 2015.

[17] D. B. Paul and J. M. Baker, „The design for the Wall Street Journal-based CSR corpus," in *Proceedings of HLT*, pp. 357–362, 1992.

[18] L. Tóth and A. Kocsor, „Training HMM/ANN hybrid speech recognizers by probabilistic sampling," in *Proceedings of ICANN*, pp. 597–603, 2005.

[19] M. Song, Q. Zhang, J. Pan, and Y. Yan, „Improving HMM/DNN in asr of under-resourced languages using probabilistic sampling," in *Proceedings of ChinaSIP*, pp. 20–24, 2015.

[20] A. Rousseau, P. Deléglise, and Y. Estéve, „TED-LIUM: an automatic speech recognition dedicated corpus," in *Proceedings of LREC*, pp. 125–129, 2012.

[21] J. Carletta, „Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[22] G. Gosztolya, T. Grósz, R. Busa-Fekete, and L. Tóth, „Detecting the intensity of cognitive and physical load using AdaBoost and Deep Rectifier Neural Networks," in *Proceedings of Interspeech*, pp. 452–456, 2014.

[23] G. Gosztolya, T. Grósz, R. Busa-Fekete, and L. Tóth, „Determining native language and deception using phonetic features and classifier combination," in *Proceedings of Interspeech*, (San Francisco, CA, USA), pp. 2418–2422, Sep 2016.

[24] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, „DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," in *Proceedings of Interspeech*, pp. 3522–3526, 2017.

[25] T. Grósz, G. Gosztolya, and L. Tóth, „A sequence training method for Deep Rectifier Neural Networks in speech recognition.," in *Proceedings of SPECOM*, pp. 81–88, Sep 2014.

[26] G. Gosztolya, T. Grósz, and L. Tóth, „GMM-free flat start sequence-discriminative DNN training," in *Proceedings of Interspeech*, (San Francisco, CA, USA), pp. 3409–3413, Sep 2016.

[27] T. Grósz, G. Gosztolya, and L. Tóth, „A comparative evaluation of GMM-free state tying methods for ASR," in *Proceedings of Interspeech*, pp. 1626–1630, 2017.

[28] T. Grósz, G. Gosztolya, and L. Tóth, „Training context-dependent DNN acoustic models using probabilistic sampling," in *Proceedings of Interspeech*, pp. 1621–1625, 2017.