Mély neuronhálós akusztikus modellek gyors adaptációja multi-taszk tanítással

Tóth László, Gosztolya Gábor*

MTA-SZTE Mesterséges Intelligencia Kutatócsoport
e-mail: {tothl, ggabor}@inf.u-szeged.hu

Kivonat A környezetfüggő mély neuronhálós akusztikus modellek gyors adaptációja különösen nehéz kihívás, mivel egy kis méretű adaptációs mintában a környezetfüggő állapotok többségére nincs tanítópélda. Nemrégiben egy olyan új mély neuronhálós tanítási séma bukkant fel, amely a hálózatot egyszerre tanítja környezetfüggő és környezetfüggetlen példákon. Ez az ún. multi-taszk technológia felveti annak a nagyon egyszerű adaptációs módszernek a lehetőségét, hogy az adaptáció során csak környezetfüggetlen címkéken tanítsunk. Jelen cikkben ezt a módszert próbáljuk ki, kombinálva egy KL-divergencia alapú regularizációs technikával. Kísérleteinkben a multi-taszk tanítási séma már önmagában 3%-os hibacsökkenést hoz egy híradós beszédfelismerési feladaton. A kombinált adaptációs módszert is bevetve további 2-5% hibaredukciót sikerült elérnünk az adaptációs minta méretének függvényében, ami 20-tól 100 másodpercig terjedt.

Kulcsszavak: mély neuronháló, akusztikus modellezés, beszédfelismerés, adaptáció

1. Bevezetés

Az utóbbi években a rejtett Markov-modellek (hidden Markov model, HMM) hagyományos Gauss-keverékmodellje (Gaussian mixture model, GMM) helyett egyre inkább a mély neuronhálókat (deep neural network, DNN) kezdik alkalmazni. Az évtizedek alatt azonban a GMM-alapú modellezésnek számos olyan finomítását találták ki, amelyek nem vihetők át triviális módon a HMM/GMM rendszerekből a HMM/DNN rendszerekbe. Az egyik ilyen finomítás a környezetfüggő (context-dependent, CD) modellek készítése és betanítása. Jelen pillanatban a HMM/DNN rendszerek környezetfüggő állapotait ugyanazzal a jól bevált technológiával szokás előállítani, mint a HMM/GMM rendszerekben. Ez azt jelenti, hogy egy mély neuronhálós felismerő készítésének első lépéseként lényegében be kell tanítani egy hagyományos GMM-alapú felismerőt [3,7,12]. Habár születtek javaslatok arra nézve, hogy a GMM-eket hogyan lehetne kihagyni a folyamatból, ezek egyelőre inkább csak kísérleti próbálkozások [1,5,14,20]. Ami a mély neuronhálók környezetfüggő állapotokkal való betanítását illeti, Bell és

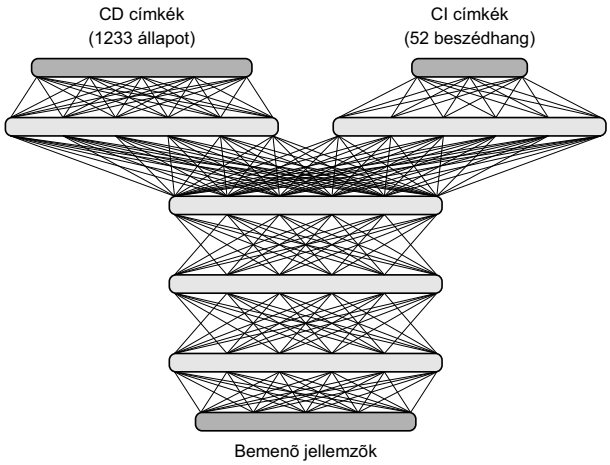
* A jelen kutatás során használt TITAN X grafikus kártyát az NVIDIA Corporation ajándékozta csoportunknak.

társai nemrégiben bemutattak egy új megoldást. Az ún. multi-taszk tanítás lényege, hogy a környezetfüggő címkékkel párhuzamosan környezetfüggetlen (context-independent, CI) címkékkel is tanítjuk a hálózatot [2]. Technikailag ezt úgy lehet megvalósítani, hogy a hálózatba két kimenő réteget veszünk fel, ahol egyikük a CD, másikuk pedig a CI címkék megtanulására törekszik [13]. A CI címkék párhuzamos tanítása egyfajta regularizációs hatást fejt ki a CD címkék tanulása során. Bell és tsai. módszerét mi is kipróbáljuk hamarosan, ami 3% hibacsökkenéshez fog vezetni a szószintű hibában.

A DNN akusztikus modellek adaptálása során a modell regularizációja kiemelt fontossággal bír. Mivel a mély neuronhálók jellemzően sok paraméterrel (réteg, ill. neuron) rendelkeznek, nagyon hajlamosak a túltanulásra, kiváltképp ha az adaptációs minta mérete kicsi. Talán a legelterjedtebb megoldás a túltanulás ellen, amikor a hálózatot kiegészítik egy lineáris réteggel, és az adaptáció során csak ezt a lineáris réteget engedik tanulni [4,16]. Hasonló megoldás a (túl)tanulás korlátozására, ha az adaptáció során csak a rétegek és/vagy súlyok csak egy kis részét engedjük tanulni [9,10]. Egy további megoldási lehetőség, ha csak a neuronok bias értékeit [17], vagy a rejtett neuronok aktivációs amplitúdóját [15] engedjük adaptálódni. A megoldások egy másik csoportja a túltanulás kockázatát valamilyen regularizációs megszorítás alkalmazásával csökkenti. Li és tsai. olyan L2-regularizáció alkalmazását javasolták, amely bünteti az adaptáció előtti és utáni hálózati súlyértékek nagy eltérését [8]. Gemello az ún. 'konzervatív tanítást' javasolta, melynek lényege, hogy az adaptációs mintában nem szereplő osztályokra az adaptálatlan hálózat kimeneteit használjuk a tanítás során célértékként [4]. Yu és tsai. egy olyan megoldást vetettek fel, amelyben a tanulási célértékek az adaptálatlan modell kimenete és az adaptációs minta címkéi közötti lineáris interpolációval állnak elő. Matematikailag ez a megoldás a Kullback-Leibler divergencia regularizációjaként formalizálható [18].

A környezetfüggő modellek használata jelentősen megnöveli a túltanulás kockázatát az adaptáció során, hiszen az állapotszám megnövelése lecsökkenti az egy állapotra eső tanítópéldák számát. Price és tsai. erre egy olyan hálózati struktúrát javasoltak, amelyben két kimeneti réteg épül egymásra, ahol az alsó a CD, a felső pedig a CI címkéknek felel meg [11]. Ezzel a megoldással betanítás és felismerés során a CD címkéket lehet használni, míg a CI kimeneti réteggel dolgozunk az adaptáció során, amikor kevés a címkézett tanítóadat.

Ebben a cikkben egy olyan megoldást javasolunk, amely alapötletében hasonlít Price és tsai. megoldásához, de az alkalmazott hálózati topológia teljesen más. Míg ők a CD és CI címkéknek megfelelő kimeneti rétegeket egymás fölé helyezték, mi egymás mellé rakjuk azokat, hasonlóan a multi-taszk tanítás során alkalmazott elrendezéshez. Ezzel a struktúrával az adaptáció módja triviálisan adódik: míg a (multi-taszk) betanítás során mind a CD, mind a CI kimeneti réteg kap mintákat, adaptáció során csak a CI kimenetet tanítjuk. Hogy tovább csökkentsük a túltanulás kockázatát, a tanítás során a Yu-féle KL-regularizációs technikát is alkalmazni fogjuk [18]. Kísérleteink azt mutatják, hogy ennek a regularizációnak kritikus szerepe van, főleg amikor az adaptációs mintahalmaz nagyon kicsi. A kombinált módszert egy felügyelet nélküli adaptációs feladaton



1. ábra. A multi-taszk neuronháló struktúrája.

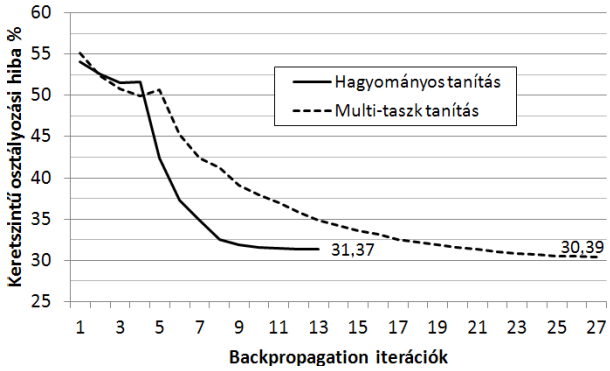
fogjuk kipróbálni, ahol az adaptációs minta mérete 20 és 100 másodperc között ingadozik. E hossz függvényében 2% és 5% közötti relatív hibacsökkenést sikerült elérni.

2. Multi-taszk tanítás

A multi-taszk tanítás lényege, hogy egy gépi tanuló algoritmusnak több megtanulandó feladatot adunk párhuzamosan, amitől jobb általánosítási képesség elérését reméljük. Tudomásunk szerint a multi-taszk tanítást DNN akusztikus modellek készítése során Seltzer és Droppo alkalmazták először. A TIMIT beszédhang-felismerési feladaton kísérletezve azt tapasztalták, hogy az aktuális adatvektor felismerési pontossága megnő, ha a hálózatnak másodlagos feladatként a fonetikai környezetet is meg kell tanulnia [13]. Bell és tsai kísérletében a CD címkéket tanuló neuronháló a CI címkék felismerését kapta másodlagos feladatként, ami 3%-10% relatív csökkentést hozott a szószintű hibában a hagyományos tanításhoz képest [2].

Az 1. ábra mutatja az általunk alkalmazott hálózati topológiát. Mint látható, a hálónak két kimeneti rétege van, egy a CD címkék és egy a CI címkék számára. A két kimeneti réteghez igazítva a legfölső rejtett réteget is kettéosztottuk, ami eltérést jelent Bell és mtsai. megoldásához képest [2]. Ezzel a struktúrával némileg jobb eredményeket kaptunk, habár a javulás nem volt szignifikáns.

Szintén a Bell-féle cikket követve a CI címkék esetén a beszédhangokat nem szedtük szét a szokásos 3 állapotra, azaz a CI címkék megfeleltek a monofón beszédhang-címkéknek. Tanítás előtt a CD állapotokat átkonvertáltuk a megfelelő monofón címkékre, és a tanítás során a hálózat mindkét fajta címkét megkapta. A tanítás folyamán minden egyes adatköteget (batch) véletlenszerűen a CD



2. ábra. A keretszintű CD hiba alakulása a validációs halmazon hagyományos és multi-taszk tanítás esetén.

vagy a CI kimeneti réteghez rendeltünk, és a hiba visszapropagálását csak a hálózat adott felén hajtottuk végre. A közös rétegek súlyait természetesen minden esetben frissítettük, míg a megosztott rejtett réteg és kimeneti réteg esetében csak az aktuális hozzárendelt hálózati ág súlyai tanultak. A következő szekcióban megmutatjuk, hogy ez a tanulási technika hogyan befolyásolja a modell konvergenciáját.

3. Multi-taszk tanítási kísérletek

A kísérletek során a „Szeged" magyar híradós beszédadatbázist használtuk [6]. A korpusz 28 órányi híradófelvételt tartalmaz nyolc tévécsatornáról. Az adatok tanító-tesztelő felosztása és a nyelvi modell ugyanaz volt, mint a korábbi munkáinkban [6]. A CD állapotok előállítására az ICASSP 2015 konferencián bemutatott módszerünket használtuk, ami 1233 trifón állapotot eredményezett [5]. A CI címkék száma 52 volt.

A kiindulási modellként alkalmazott mély neuronháló 4 rejtett réteget tartalmazott, rétegenként 2000 egyenirányított lineáris (rectified linear, ReLU) neuronnal. A multi-taszk tanítás céljaira a hálót az 1. ábrán látható módon alakítottuk át. A multi-taszk hálónak két kimenetei rétege volt, egy a CD és egy a CI címkék tanulásához, valamint a legfölső rejtett rétegnek is két változata volt, rendre 2000-2000 neuronnal. A tanítás a backpropagation algoritmussal történt, melynek során a szokványos keretenkénti keresztentrópia hibafüggvényt minimalizáltuk, az adatkötegeket a korábban leírt módon hol az egyik, hogy a másik kimeneti réteghez rendelve. A backpropagation előtt előtanítási módszert nem használtunk, mivel a korábbi eredmények azt mutatták, hogy ReLU neuronok használata esetén az előtanulásnak semmi vagy minimális haszna van csak [6,19].

A kísérletezés során megpróbáltuk belőni a CD, illetve a CI ágra irányított adatcsomagok optimális arányát. Míg Bell és tsai. az 50-50%-ot találták optimálisnak, esetünkben a 75%-25% arány (a CD kimenet javára) kicsit alacsonyabb

1. táblázat. A keretszintű (FER) és a szószintű (WER) hiba alakulása a tanító, validáló és teszthalmazokon.

Tanítási mód	FER %		WER %	
	Tanító h.	Val. h.	Val. h.	Teszth.
Hagyományos	25.9%	31.4%	17.7%	17.0%
Multi-taszk	23.5%	30.4%	17.4%	16.5%

hibaarányt adott az adatkeretek szintjén, habár ez a szószintű hibaarányt nem befolyásolta számottevően.

Ha a hálózatnak két dolgot kell egyidejűleg tanulnia, az értelemszerűen megnehezíti a tanulás konvergenciáját. Esetünkben a tanítás a backpropagation algoritmussal történt, ahol a tanulási rátát exponenciálisan felezgettük. Azt tapasztaltuk, hogy a multi-taszk tanítás során nem lehet olyan gyors léptékben csökkenteni a tanulási rátát, mint a hagyományos tanulás esetén: a 0,5-es szorzó helyett 0,8-del kaptuk a legjobb eredményt. Ugyanazt a megállási feltételt használva a multi-taszk tanulásnak körülbelül kétszer annyi iterációra volt szüksége a konvergenciához. A 2. ábrán egy példát láthatunk arra, hogy a tanítás során hogyan csökken a CD kimeneten számolt hiba a hagyományos és a multi-taszk tanítás esetén.

A 1. táblázatban összehasonlíthatjuk a kétféle tanítási móddal kapott végső hibaarányokat. Mint láthatjuk, a multi-taszk tanítással kb. 3% szószintű hibaarány-csökkenést sikerült elérni, ami nagyságrendileg megegyezik Bell és tsai. eredményeivel [2]. Azonban, míg ők azt találták, hogy a kisebb szószintű hiba ellenére a keretszintű CD hiba *nőtt*, mi ilyen ellentmondást nem tapasztaltunk. Ennek oka az lehet, hogy mi nagyobb arányban mutattunk CD példákat a hálónak, így a tanulás során nagyobb hangsúlyt kapott a CD kimeneten mért hiba.

4. Akusztikus adaptáció a multi-taszk modellel

Az állapotkapcsolt környezetfüggő modelleket előállító algoritmus zsenialitása abban rejlik, hogy a CD állapotok számát hozzá tudjuk igazítani a rendelkezésre álló tanító adatok mennyiségéhez. A gyakorlatban mindig arra törekszünk, hogy a CD állapotok számát olyan nagyra válasszuk, amennyi állapotot még biztonságosan be tudunk tanítani a túltanulás veszélye nélkül. Ha azonban a betanított modelljeinket egy új környezethez vagy beszélőhöz kell adaptálni, akkor az adaptációs tanításhoz rendelkezésünkre álló példák száma rendszerint nagyságrendekkel kisebb a teljes tanítóhalmaznál. Emiatt a CD modellek adaptációs mintára való tanítása szükségszerűen magában hordozza a túltanulás veszélyét. Azonban a multi-taszk modell egy kézenfekvő megoldást kínál a túltanulás esélyének csökkentésére: az adaptáció során a modell CD ágának nem mutatunk példákat, mivel a legtöbb CD címkére úgysem lenne példa a kis méretű adaptá-

ciós halmazban. Ehelyett az adaptáció során csakis a CI ágnak adunk példákat, amely ágat az adathiány problémája jóval kevésbé sújtja.

A mély neuronhálós akusztikus modelleknek rengeteg paraméterük (azaz súlyuk) van, ami nagyfokú rugalmasságot biztosít nagy tanítóhalmaz esetén, viszont növeli a túltanulás kockázatát egy kicsi adaptációs halmazon. Erre a legegyszerűbb megoldás, ha az adaptáció során csak a paraméterek egy részét – például egyetlen rejtett réteget – engedünk tanulni [10]. Ezzel ráadásul az adaptáció időigényét is csökkentjük. Mi a tanulást úgy korlátoztuk az adaptáció során, hogy az csak a CD és CI ágak legfelső közös rejtett rétegének súlyait frissítse (l. 1. ábra). Ezzel a megszorítással együtt is azt tapasztaltuk, hogy felügyelet nélkül adaptáció esetén nagyon nehéz megtalálni az optimális tanulási rátát. Míg kis értékek mellett stabil, de szerény hibacsökkenést kaptunk a fájlok legtöbbjére, nagyobb értékek egyes fájlokra nagy javulást adtak, másokra pedig hatalmas romlást. Arra gyanakodtunk, hogy az adaptációt a hibásan becsült adaptációs címkék viszik félre, és ezért a Yu-féle regularizációs megoldás [18] kipróbálása mellett döntöttünk. A módszer lényege, hogy a tanulás során büntetjük, ha az adaptált modell kimenete nagyon eltérne az adaptálatlan modell kimenetétől. Mivel a neuronhálók kimenete diszkrét valószínűségi eloszlásként értelmezhető, az eltérés mérésére a Kullback-Leibler divergencia adódik természetes megoldásként. Némi levezetés után (l. [18]) azt kapjuk, hogy az adaptáció során az adaptációs mintán kapott becsült címkéket simítani kell az adaptálatlan modell kimenő valószínűségeivel. Formálisan, a tanulási célok előállításához az alábbi lineáris interpolációt kell alkalmaznunk:

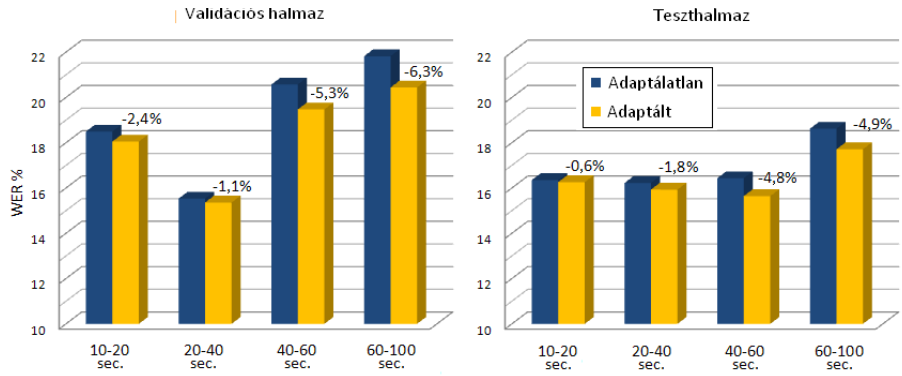
$$(1-\alpha)p(y|x) + \alpha p_{un}(y|x),$$

ahol $p(y|x)$ jelöli a 0-1 jellegű adaptációs címkézést (ez felügyelet nélküli esetben felismeréssel, felügyelt esetben kényszerített illesztéssel áll elő), $p_{un}(y|x)$ az adaptálatlan modell kimenete, az α paraméter segítségével pedig a simítás erősségét lehet állítani.

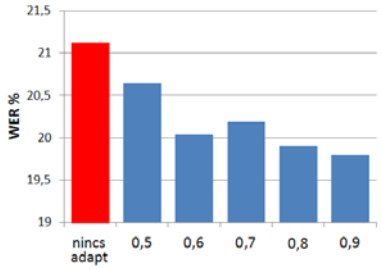
5. Kísérletek felügyelet nélküli adaptációval

A híradós tanítókorpuszunk validációs része 448 felvételt tartalmazott (kb. 2 óra összhosszban), míg a teszthalmaz 724 fájlból állt (4 óra összhosszal). Az egyes fájlok hossza egy mondat (pár másodperc) és kb. 100 másodperc között szórt. Az adaptációs kísérlet során a 10 másodpercnél rövidebb fájlokat nem használtuk. Azt biztosan lehetett tudni, hogy egy fájlon belül a beszélő személye és az akusztikai viszonyok nem változnak, tehát az adaptációnak van értelme, de ezen túl más, a beszélőre vonatkozó információ nem állt rendelkezésre. A kísérletek minden esetben felügyelet nélkül adaptációra törekedtek. Ez abból állt, hogy az adott fájlt felismertettük az adaptálatlan modellel, és a kapott becsült szöveges átiratot használtuk a modell adaptációs tanítására. Ezután a felismerést megismételtük, ezúttal már a fájlhoz adaptált modellel.

Az adaptációnak több paramétere volt, amelyeket a validációs halmazon kellett belőnünk. Ilyen paraméter volt a tanulási ráta, a tanulási iterációk száma, valamint a KL-regularizációs módszer α paramétere. A kezdeti, KL-regularizációt



3. ábra. A szószintű hiba (WER) csökkenése az adaptáció során az adaptációs minta méretének függvényében.



4. ábra. A KL-divergencián alapuló regularizációs módszer α paraméterének hatása a szószintű hibára (WER).

nem alkalmazó kísérleteinkben az optimális tanulási ráta fájlonként nagyon nagy eltéréseket mutatott, a regularizáció bevezetése után azonban az eredmények jóval stabilabbá váltak. A végső tesztekben öt tanítási iterációt mentünk minden fájlon, a normál tanítási rátához hasonló nagyságrendű rátával indulva.

A 3. ábra mutatja a szószintű hiba alakulását adaptáció előtt és után. A fájlokat a hosszuk függvényében négy csoportra osztottuk. Mint látható, a teszthalmazon 10 és 20 másodperc közti hossz esetén a hiba csak minimálisan csökkent, és még a 20-40 másodperc közti hossztartományban is csupán 2%-ot esett. Azonban a 40 másodpercnél hosszabb fájlok esetén a relatív hibacsökkenés 5-6%-ra ment fel a validációs halmazon és 5%-ra a teszthalmazon. Sajnálatos módon adatbázisunk nem tartalmazott 100 másodpercnél hosszabb fájlokat, így algoritmusunk tesztelését nem tudtuk hosszabb fájlokra is kiterjeszteni.

A 4. ábra érzékelteti a KL-regularizációs módszer hozzájárulását a jó eredményekhez. A kiértékelést a validációs készlet 40 másodpercnél hosszabb fájljain végeztük. Az ábrából nyilvánvalóan látszik, hogy a regularizációnak kulcsszerepe volt az adaptáció hatékonyságában. A legjobb eredményt mindig elég erős regu-

larizációval, $0,8 - 0,9$ közötti α értékekkel kaptuk, még a leghosszabb fájlméret (60-100 mp.) esetén is.

6. Konklúzió

A DNN akusztikus modellek adaptációja jelenleg nagyon aktív kutatási terület. A környezetfüggő DNN modellek adaptálása az adatelégtelenségi probléma miatt különösen nagy kihívás. A nemrégiben javasolt multi-taszk tanítási modell környezetfüggetlen címkéken is tanít, így kézenfekvő megoldást kínál az adaptációra. Kísérleteinkben azt tapasztaltuk, hogy mindemellett a Yu-féle regularizációs trükköt is be kellett vetnünk ahhoz, hogy stabilan viselkedő adaptációs eljárást kapjunk. Ezzel a megoldással egy híradós felismerési feladaton 3% relatív szóhiba-csökkenést értünk el csupán a multi-taszk tanítással, majd további 2%-5% hibacsökkenést az általunk javasolt adaptációs technikával.

Hivatkozások

1. Bacchiani, M., Rybach, D.: Context dependent state tying for speech recognition using deep neural network acoustic models. In: Proc. of ICASSP. pp. 230–234 (2014)
2. Bell, P., Renals, S.: Regularization of deep neural networks with context-independent multi-task training. In: Proc. ICASSP. pp. 4290–4294 (2015)
3. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. ASLP 20(1), 30–42 (2012)
4. Gemello, R., Mana, F., Scanzio, S., Laface, P., de Mori, R.: Linear hidden transformations for adaptation of hybrid ANN/HMM models. Speech Communication 49(10-11), 827–835 (2007)
5. Gosztolya, G., Grósz, T., Tóth, L., Imseng, D.: Building context-dependent DNN acoustic models using Kullback-Leibler divergence-based state tying. In: Proc. ICASSP. pp. 4570 – 4574 (2015)
6. Grósz, T., Tóth, L.: A comparison of deep neural network training methods for large vocabulary speech recognition. In: Proc. TSD. pp. 36–43 (2013)
7. Jaitly, N., Nguyen, P., Senior, A., Vanhoucke, V.: Application of pretrained deep neural networks to large vocabulary speech recognition. In: Proc. Interspeech (2012)
8. Li, X., Bilmes, J.: Regularized adaptation of discriminative classifiers. In: Proc. of ICASSP. Toulouse, France (2006)
9. Liao, H.: Speaker adaptation of context dependent Deep Neural Networks. In: Proc. of ICASSP. pp. 7947–7951. Vancouver, Canada (2013)
10. Ochiai, T., Matsuda, S., Lu, X., Hori, C., Katagiri, S.: Speaker adaptive training using deep neural networks. In: Proc. ICASSP. pp. 6399–6403 (2014)
11. Price, R., Iso, K., Shinoda, K.: Speaker adaptation of deep neural networks using a hierarchy of output layers. In: Proc. SLT. pp. 153–158 (2014)
12. Seide, F., Li, G., Chen, L., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. ASRU. pp. 24–29 (2011)
13. Seltzer, M., Droppo, J.: Multi-task learning in deep neural networks for improved phoneme recognition. In: Proc. ICASSP. pp. 6965–6969 (2013)

14. Senior, A., Heigold, G., Bacchiani, M., Liao, H.: GMM-free DNN training. In: Proc. of ICASSP. pp. 307–312 (2014)
15. Swietojanski, P., Renals, S.: Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In: Proc. SLT. 171-176 (2014)
16. Trmal, J., Zelinka, J., Müller, L.: Adaptation of feedforward artificial neural network using a linear transform. In: Proc. TSD. pp. 423–430 (2010)
17. Yao, K., Yu, D., Seide, F., Su, H., Deng, L., Gong, Y.: Adaptation of context-dependent Deep Neural Networks for Automatic Speech Recognition. In: Proc. of SLT. pp. 366–369. Miami, Florida, USA (2012)
18. Yu, D., Yao, K., Su, H., Li, G., Seide, F.: KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In: Proc. ICASSP. pp. 7893–7897 (2013)
19. Zeiler, M.D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q.V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., Hinton, G.E.: On rectified linear units for speech processing. In: Proc. ICASSP. pp. 3517–3521 (2013)
20. Zhang, C., Woodland, P.: Standalone training of context-dependent Deep Neural Network acoustic models. In: Proc. of ICASSP. pp. 5597–5601 (2014)