Mély neuronhálók az akusztikus modellezésben

Grósz Tamás, Tóth László*

MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
e-mail: groszt@sol.cc.u-szeged.hu,tothl@inf.u-szeged.hu

Kivonat A beszédfelismerők akusztikus modelljeként az utóbbi években jelentek meg, és egyre nagyobb népszerűségnek örvendenek az ún. mély neuronhálók. Nevüket onnan kapták, hogy a korábban szokványos egyetlen rejtett réteg helyett jóval többet, 3-9 réteget használnak. Emiatt – bár a hagyományos módszerekkel is taníthatók – az igazán jó eredmények eléréséhez egy új tanítóalgoritmust is ki kellett hozzájuk találni. Cikkünkben röviden bemutatjuk a mély neuronhálók matematikai hátterét, majd a mély neuronhálókra épülő akusztikus modelleket beszédhang-felismerési teszteken értékeljük ki. Az eredményeket összevetjük a korábban publikált, hagyományos neuronhálót használó eredményeinkkel.

Kulcsszavak: mély neuronháló, akusztikus modellezés, beszédfelismerés

1. Bevezetés

Az elmúlt néhány évtizedben a mesterséges neuronhálók számos változatát kipróbálták a beszédfelismerésben - annak függvényében, hogy éppen mi volt az aktuálisan felkapott technológia. Általános elismertséget azonban csak a többrétegű perceptron-hálózatokra (MLP) épülő ún. hibrid HMM/ANN modellnek sikerült elérnie, főleg a Bourlard-Morgan páros munkásságának köszönhetően [1]. Bár kisebb felismerési feladatokon a neuronhálós modellek jobb eredményt adnak, mint a sztenderd rejtett Markov-modell (HMM), alkalmazásuk mégsem terjedt el általánosan, részben mivel technikailag nehézkesebb a használatuk, másrészt mivel nagyobb adatbázisokon az előnyük elvész, köszönhetően a HMM-ekhez kifejlesztett trifón modellezési és diszkriminatív tanítási technikáknak. Így a hibrid modell az elmúlt húsz évben megmaradt a versenyképes, de igazi áttörést nem hozó alternatíva státuszában.

Mindez megváltozni látszik azonban az ún. mély neuronhálók (deep neural nets) megjelenésével. A mély neuronhálót (pontosabban tanítási algoritmusát) 2006-ban publikálták először [2], és a kezdeti cikkek képi alakfelismerési teszteket használtak demonstrációként. Legjobb tudomásunk szerint a mély hálók első beszédfelismerési alkalmazása Mohamed 2009-es konferenciaanyaga volt (ennek

* Jelen kutatást a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett.

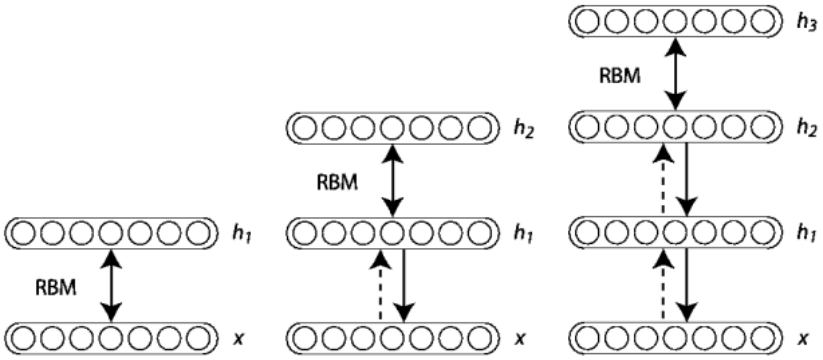
[3] az újságcikké kibővített változata) – mely cikkben rögtön sikerült megdönteni a népszerű TIMIT benchmark-adatbázison elért összes korábbi felismerési pontosságot. A modellt ráadásul hamarosan tovább javították [4]-ben. Ezek az eredmények annyira meggyőzőek voltak, hogy azóta exponenciálisan nő a témával foglalkozó cikkek száma - a legutóbbi, 2012. szeptemberi Interspeech konferencián már két szekció volt speciálisan csak a mély neuronhálóknak szentelve.

Cikkünkben először bemutatjuk a mély neuronhálók matematikai hátterét. Kitérünk a betanításuk során használt korlátos Boltzmann-gépekre, illetve a ,,kontrasztív divergencia" elnevezésű tanító algoritmusukra. A kísérleti alátámasztásra beszédhang-felismerési teszteket végzünk három adatbázison. Az angol nyelvű TIMIT-en megkíséreljük reprodukálni a [3]-ben közölt eredményeket, majd pedig két magyar nyelvű korpuszra – egy híradós adatbázis és egy hangoskönyv – terjesztjük ki a vizsgálatokat. Mindkét adatbázison közöltünk már eredményeket korábban, ezek fogják képezni a kiértékelés viszonyítási pontját.

2. Mély neuronhálók

Miben is különbözik ez az új neuronhálós technológia a megszokott többrétegű perceptronoktól? Egyrészt a hálózat struktúrájában, másrészt a tanító algoritmusban. A hagyományos hálózatok esetében egy vagy maximum két rejtett réteget szoktunk csak használni, és a neuronok számának növelésével próbáljuk a hálózat osztályozási pontosságát növelni. Emellett az az elméleti eredmény szól, miszerint egy kétrétegű hálózat már univerzális approximátor, azaz egy elég általános függvényosztályon tetszőleges pontosságú közelítésre képes [5]. Ehhez azonban a neuronok számát tetszőleges mértékben kell tudni növelni. Ehhez képest az újabb matematikai érvek és az empirikus kísérletek is amellett szólnak, hogy - *adott neuronszám mellett* - a több réteg hatékonyabb reprezentációt tesz lehetővé [6]. Ez indokolja tehát a sok, relatíve kisebb rejtett réteg alkalmazását egyetlen, rengeteg neuront tartalmazó réteg helyett.

Az ilyen sok rejtett réteges, „mély" architektúrának azonban nem triviális a betanítása. A hagyományos neuronhálók tanítására általában az ún. backpropagation algoritmust szokás használni, ami tulajdonképpen a legegyszerűbb, gradiensalapú optimalizálási algoritmus neuronhálókhoz igazított változata. Ez egy-két rejtett réteg esetén még jól működik, ennél nagyobb rétegszám mellett azonban egyre kevésbé hatékony. Ennek egyik oka, hogy egyre mélyebbre hatolva a gradiensek egyre kisebbek, egyre inkább "eltűnnek" (ún. „vanishing gradient" effektus), ezért az alsóbb rétegek nem fognak kellőképp tanulni [6]. Egy másik ok az ún. „explaining away" hatás, amely megnehezíti annak megtanulását, hogy melyik rejtett neuronnak mely jelenségekre kellene regálnia [2]. Ezen problémák kiküszöbölésére találták ki a korlátos Boltzmann-gépet (Restricted Boltzmann Machine, RBM), illetve annak tanító algoritmusát, a CD-algoritmust (kontrasztív divergencia) [2]. A korlátos Boltzmann-gép lényegében a neuronháló egy rétegpárjának felel meg, így a betanítás rétegenként haladva történik. A tanítás végén a rétegpárok egymásra helyezésével előálló többrétegű hálót "Deep Belief Network"-nek hívják az irodalomban [3]. Az elmondottakat szemlélteti a 1. ábra.



1. ábra. Korlátos Boltzmann-gép, illetve a belőle felépített DBN.

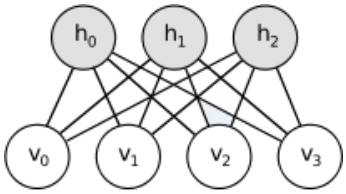
Fontos még tudni, hogy a CD-algoritmus felügyelet nélküli tanítást végez, és tulajdonképpen a „maximum likelihood" tanítás egy hatékony közelítését adja. Ezért a CD-algoritmus szerint tanítást tulajdonképpen előtanításnak tekintjük, mivel ezután következik még a címkézett tanítópéldákhoz való hozzáigazítás. E célból a hálózatot átalakítjuk korlátos Boltzmann-gépek helyett hagyományos neuronokat használó hálózattá, ráteszünk egy softmax-réteget, és ezután a megszokott backpropagation-algoritmussal végezzük a címkéken való felügyelt tanítást. A tanítás tehát két szakaszra oszlik: egyik az előtanítás, a másik pedig a hagyományos hálózatként való finomhangolás. Ha az előtanítást elhagyjuk, akkor egy teljesen hagyományos neuronhálót kapunk, így az előtanítási módszer hatékonyságának mérésére az a legjobb módszer, ha megnézzük, hogy mennyit javulnak a felismerési eredmények a használatával az előtanulást nem alkalmazó hálóhoz képest.

Az alábbi két fejezetben bemutatjuk a korlátos Boltzmann-gépeket, illetve a tanításukra szolgáló CD-algoritmust.

2.1. RBM

A korlátos Boltzmann-gép lényegében egy Markov véletlen mező (MRF), amely két rétegből áll. A korlátos jelző onnét származik, hogy két neuron csak akkor van összekapcsolva, ha az egyik a látható, a másik pedig a rejtett réteghez tartozik. Tehát a régeken belül a neuronok nem állnak kapcsolatban, ezért tekinthetünk az RBM-re úgy is, mint egy teljes páros gráf, ezt szemlélteti a 2. ábra. Az egyes kapcsolatokhoz tartozó súlyok és a neuronokhoz tartozó bias-ok egy véletlen eloszlást definiálnak a látható réteg neuronjainak állapotait tartalmazó v vektorok felett, egy energiafüggvény segítségével. Az energiafüggvény (v, h) együttes előfordulására:

$$E(v, h, \Theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j, \tag{1}$$



2. ábra. Egy RBM 4 látható és 3 rejtett neuronnal.

ahol $\Theta = (w, b, a)$, és w_{ij} reprezentálja az i. látható neuron és j. rejtett neuron szimmetrikus kapcsolatának súlyát, b_i a látható, illetve a_j pedig a rejtett neuronokhoz tartozó bias-okat. V és H a látható és rejtett egységek/neuronok száma. A modell által a v látható vektorhoz rendelt valószínűség:

$$p(v, \Theta) = \frac{\sum_h e^{-E(v,h)}}{\sum_u \sum_h e^{-E(u,h)}}, \tag{2}$$

ahol u eleme az input vektoroknak, h pedig a rejtett réteg állapotvektorainak. Mivel a korlátos Boltzmann gépben nem engedélyezett rejtett-rejtett és látható-látható kapcsolat, ezért $p(v|h)$-t és $p(h|v)$-t a következő módon definiálhatjuk:

$$\begin{aligned} p(h_j = 1|v, \Theta) &= \sigma(\sum_{i=1}^{V} w_{ij} v_i + a_j) \\ p(v_i = 1|h, \Theta) &= \sigma(\sum_{j=1}^{V} w_{ij} h_j + b_i), \end{aligned} \tag{3}$$

ahol $\sigma(x) = 1/(1 + \exp(-x))$ a szigmoid függvény.
Speciális változata az RBM-eknek az ún. Gauss-Bernoulli RBM, amely esetén a látható réteg neuronjai nem binárisak, hanem valós értékűek. Ezt valós input esetén szokás használni, és az energiafüggvény ekkor a következőképpen módosul:

$$E(v, h|\Theta) = \sum_{i=1}^{V} \frac{(v_i - b_i)^2}{2} - \sum_{i=1}^{V} \sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{j=1}^{H} a_j h_j \tag{4}$$

A v látható vektorhoz rendelt valószínűség pedig:

$$p(v_i = 1|h, \Theta) = \mathcal{N}(b_i + \sum_{j=1}^{H} w_{ij} h_j, 1), \tag{5}$$

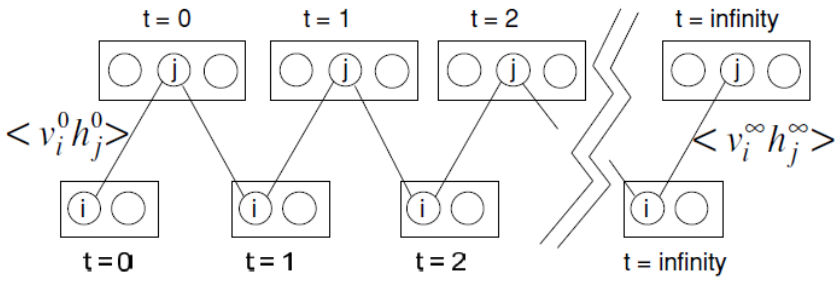
ahol $\mathcal{N}(\mu, \sigma)$ a μ várható értékű és σ varianciájú Gauss-eloszlás.
A pontos maximum likelihood tanulás alkalmatlan nagy méretű RBM esetén, ugyanis a derivált számításának időigénye exponenciálisan nő a hálózat méretével. A hatékony megoldást egy közelítő tanító algoritmus, az ún. kontrasztív divergencia (Contrastive Divergence, CD) biztosítja. Ennek a hatékony tanító algoritmusnak köszönhetően az RBM tökéletesen alkalmas arra, hogy a mély neuronhálók építőeleme legyen.

2.2. A CD-algoritmus

Hinton 2006-os cikkében javasolt egy tanító algoritmust a korlátos Boltzmann-gépekhez, amelyet kontrasztív divergenciának (Contrastive Divergence) nevezett el [2]. A javasolt módszer során a súlyok frissítési szabálya:

$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{input} - \langle v_i h_j \rangle_{rekonstrukcio}. \tag{6}$$

A (6) jobb oldalán található első tag az i. látható és j. rejtett egység korrelációja, bináris esetben annak gyakorisága, hogy mindkét neuron egyszerre aktív. A rejtett réteg állapotát adott inputvektorhoz (3) alapján számítjuk. A második tag jelentése hasonló, csak ekkor rekonstrukciós állapotokat használunk. Rekonstrukció alatt a következőt kell érteni: miután az input alapján meghatároztuk a rejtett réteg állapotait, (3) felhasználásával tudjuk (a rejtett réteg alapján) a látható réteg állapotait kiszámolni, ezután az így kapott látható réteghez generáljuk a rejtett réteget. A rekonstrukciót tetszőleges alkalommal megismételhetjük a 3. ábrán látható módon.



3. ábra. Rekonstrukciós lánc.

Mivel a rekonstrukciós lépések rendkívül időigényesek, ezért általában csak k db rekonstrukciót végzünk. A CD mohó algoritmusa $k = 1$ rekonstrukciót végez, és az alapján tanulja a súlyokat, általánosan ez a módszer terjedt el viszonylag kis időigénye és jó teljesítménye miatt. A mohó előtanítás során a súlyok frissítését a következő módon végezzük:

$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{input} - \langle v_i h_j \rangle_{t=1}. \tag{7}$$

Mint már korábban említettük, az előtanítás után a hálózatot átalakítjuk hagyományos neuronhálóvá, ami egyszerűen csak a súlyok átvitelével, illetve egy softmax-réteg felhelyezésével történik. Innentől a háló teljesen szokványosan tanítható felügyelt módon a backpropagation algoritmus segítségével. Mivel a tanításnak ez a része közismertnek tekinthető, ezért ennek az ismertetésétől eltekintünk.

3. Kísérleti eredmények

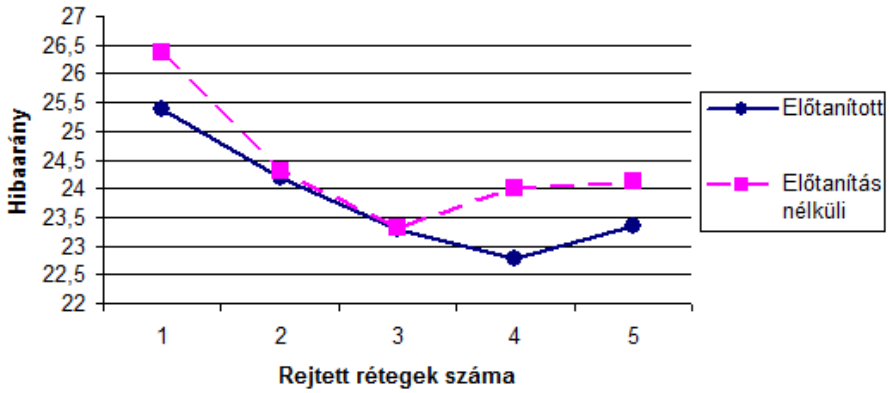
A továbbiakban kísérleti úton vizsgáljuk meg, hogy a mély neuronhálók milyen pontosságú beszédfelismerést tesznek lehetővé. Az akusztikus modellek készítése az ún. hibrid HMM/ANN sémát követi [1], azaz a neuronhálók feladata az akusztikus vektorok alapján megbecsülni a rejtett Markov-modell állapotainak valószínűségét, majd ezek alapján a teljes megfigyeléssorozathoz a rejtett Markov-modell a megszokott módon rendel valószínűségeket. Mivel a neuronhálóknak állapot-valószínűségeket kell visszaadniuk, ezért minden esetben első lépésben egy rejtett Markov-modellt tanítottunk be a HTK programcsomag használatával [7], majd ezt kényszerített illesztés üzemmódban futtatva kaptunk állapotcímkéket minden egyes spektrális vektorhoz. Ezeket a címkéket kellett a neuronhálónak megtanulnia, amihez inputként az aktuális akusztikus megfigyelést, plusz annak 7-7 szomszédját kapta meg. Az előtanítás a következő paraméterekkel történt: a tanulási ráta 0.002 volt a legalsó (Gauss-Bernoulli) rétegre, a magasabb (bináris) rétegekre 0.02. A A tanítás ún. kötegelt módon történt, ehhez a batch méretét 128-ra állítottuk, és 50 iterációt futtattunk az alsó, 20-at a többi rétegen. A backpropagation tanítás paraméterei az alábbiak voltak: a tanulási ráta 0.02-ről indult, a batch mérete ismét 128 volt. Mindegyik esetben alkalmaztuk az ún. momentum módszert, ennek paraméterét 0.9-re állítottuk.

A modellek kiértékelését háromféle adatbázison végeztük el. Mindhárom esetben azonos volt az előfeldolgozás: e célra a jól bevált mel-kepsztrális együtthatókat (MFCC) használtuk, egész pontosan 13 együtthatót (a nulladikat is beleértve) és az első-második deriváltjaikat. Közös volt még továbbá, hogy egyik esetben sem használtunk szószintű nyelvi modellt, pusztán egy beszédhangbigram támogatta a felismerést. Ennek megfelelően a felismerő kimenete is beszédhang szintű volt, ennek a hibáját (*1-accuracy*) fogjuk mérni a továbbiakban.

3.1. TIMIT

A TIMIT adatbázis a legismertebb angol nyelvű beszédadatbázis [8]. Habár mai szemmel nézve már egyértelműen kicsinek számít, a nagy előnye, hogy rengeteg eredményt közöltek rajta, továbbá a mérete miatt viszonylag gyorsan lehet kísérletezni vele, ezért továbbra is népszerű, főleg ha újszerű modellek első kiértékeléséről van szó. Esetünkben azért esett rá a választás, mert a mély neuronhálók első eredményeit is a TIMIT-en közölték [3], így kézenfekvőnek tűnt a használata az implementációnk helyességének igazolására.

A tanításhoz a szokványos tanító-tesztelő felosztást alkalmaztuk, azaz 3696 mondat szolgált tanításra és 192 tesztelésre (ez a kisebbik, ún. 'core' teszthalmaz). Az adatbázis 61 beszédhangcímkét használ, viszont sztenderdnek számít ezeket 39 címkére összevonni. Mi ezt az összevonást csupán a kiértékelés során tettük meg. Ez azt jelenti, hogy a monofón modellek tanítása során $61 \cdot 3 = 183$ címkével dolgoztunk (hangonként 3 állapot), azaz ennyi volt a neuronháló által megkülönböztetendő osztályok száma. Egy további kísérletben környezetfüggő (trifón) modelleket is készítettünk, ismét csak a HTK megfelelő eszközeit alkalmazva. Ennek eredményeként 858 állapot adódott, azaz ennyi osztályon tanítot-



4. ábra. Az előtanítás hatása a TIMIT core teszt halmazon a rejtett rétegek számának függvényében.

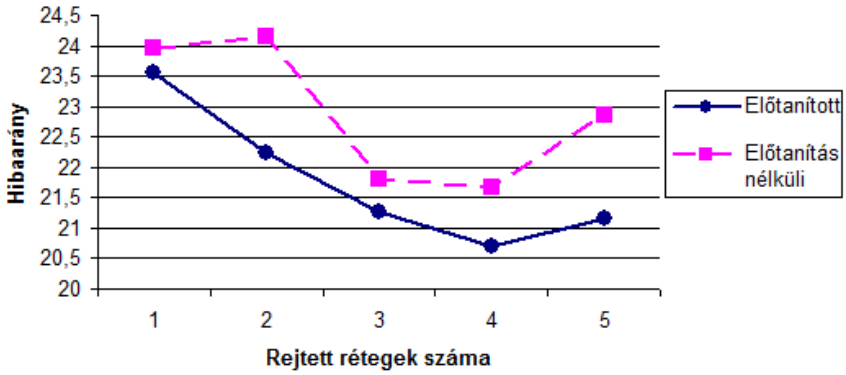
tuk a neuronhálót. A 4. ábra mutatja a monofón modellel kapott eredményeket, annak függvényében, hogy hány rejtett réteget használtunk. Az egyes rétegek neuronszáma minden esetben 1024 volt.

Az eredmények jól érzékeltetik, hogy érdemes egynél több rejtett réteget felvenni, de legfeljebb három-négyet, mert azon túl az eredmények nem javulnak számottevően (sőt, romlanak). Megfigyelhetjük továbbá, hogy az előtanítás tényleg segít, főleg mélyebb háló, azaz 4-5 réteg esetén: 4 rétegnél az eltérés az előtanítás nélküli és az előtanított háló között több mint 1% (ez kb. 5% hibacsökkenést jelent). Meg kell jegyezzük, hogy míg 4 réteg esetén az általunk kapott eredmény lényegében megegyezik az eredeti cikkben szereplővel ([3]), 5 réteg esetén nálunk már romlik az eredmény, míg ott javul. Ennek okait keressük, valószínűleg a paramétereket kell tovább hangolnunk (pl. az iterációszámot növelnünk). Azt is el kell mondanunk, hogy az itt látottaknál jobb eredményeket is el lehet érni mély neuronhálókkal (l. szintén [3]), ehhez azonban másfajta, jóval nagyobb elemszámú jellemzőkészletre van szükség. Mi most itt maradtunk az MFCC jellemzőknél, mivel ez a legáltalánosabban elfogadott jellemzőkészlet.

Rejtett rétegek száma	Hibaarány
3	22,04%
4	22,09%
5	21,91%

1. táblázat. Beszédhang-felismerési hibaarány a TIMIT adatbázison trifón címkék használata esetén.

A 1. táblázat a környezetfüggő címkékkel kapott eredményeket mutatja a TIMIT adatbázison (csak előtanításos esetre). Látható, hogy itt már öt rejtett réteg esetén kapjuk a legjobb eredményt, és az is látszik, hogy a monofón címkés eredményekhez képest kb. 1% javulás mutatkozik.



5. ábra. Az előtanítás hatása a híradós adatbázison a rejtett rétegek számának függvényében.

3.2. Híradós adatbázis

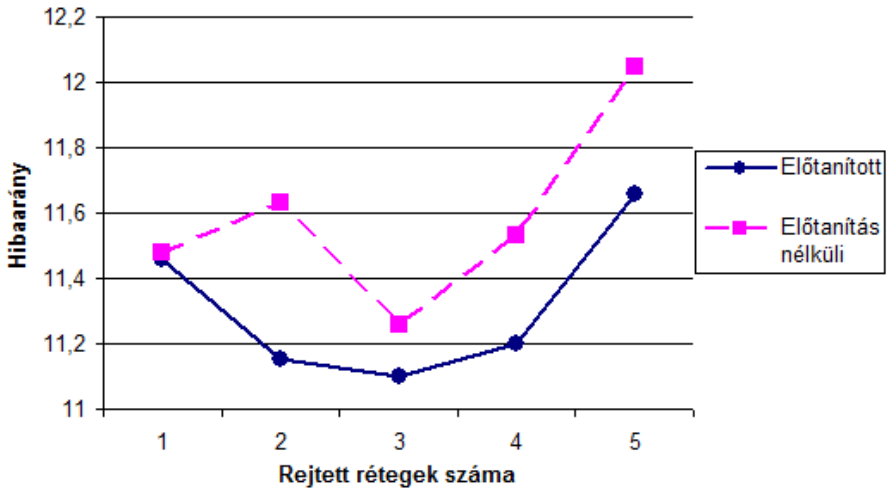
A magyar nyelvű felismerési kísérletekhez felhasznált híradós adatbázis megegyezik a [9]-ben ismertetettel. Az adatbázisnak ismét csak a a „tiszta” címkét kapott részeit használtuk fel, ami egy kb. öt és fél órás tanító és egy egyórás tesztelő részt eredményezett. Egy kétórás blokkot fenntartottunk a meta-paraméterek belövésére. Az adatbázis csak ortografikus átiratot tartalmaz, ezt egy egyszerű fonetikus átíróval alakítottuk át fonetikai címkékre, mely címkekészlet 52 elemből állt. Ebből a TIMIT adatbázisnál ismertetett módon készítettünk HMM-állapotoknak megfelelő címkézést.

A 5. ábra mutatja a monofón modellekkel elért eredményeket, különféle rétegszám mellett, ismét csak rétegenként 1024 neuronnal. Ezen az adatbázison az előtanítás kedvező hatása sokkal egyértelműbben megmutatkozik. A legjobb eredményt ismét csak négy rejtett réteggel kapjuk, a különbség az előtanítás nélküli és az előtanított rendszer között közel 1% (hibacsökkenésben kifejezve ez közel 5%). Összehasonlításképpen, korábban egy hagyományos, azaz egyetlen rejtett réteget használó hibrid modellel 23,07%-os eredményt közöltünk [9], ahhoz képest az itt szereplő 20,7% több mint 10%-os javulást jelent.

Rejtett rétegek száma	Hibaarány
3	17,94%
4	17,95%
5	18,51%

2. táblázat. Beszédhang-felismerési hibaarány a híradós adatbázison trifón címkék használata esetén

Ezen az adatbázison is megismételtük a kísérleteket környezetfüggő, azaz trifón címkékkel is (ismét csak előtanítással). Az eredmények a 2. táblázatban



6. ábra. Az előtanítás hatása a hangoskönyv-adatbázison a rejtett rétegek számának függvényében.

láthatóak. A legjobb értékeket ismét csak három és négy rejtett réteggel kaptuk, öt réteg esetén már romlás figyelhető meg. Az eredmények közel 3%-kal jobbak, mint monofón címkék esetén, ami hibacsökkenésben kifejezve 13%-os javulást jelent. Összehasonlításképp, a [9]-ben közölt legjobb trifónos korábbi eredmény 16.67% volt, tehát jobb a mostani eredménynél, de az összehasonlításhoz figyelembe kell venni, hogy ott egy ún. kétfázisú modellt alkalmaztunk, azaz két neuronháló volt egymásra tanítva, és a tanítás módja is jóval komplikáltabb volt az itt ismertetettnél. Semmi elvi akadálya nincs annak, hogy az ott közölt technológiát mély neuronhálókkal kombináljuk, ez várhatóan további javulást eredményezne.

3.3. Hangoskönyv

2009-ben beszédfelismerési kísérleteket végeztünk egy hangoskönyvvel, hogy lássuk, mit tudnak elérni a beszédfelismerők közel ideális beszédjel esetén [10]. Most ugyanazt az adatbázist vettük elő, ugyanazokkal az előkészítő lépésekkel és train-teszt felosztással. A felhasznált címkézés is ugyanaz volt.

A 6. ábra mutatja a különféle rétegszámmal elért eredményeket előtanulással és előtanulás nélkül, ismét csak rétegenként 1024 neuronnal. Érdekes módon ebben az esetben minimális volt csak az eltérés a 2-3-4 rétegszámú hálózatok eredményei között, és a legjobb eredményt három rejtett réteggel kaptuk. Az előtanulás ismét csak javított az eredményeken, de ennek hatása is kevésbé jelentős. A magyarázat valószínűleg az, hogy ez a tanulási feladat lényegesen könnyebb a másik kettőnél, és emiatt kevesebb rejtett réteg is elegendő a tanuláshoz.

Végezetül, a 3. táblázat mutatja a trifón címkézéssel kapott eredményeket. Ez esetben is a három rejtett réteges hálózat bizonyult a legjobbnak, és az

eredmények körülbelül egy százalékkal jobbak, mint a monofón címkék esetében. Ez relatív hibában kifejezve majdem tíz százalék, tehát szignifikáns javulás. Azt is elmondhatjuk továbbá, hogy az itt bemutatott eredmények lényegesen jobbak, mint a korábban tandem technológiával elért 13,16% ugyanezen adatbázison [10].

Rejtett rétegek száma	Hibaarány
3	10,24%
4	10,77%
5	11,32%

3. táblázat. Beszédhang-felismerési hibaarány a hangoskönyv-adatbázison trifón címkék használata esetén.

4. Konklúzió

Cikkünkben bemutattuk a mély neuronhálókra épülő akusztikus modelleket. A kísérleti eredmények egyértelműen igazolják, hogy a több rejtett réteg használata számottevően tud javítani az eredényeken. A „kontrasztív divergencia" előtanító algoritmus is egyértelműen hasznosnak bizonyult, bár ennek már most is sokan keresik a továbbfejlesztési lehetőségeit, főleg a nagy műveletigénye miatt. Mivel az egész témakör nagyon friss, bizonyosak lehetünk benne, hogy még számos újdonsággal fogunk találkozni e témában.

Hivatkozások

1. Bourlard, H., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach. Kluwer (1994)
2. Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation, Vol. 18 (2006) 1527–1554
3. Mohamed, A., Dahl, G. E., Hinton, G.: Acoustic modeling using deep belief networks. IEEE Trans. ASLP, Vol. 20, No. 1 (2012) 14–22
4. Dahl, G. E., Ranzato, M., Mohamed, A., Hinton, G.: Phone recognition with the mean-covariance restricted boltzmann machine. In: NIPS (2010) 469–477
5. Bishop, C. M.: Pattern Recognition and Machine Learning. Springer (2006)
6. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proc. AISTATS (2010) 249–256
7. Young, S. et al.: The HTK Book. Cambridge University Engineering Department (2005)
8. Lamel, L., Kassel, R., Seneff, S.: Speech database development: Design and analysis of the acoustic-phonetic corpus. In: Proc. DARPA Speech Recognition Workshop (1986) 121-124
9. Gosztolya G., Tóth L.: Kulcsszókeresési kísérletek hangzó híranyagokon beszédhang alapú felismerési technikákkal. In: MSZNY 2010 (2010) 224–235
10. Tóth L.: Beszédfelismerési kísérletek hangoskönyvekkel. In: MSZNY 2009 (2009) 206–216