Mély neuronhálós beszédfelismerők GMM-mentes tanítása

Grósz Tamás[1], Gosztolya Gábor[1,2], Tóth László[2]

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
e-mail: { groszt, ggabor, tothl } @ inf.u-szeged.hu

Kivonat Az utóbbi pár évben a beszédfelismerőkben használt rejtett Markov modellekben (hidden Markov model, HMM) az ún. Gauss-keverékmodell (gaussian mixture model, GMM) komponenst leváltották a mély neuronhálók (deep neural network, DNN). Ugyanakkor ezek az új, neuronálókra épülő hibrid HMM/DNN felismerők számos olyan algoritmust megörököltek, melyeket eredetileg GMM-alapú rendszerekhez fejlesztettek ki, és így optimalitásuk az új környezetben nem garantált. A HMM/DNN modellek 'GMM-mentes' tanításához két részfeladatra kell új megoldást adnunk. Az egyik, hogy a mély hálók időben illesztett tanítócímkéket igényelnek, a másik pedig a környezetfüggő állapotok előállítása, amelyre a klasszikus megoldás egy GMM-alapú klaszterezési algoritmus. Bár a HMM/DNN hibridek tanítására léteznek teljes mondatokon dolgozó ún. szekvencia-diszkriminatív tanítóalgoritmusok, ezeket jellemzően csak a tanítás legutolsó fázisában, a modellek finomhangolására szokták bevetni, míg a tanítás elején HMM/GMM modellekel előállított és illesztett címkékből indulnak ki. Jelen cikkünkben viszont megmutatjuk, hogy megfelelő odafigyeléssel a szekvenciatanuló algoritmusok a tanítás legelejétől használhatóak. Az állapotklaszterezési lépésre korábban már javasoltunk egy GMM-mentes megoldást, így a címkeillesztési feladat megoldásával egy teljesen GMM-mentes tanítási sémához jutottunk. Kísérleti eredményeink azt mutatják, hogy a javasolt megoldás nemcsak gyorsabb, mint a hagyományos tanítási módszer, hanem valamivel jobb felismerési pontosságot is eredményez.

Kulcsszavak: mély neurális hálók, szekvencia-diszkriminatív tanítás

1. Bevezetés

A beszédfelismerésben a mély neuronhálók (deep neural network, DNN) áttörésével a hagyományos, Gauss-keverékmodelleken (gaussian mixture model, GMM) alapuló rejtett Markov-modellek (hidden Markov model, HMM) helyett most már az ún. HMM/DNN hibridek számítanak a csúcstechnológiának. Ezen modellek betanítása azonban jelenleg még több ponton is a hagyományos HMM/GMM modellhez kidolgozott tanítási algoritmusokon alapul. Jelenleg a neuronhálós

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

HMM/DNN modell tanítását egy hagyományos HMM/GMM rendszer betanításával kell kezdeni. Ebből a rendszerből nyerjük ki azután azokat a keretszinten illesztett, környezetfüggő állapotcímkéket, amelyek a DNN betanítása során tanítási célként szolgálnak. Ez az eljárás egyrészt erőforrás-pazarló (a HMM/GMM rendszert a tanítócímkék kinyerése után eldobjuk), másrészt semmi sem garantálja, hogy a GMM használatával kialakított és illesztett címkék a DNN számára is optimálisak lesznek. A két feladat – az állapotcímkék időbeli illesztése és környezetfüggő címkékké való konvertálása – közül az utóbbira korábban már adtunk egy GMM-mentes megoldást [1], így ebben a cikkben a másik problémára, azaz az állapotcímkék kezdeti időbeli illesztésére koncentrálunk.

A HMM/DNN modellek DNN komponensének betanítása legegyszerűbben úgy történhet, ha rendelkezésre állnak időben illesztett tanítócímkék, ekkor ugyanis a tanulás során használhatunk olyan klasszikus hibafüggvényeket, mint például a keresztentrópia (cross-entropy, CE). A legtöbbször azonban a tanítóadatokhoz csak mondatszintű átiratokat kapunk, a beszédhangok időbeli illesztése nem áll rendelkezésre. A HMM/GMM modelleknek megvan a tecnhológiája az időbeli illesztések előállítására, melyet gyakran 'flat start' tanításként emlegetnek [2]. Ez az összes beszédhang-modellt azonos paraméterekkel inicializálja, ami lényegében megfelel a hanghatárok időben egyenletes felosztásának. Innen kiindulva a HMM-ek klasszikus Baum-Welch tanítóalgoritmusa iteratívan tanítja és újrailleszti a modell címkéit. Hasonló, iteratív tanításon és újraillesztésen alapuló procedúrát természetesen ki lehet alakítani a DNN-tanításhoz is, akár a jól bevált CE-hibafügvényre építkezve. Senior és tsai. például véletlenszerűen inicializált neuronhálóval teszik ezt [3], míg Zhang és tsai. kiindulásként egyenletes beszédhang-szegmentálást alkalmaznak [4]. Ezek a megoldások működőképesek, de mint látni fogjuk, relatíve lassan konvergálnak, azaz sok tanítási-újraillesztési ciklust igényelnek.

A fenti eljárások megoldják ugyan a címkék illesztését, de továbbra is egy adatkeretek szintjén definiált hibafüggvényt használnak. Ez nem optimális, mivel a felismerés és a kiértékelés is mondatszinten történik. A HMM/GMM-ek körében számos mondatok szintjén definiált, más szóval szekvencia-diszkriminatív hibafüggvényt javasoltak, és ezek jó részét adaptálták is HMM/DNN hibridekre [5,6,7]. A legismertebb ilyen tanítási kritérium a kölcsönös információ maximalizálásán alapuló 'maximum mutual information', vagy röviden MMI-hibafüggvény [5]. A legtöbb szerző azonban a szekvencia-diszkriminatív tanítást csak a tanítási folyamat legvégén, a már betanított modellek finomhangolására alkalmazza. Magyarul, az első lépés mindig egy CE-hibafüggvényen alapuló tanítás (pl. [5,6,8,9,10,11]).

Az ún. 'neuronhálós időbeli osztályozás' (connectionist temporal classification, CTC) az utóbbi néhány évben vált népszerűvé DNN-ek sorozatokon való tanítására olyan esetben, amikor időben illesztett címkék nem állnak rendelkezésre [12]. Rao és tsai. javasoltak is egy 'flat start' tanítási eljárást, amely a CTC-n alapul [13]. A CTC technológiának azonban több hátránya is van az MMI-tanításhoz képest. Először is, a CTC a szokványos állapotcímkék mellett üres címkéket is elhelyez, amelyekkel aztán valamit kezdeni kell később, a környe-

zetfüggő állapotok kialakítása során. Másodszor, a CTC maga nem szekvencia-diszkriminatív módszer, így a legjobb eredményeket akkor adja, ha ilyen hibafüggvényekkel kombinálva használják [12,13].

A korábbi szerzőkkel ellentétben mi egy olyan tanítási eljárásra teszünk javaslatot, amely a tanítás legelejétől kezdve szekvencia-diszkriminatív hibafüggvényt használ. Ehhez a szokványos alkalmazáshoz képest több apró módosításra lesz szükség, amelyeket részletesen bemutatunk. A kísérletek során az általunk javasolt megoldást a Zhang és tsai. cikke alapján megvalósított, CE-hibafüggvényen alapuló iteratív újratanítási-újraillesztési megoldással vetjük össze [4]. Eredményként azt kapjuk, hogy a mi megoldásunk gyorsabb, és az elért szószintű hibaarány is valamivel kisebb. Tanítási módszerünket kombináljuk a korábban javasolt állapotklaszterezési algoritmusunkkal [1], így a végeredményként kapott tanítási eljárás összes lépése mentes lesz a GMM-alapú technológiától.

2. HMM/DNN felismerők 'flat start' tanítása

A HMM/DNN felismerők tanítása előtt egy HMM/GMM rendszert szokás betanítani, és ezzel állíthatóak elő a DNN tanításához szükséges, időben illesztett állapotcímkék. A cikkben két olyan módszert fogunk összehasonlítani, amelyek GMM használata nélkül képesek ugyanezt a feladatot elvégezni. Összehasonlítási alapként egy olyan algoritmus fog szolgálni, amely iteratívan ismétlődő tanítási-újraillesztési ciklusokat végez a HMM/DNN modellel, melynek DNN komponensét hagyományos, keretalapú CE-hibafüggvénnyel tanítja. Saját megoldási javaslatunk ezzel szemben a DNN tanítására szekvencia-diszkriminatív hibafüggvényt fog használni, mégpedig a talán legismertebb ilyet, a korábban már említett MMI-hibafüggvényt [5]. Az MMI-hiba 'flat start' tanításra való használata több apró módosítást fog igényelni, ezeket a 3. fejezetben be fogjuk mutatni.

2.1. Iteratív CE-tanítás és újraillesztés

Az összehasonlítási alapként szolgáló megoldás a CE tanulási kritériumot használja a DNN tanítására oly módon, hogy a címkéket időnként újraillészteni, majd a tanítást megismétli. Az algoritmus implementálása során Zhang és tsai. cikkét próbáltuk követni [4]:

1. A hangfájlokhoz a címkéket egyenletes időközökre bontással rendeljük hozzá, majd betanítjuk a DNN-t.
2. Az aktuális DNN-t használva újraillesztjük a címkéket a HMM/DNN modellel.
3. A régi DNN-t eldobva új hálót tanítunk az új címkehatárokkal.
4. A 2–3 lépéseket konvergenciáig ismételgetjük.

A fenti eljárás végén kapott DNN-t használjuk a címkék időbeli illesztésére, ez alapján a környezetfüggő modellek kialakítására, majd ezek segítségével a végleges DNN betanítására.

A fent ismertetett eljárás előnye, hogy a szokványos CE-hibafüggvény mellett nem igényli új hibafüggvény implementálását a tanításhoz, az újraillesztést pedig standard beszédfelismerési eszközökkel meg lehet oldani. A módszer hátránya, hogy az újratanítás-újraillesztés ismételgetése elég időigényes, amint majd azt a 6. fejezetben látni fogjuk.

2.2. Szekvencia-diszkriminatív tanítás az MMI-hibafüggvénnyel

A hagyományos HMM/GMM modellek szekvencia-diszkriminatív tanítása ma már sztenderdnek számít. Többféle hibafüggvényt is javasoltak e célra [14], és ezeket már a HMM/DNN modellekre is átültették [5,6,10,15]. A legrégebbi és legegyszerűbb ilyen hibakritérium a maximális kölcsönös információ (maximum mutual information, MMI) hibafüggvény. Az MMI függvény a jellemzővektor-sorozat és a hozzárendelt allapotsorozat kölcsönös információját méri. A jellemzővektorok sorozatára az $O_u = o_{u1}, \ldots, o_{uT_u}$, az u mondathoz tartozó szósorozatra pedig a W_u jelölést használva, az MMI-hibafüggvényt az alábbi módon formalizálhatjuk:

$$F_{MMI} = \sum_u \log \frac{p(O_u|S_u)^\alpha p(W_u)}{\sum_W p(O_u|S)^\alpha p(W)}, \tag{1}$$

ahol $S_u = s_{u1}, \ldots, s_{uT_u}$ a W_u-hoz tartozó állapotsorozat, α pedig az akusztikus modell súlya. A nevezőben található összegzés az u mondatra felismerési kimenetként kapott legvalószínűbb beszédhang-sorozatokat tartalmazza – ezt úgy kaphatjuk meg, hogy egyetlen kimenet helyett ún. szóhálót (lattice) generáltatunk a felismerővel. Az (1) egyenletet deriválva a $\log p(o_{ut}|r)$ log-likelihood érték szerint r állapotban és t időpillanatban, azt kapjuk, hogy

$$\begin{aligned} \frac{\partial F_{MMI}}{\partial \log p(o_{ut}|r)} &= \alpha\delta_{r;s_{ut}} - \frac{\alpha \sum_{W:s_t=r} p(O_u|S)^\alpha p(W)}{\sum_W p(O_u|S)^\alpha p(W)} \\ &= \alpha\big(\delta_{r;s_{ut}} - \gamma_{ut}^{DEN}(r)\big), \end{aligned} \tag{2}$$

ahol $\gamma_{ut}^{DEN}(r)$ a t időpillanatban az r állapotban való tartózkodás valószínűsége a nevezőhöz tartozó felismerési szóhálón számolva – amit a HMM-ek szokványos 'előre-hátra' algoritmusával kaphatjuk meg –, a $\delta_{r;s_{ut}}$ pedig a Kronecker-delta függvény (ez adja meg a 0-1 jellegű tanítási célvektorokat).

3. Flat start tanítás az MMI-hibakritériummal

A szekvencia-diszkriminatív tanítási kritériumokat, így például az MMI hibafüggvényt mostanra már széles körben használják a HMM/DNN hibridek tanítására. Tapasztalatunk szerint azonban a tanítást minden szerző a CE-hibakritériummal kezdi el, és a szekvencia-diszkriminatív hibakritériumot csak a tanítás végső fázisában vetik be, pusztán a modellek finomhangolására használva

azt [6,10]. Ez esetben viszont a CE-tanítás miatt mindenképpen szükség van valamilyen módszerre az időillesztett tanítási célvektorok előállítására. Ezekkel a szerzőkkel szemben mi azt állítjuk, hogy az MMI célfüggvényt rögtön a tanítás elejétől kezdve lehet használni, így a CE-tanulás, illetve ezáltal az ehhez szükséges illesztett címkék előállítása kihagyható. A módszerünk működőképessége érdekében az alábbi apró változtatásokat kellett elvégeznünk.

Elsőként, a (2) egyenlet számlálójában a $\delta_{r;s_{ut}}$ értékek helyett a $\gamma_{ut}^{NUM}(r)$ értékeket fogjuk használni, amit az előre-hátra algoritmussal számolunk ki. Ennek előnye, hogy bináris értékek helyett 0-1 közötti valószínűségi értékekkel dolgozhatunk, így kihagyhatjuk a (szokásosan GMM-alapú) címkeillesztési lépést. Ezt a megoldási lehetőséget több tanulmányban is említik (pl. [6,15]), de egyedül Zhou és tsai. cikkében találtuk nyomát, hogy valaki meg is valósította [8]. Azonban a tanítási folyamatot ők is CE-tanítással indítják, azaz az általunk javasolt flat start MMI-tanítást nem próbálják ki.

Mivel a szekvencia-diszkriminatív tanítási kritériumot a kész rendszer finomítására szokták használni, az MMI-célfüggvényt a teljes felismerővel, azaz környezetfüggő beszédhang-modellek és szószintű nyelvi modell mellett számolják ki. A (2) egyenlet nevezőjének kiszámolása a teljes felismerési procedúra lefuttatását igényli, ami a teljes modell használata mellett nagyon lassú. Emiatt a számlálóhoz és nevezőhöz szükséges hálók leszámolását csak egyszer szokták elvégezni, méghozzá az MMI-tanítás elindítása előtt. Ezzel szemben mi a szekvencia-diszkriminatív tanulást szószintű helyett pusztán fonetikai szintű szótárral végezzük, ráadásul környezetfüggő helyett környezetfüggetlen beszédhang-modellekkel. E két változtatás nagyon gyors dekódolást tesz lehetővé, így a számlálót és nevezőt minden egyes mondat után újra tudjuk számolni. Ez a módosítás kulcsfontosságú az eljárásunk gyors konvergenciája szempontjából. A szószintű átiratok fonetikai átirattá konvertálására a HTK rendszerben javasolt technikát használtuk, azaz első körben a hangsorozatot az egyes szavak fonetikai átiratát behelyettesítve kapjuk meg, a szavak közé sehol sem rakunk csendet. Az esetleges kiejtésvariánsokat, illetve a szavak közti csendet néhány iteráció után illesztjük be, újraillesztést végezve a már relatíve elfogadható szinten betanult modellel [2].

További finomítás, hogy a fonetikai dekódolás során nem használjuk sem a hangok a priori valószínűségét, sem bigramot vagy egyéb, összetettebb nyelvi modellt, emiatt a (2) egyenletből az α tag is elhagyható. Emellett, a számítási igény további csökkentése érdekében a $\gamma_{ut}^{DEN}(r)$ érték közelítésére a hálózat összes útvonalának figyelembe vétele helyett csak a legvalószínűbb felismerési útvonalat használtuk fel (ezt a közelítést jelöli a $\hat{\gamma}_{ut}^{DEN}(r)$ formula).

Ezekkel a módosításokkal a célfüggvény gradiense az alábbi módon alakul:

$$\begin{aligned}\frac{\partial F_{MMI}}{\partial a_{ut}(s)} &= \sum_r \frac{\partial F_{MMI}}{\partial \log p(o_{ut}|r)} \frac{\partial \log p(o_{ut}|r)}{\partial a_{ut}(s)} \\ &= \gamma_{ut}^{NUM}(s) - \hat{\gamma}_{ut}^{DEN}(s),\end{aligned} \tag{3}$$

amit pedig már közvetlenül tudunk használni a DNN tanítása során. Neuronhálók tanításánál jól ismert technika, hogy a tanítóhalmaz egy kis részét félretesszük validálási célra. Ha az aktuális tanítási iteráció után a hiba növekedne,

(1) A keretek tanítási célértékét ($\gamma_{ut}^{NUM}(r)$-t) az előre-hátra algoritmussal határozzuk meg.
(2) Beszédhang-szintű átiratokkal és környezetfüggetlen beszédhang-modellekkel dolgozunk.
(3) Nem használunk a priori valószínűségeket, sem nyelvi modellt.
(4) $\gamma_{ut}^{DEN}(r)$ értékét a legvalószínűbb felismerési útvonal valószínűségével ($\hat{\gamma}_{ut}^{DEN}(r)$) 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.

1. táblázat. *A 'flat start MMI' tanításhoz javasolt módosításaink összegzése.*

akkor a súlyokat viszaállítjuk az iteráció előttire, és a tanítást innen folytatjuk egy kisebb tanulási rátával. Ez a módszer szekvencia-diszkriminatív tanítás esetén is természetes módon alkalmazható [5], sőt, úgy találtuk, hogy a flat-start tanítási módszerünk stabilitásában ennek a lépésnek nagyon fontos szerepe van, mivel segít elkerülni az elakadásokat.

Az MMI-kritérium használatához javasolt módosításainkat a 1. táblázatban összszegezzük. Az (1)-(4) módosítási javaslatok egyrészt gyorsítják a felismerési folyamatot, másrészt növelik annak hibákkal szembeni robusztusságát. A (2) pont kulcsfontosságú szerepű abban, hogy a szekvencia-diszkriminatív tanulást a tanulási folyamat elejétől, még a környezetfüggő modellek kialakítása előtt alkalmazni tudjuk. Végezetül, az (5) pont segít az elakadási problémák kikerülésében, feloldásában.

4. KL-divergencia alapú állapotkapcsolás

Amikor a flat start tanítás konvergált, azaz megkaptuk a környezetfüggetlen (context-independent, CI) modellek legjobb időbeli illesztését, következhet a környezetfüggő (context-dependent, CD) modellek kialakítása. Jelenleg erre a legelterjedtebb megoldás az ún. döntési fa-alapú állapotklaszterező algoritmus [16]. Ez az algoritmus összegyűjti az egyes beszédhang-állapotok összes, különböző kontextusokban előforduló példányát, majd minden egyes csomópontban kettéosztva ezt a halmazt, felépít egy döntési fát, bizonyos előredefiniált kérdéseket követve. A kettéosztáshoz Gauss-görbét illeszt az aktuális adatok eloszlására, majd az alapján a kérdés alapján osztja ketté a csomópontot, amelyik a legnagyobb növekedést eredményezi a Gauss-görbék illeszkedésében (likelihood-értékében). Habár ez az algoritmus remekül működik GMM-alapú akusztikus modellek esetén, megkérdőjelezhető, hogy a Gauss-görbék illeszkedése mennyire alkalmas a mély neuronhálókkal való megtanulhatóság mérésére.

A fentiek miatt javasoltunk egy olyan alternatív megoldást, amely Gauss-görbék illesztése helyett betaníthat egy segéd-neuronhálót, majd ennek kimeneti

értékei alapján végzi el a döntési fa felépítését. Mivel a neuronháló-kimenetek egy diszkrét valószínűségi eloszlásból vett mintáknak tekinthetők, ezen kimeneti vektorok összehasonlítására természetes módon adódik az ún. Kullback-Leibler (KL) divergencia. Így az állapotklaszterezési algoritmust vezérlő, Gauss-görbékre felírt távolságfüggvényt lecseréltük egy KL-divergencián alapuló döntési kritériumra, Imseng és társainak cikkét követve [17]. A döntési függvény lecserélésén túl a döntésifa-építési mechanizmus változatlan marad, így a korábbi implementációk könnyen módosíthatók. Ezzel a megoldással nemcsak eliminálták a Gauss-görbéket az állapotklaszterezési folyamatból, de még 4% relatív javulást is elértünk a szószintű hibában. Az algoritmus részleteit korábban már publikáltuk, lásd [1].

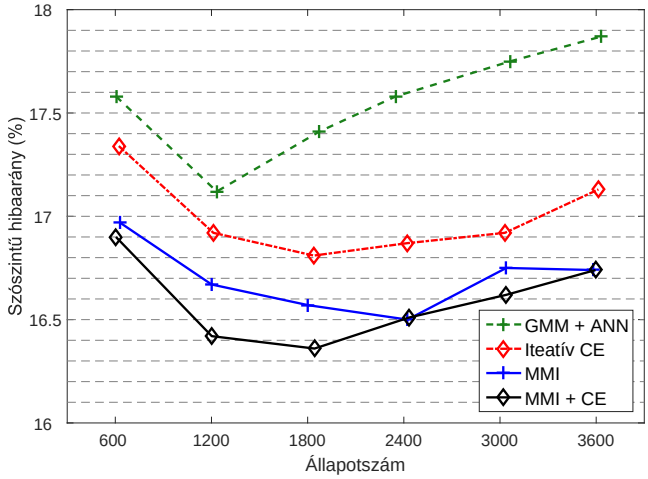
5. Kísérleti beállítások

Kísérleteink paraméterezése lényegében megegyezik a korábbi cikkeinkben leírtakkal [1]. Akusztikus modellként egy öt rejtett rétegű mély neuronhálót használtunk, melynek minden rétege 1000 'egyenirányított' (rectifier) neuront tartalmazott [18], míg a kimeneti rétegben softmax aktivációs függvényt alkalmaztunk. A modell saját neuronhálós csomagunkra épült, mellyel korábban kiemelkedő eredményeket értünk el több különböző feladaton is ([19,20,21,22]). Jellemzőkészletként egy 40-sávos mel-szűrőkészlet energiakimeneteit használtuk, a szokványos első és második derivált értékeivel kiegészítve. A felismerést és kiértékelést a HTK programcsomag mély hálókhoz igazított verziójával végeztük [2].

Beszédkorpuszként a 'Szeged' híradós beszédadatbázist használtuk, amely 28 órányi híradófelvételt tartalmaz nyolc tévécsatornáról rögzítve [23]. Tanítóhalmazként egy kb. 22 órányi részt különítettünk el, míg 2 órányi adatot használtunk validációs avagy fejlesztési (development) halmazként, 4 órányit pedig tesztelésre. Nyelvi modellként egy sztenderd trigram modell szolgált, a kiejtési szótár szűk ötszázezer szóalakot tartalmazott. Az állapotklaszterző algoritmus paramétereit úgy állítottuk be, hogy a különböző kísérletekben nagyjából 600, 1200, 1800, 2400, 3000, illetve 3600 kapcsolt állapotot kapjunk.

A beszédhang-modellek kezdeti illesztésére négyféle módszert próbáltunk ki és hasonlítottunk össze. Elsőként egy hagyományos, GMM-alapú rendszert tanítottunk be, és ezzel állítottuk elő az időben illesztett CI címkéket. Ezután az így kapott állapotcímkéken betanítottunk egy szimpla (azaz nem mély) neuronhálót a CE-kritériummal, és az így kapott hálóval újraillesztettük a címkéket (korábbi tanulmányunkban azt kaptuk, hogy szimpla helyett mély hálót használva nem javulnak az eredmények [1]). A táblázatokban erre a módszerre *„GMM + ANN"* jelöléssel fogunk hivatkozni. Az újraillesztés után a CD modellek előállítására mind a GMM-alapú, mind a KL-kritérium alapú megoldást kipróbáltuk, ahol az utóbbi esetben természetesen a neuronháló kimenete szolgált inputként.

Míg a fenti megoldás egy GMM-alapú rendszerből indult ki, 'GMM-mentes' megoldásként a 2. és 3. fejezetekben ismertetett algoritmusokat vetettük be. Ezekben a kísérletekben a neuronháló mindig mély háló volt, öt rejtett réteggel. Az iteratív CE-tanításon és újraillesztésen alapuló módszer esetében (a táblázat-



1. ábra. Szószintű hibaarány a KL-klaszterezéssel kapott állapotok számának függvényében, a fejlesztési halmazon.

ban *„Iteratív CE”*) négy tanítási-újraillesztési ciklust futtattunk, az ezt követő állapotklaszterezés során pedig a KL-divergencia alapú módszert hajtottuk végre a végső neuronháló által adott illesztésen. Az MMI-tanítás esetén (a táblázatban *„MMI”*) szintén véletlen súlyokkal inicializált mély hálóból indultunk ki, melyet a korábban ismertetett módon tanítottunk. A végeredményként előálló DNN szolgáltatta az inputot a rákövetkező, KL-divergencia alapú klaszterezési lépéshez. Végezetül, a negyedik kísérletben a szekvencia-diszkriminatív MMI-tanítással kapott illesztett címkéken lefuttattunk még egy CE-tanítást, és ennek kimenetén végeztük el a KL-kritérium alapú klaszterezést (*„MMI + CE”*). Tettük ezt azért, mert azt tapasztaltuk, hogy a CE, illetve az MMI kritérium eléggé eltérő valószínűségi eloszlásokat eredményez, ezért kíváncsiak voltunk, hogy vajon a klaszterezést ez hogyan befolyásolja.

Cikkünk fő célja a ‘flat-start’ lépés, azaz a kezdeti címkeillesztéseket előállító lépés különböző változatainak összehasonlítása volt. Ezért az állapotklaszterezés után előálló CD-modelleket már csak az egyszerűbb CE-kritériummal tanítottuk. Természetesen ezeket a modelleket tanítás után tovább lehetne finomítani a szekvencia-diszkriminatív tanítás bevetésével. Ezzel vélhetően kicsit jobb eredményeket kapnánk ugyan, de mivel ez egy sztenderd eljárás, ezért ettől jelen cikkben eltekintettünk.

6. Kísérleti eredmények

A különböző módszerekkel kapott szószintű hibaarányok alakulását a fejlesztési halmazon az 1. ábra mutatja, különböző állapotszámok esetére. Mint látható, a GMM-alapú módszer messze a legrosszabbul teljesített, míg az MMI-alapú

Flat start módszer	Állapotkapcsolási módszer	Szóhiba (%)		Iterációk száma
		Dev.	Teszt	
GMM + ANN	GMM	18.83%	17.27%	—
GMM + ANN	KL	17.12%	16.54%	—
Iteratív CE	KL	16.81%	16.50%	48
MMI		16.50%	15.96%	13
MMI + CE		16.36%	15.86%	29

2. táblázat. *Szószintű hibaarány a különféle 'flat start' illetve állapotkapcsolási stratégiák esetén.*

flat start eljárás minden esetben kissé jobb eredményeket adott, mint az iteratív megoldás. Habár az MMI-t követő CE tanítás (az 'MMI+CE'-vel jelölt modell) kisebb állapotszám mellett némileg jobb eredményeket adott, ez a javulás nem jelentős annyira, hogy megérje a többletidőt. Mindez azt mutatja, hogy a szekvencia-diszkriminatív tanítás egyaránt pontos időillesztéseket és jó valószínűségi becsléseket eredményez.

A 2. táblázat összesíti a különböző konfigurációkkal elért legjobb szóhibaarányokat a fejlesztési és teszthalmazokon. Az állapotklaszterezési módszerek közül a KL-divergencia alapú megoldás minden esetben egyértelműen túlszárnyalta a GMM-alapú módszert. Az illesztési technikákat összevetve azt láthatjuk, hogy a HMM/GMM rendszerre támaszkodó megoldás bizonyult a legrosszabbnak, amin a neuronhálós újraillesztés sem segített. Az iteratív CE-alapú tanítási módszer kicsivel rosszabb lett a két MMI alapú megoldásnál. E módszer esetén sajnos elég nehéz megmondani az optimális iterációszámot. Zhang és társai 20 lépésen át végezték az iterációt [4], míg mi csak 4 lépésig futtattuk. Emiatt érdemes a futási időket is összevetni, mely értékek a 2. táblázat jobb szélső oszlopában láthatók (a tanítási iterációk számát a „GMM + ANN" rendszer esetében nem tüntettük fel, mivel ott a tanítás egy radikálisan eltérő procedúrán alapult). Az iteratív CE-tanítás 4 iterációt igényelt, összesen 48 DNN-tanítási ciklust eredményezve, míg az MMI-tanítás ennek csak kb. a negyedét. Habár az utóbbihoz az előre-hátra algoritmus lefuttatásának költségét is hozzá kell adni, ezzel együtt is egyértelmű, hogy az MMI-tanítás műveletigénye jóval kisebb.

Ha a futási időt DNN-tanítási ciklusok helyett egyszerűen CPU/GPU időben mérjük, akkor még nagyobb különbségeket kapunk az MMI módszer javára (3 óra 16-tal szemben). Ennek oka, hogy a CE-tanítás során 100-as minibatch-méretet használtunk, míg az MMI-tanítás során a kötegméret az egyes felvételek méretével egyezett meg, ami átlagosan 1000 körüli batch-méretet, és így a GPU-k struktúrája miatt gyorsabb végrehajtást eredményezett.

Álláspontunk szerint módosításaink közül kettő kulcsfontosságú a javasolt algoritmusunk sebessége és futásideje szempontjából. Az első módosítás, hogy az illesztést környezetfüggetlen beszédhang-modellekkel, nyelvi modell nélkül végezzük. Ez teszi lehetővé a gyors számítást, és így a célfüggvényben található

szóhálók frissítését minden egyes mondat feldolgozása után. Az irodalomban egyetlen olyat cikket találtunk, amely nem csak a tanulási iterációk végén frissíti ezeket a hálókat, ebben a cikkben azonban egy masszívan párhuzamosított architektúrát írnak le, ami nagyon nehezen összevethető a mi szekvenciális algoritmusunkkal [24].

A stabilitást illetően közismert, hogy a szekvencia-diszkriminatív módszerek erősen hajlamosak a túltanulásra. Az állapotcímkék és azok illesztésének egyidejű tanulása gyakran vezet az ún. „run-away silence model" esetéhez, amikor a hosszú csendszakaszok miatt a csendhez tartozó kimenet egyre dominánsabbá válik, majd az illesztést is elrontva 'megeszi' a beszédhang-szakaszokat is [25]. A hasonló esetek elkerülésére egy független validációs halmazon mértük a neuronháló hibáját, és ha a hiba az aktuális iteráció után megugrott, akkor a korábbi súlyok visszaállítása után egy kisebb tanulási rátával újrapróbáltuk a tanulást. Tapasztalatunk szerint ez az egyszerű trükk sokat segített a hasonló elakadási jelenségek megakadályozásában.

7. Konklúzió

Cikkünkben megmutattuk, hogy a HMM/DNN modellek szekvencia-diszkriminatív tanítását a tanítás legelső, ún. 'flat start' fázisában is sikeresen lehet használni. E célra a szokványos MMI tanítási kritériumot alkalmaztuk, míg a tanítási folyamatban néhány apró módosítást vezettünk be. Kísérleti eredményeink azt mutatták, hogy – a CE tanítási kritériumon alapuló újratanítás-újraillesztés stratégiával összevetve – az általunk javasolt megoldás lényegesen gyorsabb, és még a szóhiba-arányt is csökkenti valamelyest. A korábban javasolt KL-divergencia alapú állapotklaszterezési megoldást is bevonva, összességében egy olyan HMM/DNN tanítási algoritmust adtunk, amely egyáltalán nem igényli a hagyományos HMM/GMM modellek használatát.

Hivatkozások

1. Gosztolya, G., Grósz, T., Tóth, L., Imseng, D.: Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying. In: Proceedings of ICASSP. (2015) 4570–4574
2. 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)
3. Senior, A., Heigold, G., Bacchiani, M., Liao, H.: GMM-free DNN acoustic model training. In: Proceedings of ICASSP. (2014) 5639–5643
4. Zhang, C., Woodland, P.: Standalone training of context-dependent Deep Neural Network acoustic models. In: Proceedings of ICASSP. (2014) 5634–5638
5. Kingsbury, B.: Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In: Proceedings of ICASSP. (2009) 3761–3764
6. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: Proceedings of Interspeech. (2013) 2345–2349

7. Grósz, T., Gosztolya, G., Tóth, L.: A sequence training method for Deep Rectifier Neural Networks in speech recognition. In: Proceedings of SPECOM, Novi Sad, Serbia (2014) 81–88
8. Zhou, P., Dai, L., Jiang, H.: Sequence training of multiple Deep Neural Networks for better performance and faster training speed. In: Proceedings of ICASSP. (2014) 5664–5668
9. Saon, G., Soltau, H.: A comparison of two optimization techniques for sequence discriminative training of Deep Neural Networks. In: Proceedings of ICASSP. (2014) 5604–5608
10. Wiesler, S., Golik, P., Schüter, R., Ney, H.: Investigations on sequence training of neural networks. In: Proceedings of ICASSP. (2015) 4565–4569
11. Chen, D., Mak, B., Sivadas, S.: Joint sequence training of phone and grapheme acoustic model based on multi-task learning Deep Neural Networks. In: Proceedings of Interspeech. (2014) 1083–1087
12. Graves, A., Mohamed, A.R., Hinton, G.E.: Speech recognition with Deep Recurrent Neural Networks. In: Proceedings of ICASSP. (2013) 6645–6649
13. Rao, K., Senior, A., Sak, H.: Flat start training of CD-CTC-SMBR LSTM RNN acoustic models. In: Proceedings of ICASSP, Shanghai, China (2016) 5405–5409
14. He, X., Deng, L.: Discriminative Learning for Speech Recognition. Morgan & Claypool, San Rafael, CA, USA (2008)
15. Yu, D., Deng, L.: Chapter 8: Deep neural network sequence-discriminative training. In: Automatic Speech Recognition — A Deep Learning Approach. Springer (2014)
16. 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
17. Imseng, D., Dines, J.: Decision tree clustering for KL-HMM. Technical Report Idiap-Com-01-2012, IDIAP Research Institute (2012)
18. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier networks. In: Proceedings of AISTATS. (2011) 315–323
19. Tóth, L.: Convolutional deep maxout networks for phone recognition. In: Proceedings of Interspeech. (2014) 1078–1082
20. 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: Proceedings of Interspeech. (2015) 1339–1343
21. Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M., Kálmán, J.: Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: Proceedings of Interspeech, Dresden, Germany (2015) 2694–2698
22. Kovács, Gy., Tóth, L.: Joint optimization of spectro-temporal features and Deep Neural Nets for robust automatic speech recognition. Acta Cybernetica **22** (2015) 117–134
23. Grósz, T., Tóth, L.: A comparison of Deep Neural Network training methods for Large Vocabulary Speech Recognition. In: Proceedings of TSD, Pilsen, Czech Republic (2013) 36–43
24. Bacchiani, M., Senior, A., Heigold, G.: Asynchronous, online, GMM-free training of a context dependent acoustic model for speech recognition. In: Proceedings of Interspeech, Singapore, Singapore (2014) 1900–1904
25. Su, H., Li, G., Yu, D., Seide, F.: Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription. In: Proceedings of ICASSP. (2013) 6664–6668