Mély neuronhálóba integrált spektro-temporális jellemzőkinyerési módszer optimalizálása

Kovács György[1,2], Tóth László[2]

[1]Magyar Tudományos Akadémia, Nyelvtudományi Intézet,
Budapest VI., Benczúr utca 33.
[2]MTA-SzTE Mesterséges Intelligencia Kutatócsoport
Szeged Tisza Lajos körút 103,
e-mail:{gykovacs,tothl}@inf.u-szeged.hu

Kivonat Korábbi munkáinkban szignifikánsan javítottuk a fonémafelismerés pontosságát a jellemzőkinyerés mély neuronhálókba történő integrálásával. Az elkészült keretrendszerben azonban maradtak megválaszolandó kérdések, mint például a jellemzőkinyerési lépés paramétereinek optimalizálása, vagy a Δ és $\Delta\Delta$ együtthatók használata. Jelen munkánkban ezen kérdések megválaszolásával foglalkozunk a TIMIT fonémafelismerési valamint az Aurora-4 szófelismerési feladatokon. Először a TIMIT adatbázist felhasználva próbáljuk a jellemzőkinyerési paramétereket javítani, majd a kapott paramétereket a TIMIT és az Aurora-4 adatbázison értékeljük ki. Megmutatjuk, hogy mind a jellemzőkinyerési paraméterek módosítása, mind pedig a Δ és $\Delta\Delta$ együtthatók általunk javasolt felhasználási módja szignifikánsan javítja az elérhető hibaarányokat.
Kulcsszavak: TIMIT, Aurora-4, spektro-temporális jellemzőkinyerés

1. Bevezetés

Korábbi munkáinkban bemutattunk egy keretrendszert, amely a jellemzőkinyerési szakaszt integrálja a neuronhálóba [1]. Itt meghatározott neuronok bementének megválasztásával, súlyainak megfelelő inicializálásával, valamint azáltal, hogy lineáris aktivációs függvényt rendelünk hozzájuk, elértük, hogy az alsó rétegekben elhelyezkedő neuronok valósítsák meg a jellemzőkinyerés lépését, mintegy szűrőként viselkedve. Később ezt a megoldást fejlesztettük tovább egyenirányított mély neuronhálók valamint konvolúció alkalmazásával [2]. A jellemzőkinyerést végző szűrők méretének optimalizálására azonban kevés figyelmet fordítottunk. Ezt a kérdést két úton tudjuk megközelíteni. Egyrészt módosíthatjuk a szűrők méretét azok „technikai" méretén keresztül, azaz a bemenetet változatlanul hagyva, a szűrők mátrixának sor- és oszlopszámát változtatva. Másrészt módosíthatjuk a szűrők méretét pusztán fizikai méretük változtatásával, azaz a bemenet felbontását úgy variálva, hogy egy ugyanannyi sort és oszlopot tartalmazó szűrőmátrix eltérő nagyságú frekvenciatartományt és időintervallumot fedjen le. Annak érdekében, hogy a korábban létrehozott szűrőket továbbra is alkalmazni tudjuk, ez utóbbi megoldás mellett döntöttünk.

Az általunk használt neuronhálók mel skála szerinti sávszűrőket (mel filterbank) használnak bemenetként. Így azt, hogy adott számú sor eltérő nagyságú frekvenciatartományt fedjen le, könnyen elérhetjük a sávszűrők számának módosításával. Az adott számú oszlop által lefedett időintervallum módosítására a keretezés során alkalmazott keretek méretének, valamint a keretek közti lépésköz méretének módosításával is lehetőségünk nyílik. A keretek közti lépésköz hossza általában 10 és 20 ezredmásodperc között mozog [3], ám újabban mások érdekes eredményeket értek el ezen tartományon kívül eső lépésközök vizsgálatával [4], ezért mi is ez utóbbi megoldás mellett döntöttünk. Az ezen vizsgálatokhoz szükséges matematikai formalizmust a 2. fejezetben vezetjük be. Majd a kísérleti eszközök (adatbázisok és neuronhálók – 3. fejezet) leírása után a 4. fejezetben bemutatjuk a paraméterek optimalizálásához elvégzett kísérleteket.

A szűrők méretének változtatása mellett egy másik módszer, amivel a keretrendszer felismerési eredményeit próbáltuk javítani, a delta és gyorsulási ($\Delta\Delta$) együtthatók felhasználása volt. Ezek keretrendszerünkbe integrálásához (ahogy azt majd részletesen látjuk a 5. fejezet) szükség volt arra, hogy átfogalmazzuk az együtthatók kinyerésének problémáját.

A javasolt változtatásokat a TIMIT fonémafelismerési, valamint az Aurora-4 szófelismerési feladatán teszteltük. Az elvégzett tesztek eredményeit a 6. fejezetben ismertetjük, majd a 7. fejezetben konklúziók levonásával és a jövőbeni tervek ismertetésével zárjuk cikkünket.

2. Jelölések és paraméterek

2.1. Konvolúciós paraméterek

Korábbi munkáinkban két fontos változtatást vezettünk be keretrendszerünkbe [2,5]. Egyrészt, ahogy az napjainkban gyakori [6], egyenirányított neuronokat használtunk, ami azt jelenti, hogy a neuronok a rejtett rétegben hagyományos szigmoid aktivációs függényt helyett a következő függvényt valósítják meg: $max(0, x)$. Másrészt, az általunk alkalmazott neuronhálók Vesely és tsai. [7] nyomán időbeli konvolúciót alkalmaznak. Ennek magyarázatához talán az a legegyszerűbb, ha a konvolúciós réteget több különálló rétegként képzeljük el, amelyek osztoznak súlyaikon. Így a súlyok száma nem változik a konvolúcióval, a bemenetek és kimenetek száma viszont úgy viselkedik, mintha több különálló rétegünk lenne. Az R réteget leíró jellemzők tehát a következők:

- $\mathbf{R^I}, \mathbf{R^O}$: $\mathbf{R}$ réteg be- és kimenete
- $\mathbf{R}^{\#_\mathbf{n}}$: az $\mathbf{R}$ rétegbeli neuronok száma
- $\mathbf{R}^{\mathbf{W_i}}$: az $\mathbf{R}$ réteg i-edik neuronjához tartozó súlyvektor ($0 < i \leq \mathbf{R}^{\#_\mathbf{n}}$)
- $\mathbf{R_{C_m}}$: Ha az $\mathbf{R}$ réteg konvolúciót használ az időtartományban (azaz $\mathbf{R}$ konvolúciós réteg), és bemenetetét X eltérő időpontból veszi ($1 \leq m \leq X$), $\mathbf{R}$ rétegre úgy tekinthetünk, mint X darab különböző rétegre. Ebben az esetben $\mathbf{R_{C_m}}$ jelöli az m-edik ilyen réteget, melynek bemenete $\mathbf{R^I_{C_m}}$, és kimenete $\mathbf{R^O_{C_m}}$ (mivel a súlyokat ezek a rétegek megosztják egymás közt, az X réteg mindegyikének súlyai továbbra is $\mathbf{R^W}$ jelölést kapnak, és az X azonos méretű réteg neuronszámára továbbra is az $\mathbf{R}^{\#_\mathbf{n}}$ jelöléssel hivatkozunk).

2.2. Spektrogram- és ablak-paraméterek

A mel-skálás spektrális ábrázolás vagy röviden spektrogram (S) létrehozásánál többek között az alábbi paraméterek játszanak fontos szerepet:

- $\mathbf{S}^{\delta}_{\mathbf{W}}$: az S létrehozásához használt keret mérete (ezredmásodpercben)
- $\mathbf{S}^{\nu}_{\mathbf{W}}$: az egyes keretek kezdőpontja közti időintervallum (ezredmásodpercben)
- $\mathbf{S}^{\#}_{\mathbf{F}}$: a szűrőkészlet szűrőinek száma

A szűrőkhöz használt ablakok (P - Patch) paraméterei a következők:

- $\mathbf{P}^{\delta_{\mathbf{P}}}_{\mathbf{t}}$: $\mathbf{P}$ fizikai mérete az időtartományban. Azt jelzi, hány ezredmásodpercet fed le $\mathbf{P}$.
- $\mathbf{P}^{\delta_{\mathbf{P}}}_{\mathbf{f}}$: $\mathbf{P}$ fizikai mérete a frekvenciatartományban.
- $\mathbf{P}^{\delta_{\mathbf{t}}}_{\mathbf{t}}$: $\mathbf{P}$ „technikai" mérete az időtartományban. Azt jelzi, hány keretet fed le $\mathbf{P}$ (megjegyezzük, hogy ez meghatározható az $\mathbf{S}^{\nu}_{\mathbf{W}}$, $\mathbf{S}^{\delta}_{\mathbf{W}}$ és a $\mathbf{P}^{\delta_{\mathbf{P}}}_{\mathbf{t}}$ paraméterekből, ahogy $\mathbf{P}^{\delta_{\mathbf{P}}}_{\mathbf{t}}$ is meghatározható az $\mathbf{S}^{\nu}_{\mathbf{W}}$, $\mathbf{S}^{\delta}_{\mathbf{W}}$ és a $\mathbf{P}^{\delta_{\mathbf{t}}}_{\mathbf{t}}$ paraméterekből).
- $\mathbf{P}^{\delta_{\mathbf{t}}}_{\mathbf{f}}$: $\mathbf{P}$ „technikai" mérete a frekvenciatartományban. Azt jelzi, hány melszűrőt fed le $\mathbf{P}$ (megjegyezzük, hogy ez meghatározható $\mathbf{S}^{\#}_{\mathbf{F}}$ és $\mathbf{P}^{\delta_{\mathbf{P}}}_{\mathbf{f}}$ paraméterekből, ahogy $\mathbf{P}^{\delta_{\mathbf{P}}}_{\mathbf{f}}$ is meghatározható $\mathbf{P}^{\delta_{\mathbf{t}}}_{\mathbf{f}}$ és $\mathbf{S}^{\#}_{\mathbf{F}}$ paraméterekből).
- $\mathbf{P}^{\nu_{\mathbf{P}}}_{\mathbf{f}}$: a szomszédos ablakok átfedési aránya a frekvenciatartományban
- $\mathbf{P}^{\#}_{\mathbf{f}}$: az ablakok száma, amelyek szükségesek a használt frekvenciatartomány lefedéséhez (feltételezve, hogy a szomszédos ablakok átfedése: $\mathbf{P}^{\nu_{\mathbf{P}}}_{\mathbf{f}}$)
- $\overline{\mathbf{P}}$: $\mathbf{P}$ (ami egy 2-dimenziós mátrix) vektor formában felírva
- $\overline{\mathbf{P}}^{\delta_{\mathbf{t}}}$: $\overline{\mathbf{P}}$ vektor hossza. Kiszámítható az alábbi formulával: $\mathbf{P}^{\delta_{\mathbf{t}}}_{\mathbf{t}} \cdot \mathbf{P}^{\delta_{\mathbf{t}}}_{\mathbf{f}}$

Mivel a konvolúciót az időtartományban alkalmazzuk, így neuronhálónk nem csak a frekvencia- de az időtartományban is több ablakot használ bemenetként. Az ehhez kapcsolódó paraméterek és jelölések a következők:

- $\delta_{\mathbf{T}}$: az időtartam (ezredmásodpercben) amit le kívánunk fedni az egymást fedő ablakok használatával
- $\mathbf{P}^{\#_{\mathbf{i}}}_{\mathbf{t}}$: az ablakok száma, amelyek szükségesek a megadott $\delta_{\mathbf{T}}$ időtartam lefedéséhez (feltételezve, hogy közvetlen szomszédos ablakokat használunk)
- $\mathbf{P}^{\nu_{\mathbf{P}}}_{\mathbf{t}}$: a szomszédos ablakok átfedési aránya az időtartományban
- $\mathbf{P}^{\#}_{\mathbf{t}}$: az ablakok száma, amelyek szükségesek a megadott $\delta_{\mathbf{T}}$ időtartam lefedéséhez (feltételezve, hogy a szomszédos ablakok átfedése: $\mathbf{P}^{\nu_{\mathbf{P}}}_{\mathbf{t}}$)
- $\mathbf{P}^{\nu_{\mathbf{t}}}_{\mathbf{t}}$: azon közvetlenül szomszédos ablakok száma, melyeket kihagyunk, hogy $\mathbf{P}^{\nu_{\mathbf{P}}}_{\mathbf{t}}$-nek megfelelő átfedést érjünk el a az ablakok között (ahogy a jelölés sugallja, ez a „technikai" oldala az átfedésnek, amit keretekben adunk meg, míg a fizikai oldala – $\mathbf{P}^{\nu_{\mathbf{P}}}_{\mathbf{t}}$ – értelmezhető ezredmásodpercekben is)

Ha a rendszer által egy adott időpillanatban használt ablakok számát egy mátrixként fogjuk fel, ahol az azonos időtartamból és frekvenciatartományból származó ablakok alkotják a mátrix oszlopait, és sorait, az ablakot amely a mátrix i-edik sorából ($0 < i \leq \mathbf{P}^{\#}_{\mathbf{f}}$) és a j-edik oszlopából ($0 < j \leq \mathbf{P}^{\#_{\mathbf{i}}}_{\mathbf{t}}$) származik jelölhetjuk $\mathbf{P_{ij}}$-vel. Ekkor a korábban bemutatott $\overline{\mathbf{P}}^{\delta_{\mathbf{t}}}$ paraméter a következő jelöléssel kellene rendelkezzen: $\overline{\mathbf{P}}^{\delta_{\mathbf{t}}}_{\mathbf{ij}}$. Mivel azonban kísérleteinkben a különböző idő- és frekvenciatartományból vett ablakok „technikai" mérete megegyezik, a továbbiakban maradunk a korábban bevezetett jelölésnél.

3. Kísérleti eszközök

3.1. TIMIT

A paraméterek hangolásához a TIMIT beszédadatbázist használtuk [8]. A neuronhálók súlyait a 3969 mondatból álló tanítóhalmaz kilencven százalékán tanítottuk, a fennmaradó tíz szálékot pedig a megállási feltétel kiértékelésére, és a paraméterek kiválasztására használtuk. Ezen paraméterek használatával azt követően újabb neuronhálókat tanítottunk, amelyeket a 192 mondatot tartalmazó „mag" (core) teszthalmazon értékeltünk ki. Kiértékelés előtt a fonémacímkéket 39 kategóriába vontuk össze, a bevett gyakorlatnak megfelelően [9].

3.2. Aurora-4

Az Aurora-4 a Wall Street Journal beszédadatbázis zajosított változata [10]. Két 7138 mondatból álló tanítóhalmazt, és 14, egyenként 330 mondatból álló teszthalmazt tartalmaz. A tanítóhalmaz első (tiszta) változata a mondatok zaj nélküli változatát tartalmazza Sennheiser mikrofonnal rögzítve, míg a második változatban az egyes mondatok különböző zajokkal szennyezettek, illetve rögzítésük eltérő mikrofonnal történt. Jelen cikkünkben csak a második (multi-condition) tanítóhalmazt használtuk. Ennek kilencven százalékán tanítottuk a neuronhálók súlyait, míg a fennmaradó részt a megállási feltétel kiértékelésére használtuk.

A kiértékelést az összes teszthalmazon végeztük. Ezen teszthalmazok ugyanazt a 330 mondatot tartalmazzák különböző verziókban: az első hét teszthalmazban lévő hangfájlok rögzítése a Sennheiser mikrofonnal történt, míg a második hét teszthalmazban ettől eltérő mikrofonnal rögzített felvételeket találunk. Mindkét csoport belső felosztása azonos: az első halmaz tiszta beszédet tartalmaz, míg a következő hatban hat különböző zajjal szennyezett beszéd található.

3.3. Neuronháló

A neuronhálók ismertetésének egyszerűsítése céljából leírásukat a különböző funkciókat ellátó rétegek leírására bontjuk.

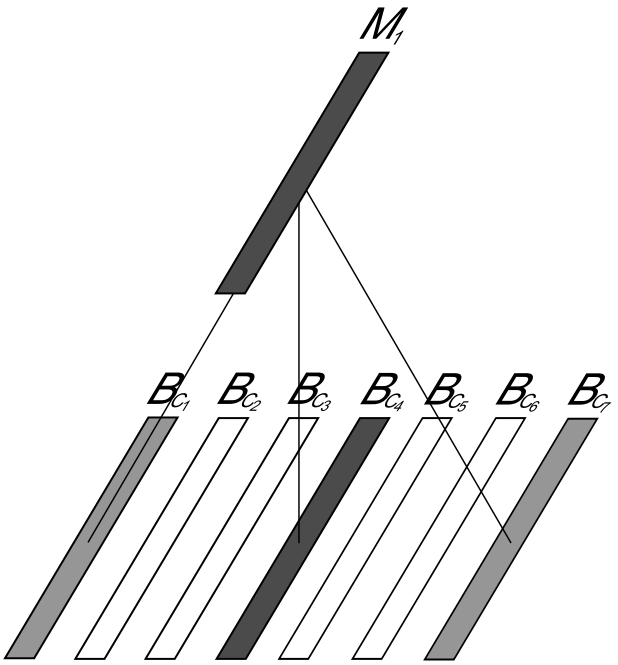
Szűrőrétegek. Minden, a szűrés (Filtering) megvalósításáért felelős réteg ($\mathbf{F_i}$) egy megadott frekvenciatartományból veszi az input-ablakokat. A réteg be- és kimenete a következőképp írható le:

$$\mathbf{F^{I}_{iC_j}} = \overline{\mathbf{P_{ij}}},\ ahol \quad \begin{matrix} 0 < i \leq \mathbf{P^{\#}_{f}} \\ 0 < j \leq \mathbf{P^{\#_i}_{t}} \end{matrix} \tag{1}$$

$$\mathbf{F^{O}_{iC_j}} = [o_{1j}, ..., o_{\mathbf{F^{\#_n}_i}\,j}],$$

$$o_{kj} = \sum_{x=1}^{\overline{\mathbf{P}}^{\delta_t}} \overline{\mathbf{P_{ij}}}[x] \cdot \mathbf{F^{W_k}_{i}}[x] + b_k, \quad ahol\ 0 < k \leq \mathbf{F^{\#_n}_{i}}\,^1 \tag{2}$$

[1] Mivel kísérleteink során minden szűrőréteg ugyanannyi neuront tartalmaz, erre az értékre a későbbiekben a $\mathbf{F}^{\#_n}$ szimbólummal fogunk hivatkozni



1. ábra. A konvolúciós bottleneck réteg ($\mathbf{B}$), valamint az első egyszerű egyenirányított réteg ($\mathbf{M_1}$) illusztrációja. Azt az esetet mutatja be, ahol $\mathbf{P_t^{\#_i}} = 7$, és $\mathbf{P_t^{\nu_t}} = 2$. Egyben illusztrálja azt a tényt is, hogy ablakok megfelelő átfedéséről az időtartományban a bottleneck réteget követő réteg gondoskodik azáltal, hogy a bottleneck réteg megadott számú konvolúciós kimenetét átugorja bemenete beolvasása közben.

„Konvolúciós" réteg. A szűrők megvalósításáért felelős réteget egy (vagy több) további konvolúciós réteg követi, melyek közül az utolsó az „üvegnyak" (bottleneck) réteg. Számunkra itt az első réteg bemenete érdekes:

$$\mathbf{Conv^I_{C_j}} = [\mathbf{F^O_{1C_j}}, ..., \mathbf{F^O_{P_f^{\#}C_j}}],\ ahol\ \ 0 < j \leq \mathbf{P_t^{\#_i}} \tag{3}$$

Egyszerű egyenirányított réteg. Az első nem-konvolúciós egyenirányított réteg ($\mathbf{M_1}$ – lásd 1. ábra) kombinálja a konvolúciós bottleneck ($\mathbf{B}$) réteg kimeneteit, azáltal, hogy kimenetét a következő módon használja fel bemenetként:

$$\mathbf{M_1^I} = [\mathbf{B^O_{C_1}}, \mathbf{B^O_{C_{1+P_t^{\nu_t}+1}}}, ..., \mathbf{B^O_{C_j}}],\ ahol\ \ j = \mathbf{P_t^{\#_i}} \tag{4}$$

Innentől a konvolúciós mély neuronháló (beleértve az $\mathbf{O}$ kimeneti réteget is) ugyan úgy működik, mint bámely hagyományos mély háló, így a további rétegek részletes leírásától eltekintünk.

1. táblázat. Paraméterbeállítások a frekvenciatartományra vonatkozóan

paraméterek	F1	F2	F3	F4	F5
$\mathbf{S}_{\mathbf{F}}^{\#}$	18	26	34	42	50
$\mathbf{P}_{\mathbf{f}}^{\delta_{\mathbf{P}}}$	~1420 mel	~980 mel	~750 mel	~610 mel	~510 mel
$\mathbf{P}_{\mathbf{f}}^{\#}$	3	5	7	9	11

4. Paraméterek optimalizálása

A kísérletek célja az volt, hogy megvizsgáljuk, milyen hatással van a szűrők fizikai méretének változása (a „technikai" méret változtatása nélkül) a felismerési eredményekre. Ezen kísérletekhez bizonyos paramétereket rögzítettnek vettünk:

- $\mathbf{S}_{\mathbf{W}}^{\delta} = 25$ ms
- $\delta_{\mathbf{T}} \approx 265$ ms
- $\overline{\mathbf{P}}^{\delta_{\mathbf{t}}} = 81$ ($\mathbf{P}_{\mathbf{t}}^{\delta_{\mathbf{t}}} = 9$, $\mathbf{P}_{\mathbf{f}}^{\delta_{\mathbf{t}}} = 9$)
- $\mathbf{P}_{\mathbf{t}}^{\nu_{\mathbf{P}}} = \frac{\left\lceil \mathbf{P}_{\mathbf{t}}^{\delta_{\mathbf{t}}}/2 \right\rceil}{\mathbf{P}_{\mathbf{t}}^{\delta_{\mathbf{t}}}}$ ($\mathbf{P}_{\mathbf{t}}^{\delta_{\mathbf{t}}} = 9$ felhasználásával adódik, hogy $\mathbf{P}_{\mathbf{t}}^{\nu_{\mathbf{t}}} = 3$ keret)
- $\mathbf{P}_{\mathbf{f}}^{\nu_{\mathbf{P}}} = \frac{\left\lceil \mathbf{P}_{\mathbf{f}}^{\delta_{\mathbf{t}}}/2 \right\rceil}{\mathbf{P}_{\mathbf{f}}^{\delta_{\mathbf{t}}}}$.

A különböző paraméter-beállítások hatásának tanulmányozására 5 beállítást hoztunk létre a frekvenciatartományra vonatkozó paraméterekre nézve (leolvashatók a 1. táblázatból), és 3 beállítást az időtartományra vonatkozó paraméterekre nézve (leolvashatók a 2. táblázatból). Így összesen 15 párt vizsgáltunk a kísérleteink során.

A vizsgálatához használt neuronhálók bizonyos paraméterei a táblázatokban leírt paraméterek függvényében alakultak, míg mások kötöttek voltak. Ez utóbbiak a következők:

- $\mathbf{F}^{\#_{\mathbf{n}}} = 9$
- $\mathbf{B}^{\#_{\mathbf{n}}} = 200$
- $\mathbf{M}_{\mathbf{1}}^{\#_{\mathbf{n}}} = \mathbf{M}_{\mathbf{2}}^{\#_{\mathbf{n}}} = 1000$
- $\mathbf{O}^{\#_{\mathbf{n}}} = 183/61$ (a három- és egyállapotú fonémamodellekhez)

A paraméterek optimalizálásáért végzett kísérletek során használt neuronhálóknak a szűrők megvalósításáért felelős ($\mathbf{F_i}$) rétegein kívül egyetlen konvolúciós rétegük volt, a bottleneck ($\mathbf{B}$) réteg.

2. táblázat. Paraméterbeállítások az időtartományra vonatkozóan

paraméterek	T1	T2	T3
$\mathbf{S}_{\mathbf{W}}^{\nu}$	10 ms	8 ms	6 ms
$\mathbf{P}_{\mathbf{t}}^{\delta_{\mathbf{P}}}$	~105 ms	~89 ms	~73 ms
$\mathbf{P}_{\mathbf{t}}^{\#_{\mathbf{i}}}$	17	25	33
$\mathbf{P}_{\mathbf{t}}^{\#}$	5	7	11

3. táblázat. Fonémafelismerési hibaarányok (10 függetlenül tanított neuronháló eredményeinek átlaga) a TIMIT validációs halmazán egy- és háromállapotú fonémamodellek esetére (a legjobb eredmények, és az azoktól szignifikánsan nem eltérő eredmények vastagon szedve mindkét esetben).

	61 egyállapotú modell			61 háromállapotú modell		
	T1	T2	T3	T1	T2	T3
F1	21,36%	20,78%	20,68%	21,29%	20,39%	20,04%
F2	*20,65%*	20,20%	20,12%	*20,26%*	19,64%	19,28%
F3	20,18%	19,81%	20,00%	20,05%	19,32%	19,03%
F4	19,89%	**19,65%**	19,79%	19,61%	19,15%	**18,88%**
F5	19,85%	**19,52%**	19,84%	19,64%	19,08%	**18,86%**

4.1. Eredmények

A tanítóhalmazból lehasított tíz százalékra, mint validációs halmazra, megvizsgáltuk a fonémafelismerési eredményeket abban az esetben, ha egyállapotú vagy háromállapotú fonémamodellek segítségével végeztük a tanítást.

Az így kapott eredmények leolvashatók a 3. táblázatból. Korábbi kísérleteink során az F2/T1 paramétereinek megfelelő beállításokat használtuk (az ehhez kapcsolódó beállításokat dőlt betűkkel emeltük ki a táblázatban). Látható, hogy a két esetben a legjobb eredményt adó beállítások eltérnek egymástól. Az abszolút értékben legjobb eredményeket a táblázat jobb oldalán található F4/T3 és F5/T3 beállításokkal kaptuk. Azaz a legjobb felismerési eredményt akkor értük el, ha a keretek közti lépésközt 10 ezredmásodpercről 6 ezredmásodpercre csökkentettük, és a mel-szűrők számát 26-ról 42-re vagy 46-ra növeltük. E két beállítás közül választottunk az előbbit, mivel a használatukkal kapott felismerési eredménye között szignifikáns különbséget nem találtunk, így a szűrők számának további növelését nem láttuk indokoltnak. A 6. fejezetben végzett tesztek során tehát ezt a paraméterbeállítást (F4/T3) fogjuk összehasonlítani az eredeti (F2/T1) beállításokkal, azt vizsgálandó, hogy a javasolt változtatások valóban jobb eredményre vezetnek-e.

5. Delta és gyorsulási együtthatók

Korábbi publikációinkban többször előkerült, hogy a Δ és $\Delta\Delta$ együtthatók hozzáadása a keretrendszerünkhöz hasznos lenne [1,5]. Ezen véleményünket arra alapoztuk, hogy korábban a delta és gyorsulási együtthatók hozzáadása a jellemzőkészlethez javította az elért eredményeinket [11]. A megadott együtthatók használata a jelen keretrendszerben kivitelezhető lenne oly módon, hogy új neuronokat veszünk fel a jellemzőkinyerési réteg után melyek Δ és $\Delta\Delta$ együtthatók kinyerését valósítják meg, és megoldjuk, hogy a hibavisszaterjesztés áthaladjon ezeken a neuronokon. Van azonban egy egyszerűbben kivitelezhető megoldás. A

Δ együtthatók a következő formulával állnak elő:

$$d_T = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{T+\theta} - c_{T-\theta})}{2 \cdot \sum_{\theta=1}^{\Theta} \theta^2}, \tag{5}$$

ahol d_T a Δ együttható T időpontban, a $c_{T-\theta}$ és $c_{T+\theta}$ közötti konstans együtthatókból számítva [12]. Ha ezt a formulát a szűrők alkalmazásával kapott jellemzőkre alkalmazzuk,

$$c_T = \sum_{t=1}^{N} \sum_{f=1}^{M} P_T(f,t) \cdot F(f,t), \tag{6}$$

ahol P mintákat az S spektrogramból a következő módon nyerjük ki:

$$P_T(f,t) = S(f, T+t), \tag{7}$$

a következő egyenletet kapjuk:

$$d_T = \frac{\sum_{\theta=1}^{\Theta} \theta \left(\sum_{t=1}^{N} \sum_{f=1}^{M} F(f,t) \cdot \left(S(f,T+\theta+t) - S(f,T-\theta+t) \right) \right)}{2 \cdot \sum_{\theta=1}^{\Theta} \theta^2}, \tag{8}$$

Mivel az osztó Θ megválasztása után nem függ egyéb paramétertől, adott Θ esetén konstans. Vezessük be tehát a következő konstanst: $\vartheta = 2 \cdot \sum_{\theta=1}^{\Theta} \theta^2$. Ezt használva, valamint újrarendezve (8) egyenletet, a következő formulát kapjuk a Δ együttható számítására, egy jellemző esetén, amit egy S spektrogramra alkalmazott szűrő kimenetként kapunk:

$$d_T = \sum_{t=1}^{N} \sum_{f=1}^{M} F(f,t) \cdot \frac{\sum_{\theta=1}^{\Theta} \theta \left(S(f,T+\theta+t) - S(f,T-\theta+t) \right)}{\vartheta}, \tag{9}$$

Másrészről, ha először alkalmazzuk a (5) egyenletet az S spektrogramra, a spektrogramnak egy Δ változatát kapjuk, ahol az f frekvenciához és t időponthoz tartozó elemet a következőképp kapjuk:

$$S_\Delta(f,t) = \frac{\sum_{\theta=1}^{\Theta} \theta \left(S(f,t+\theta) - S(f,t-\theta) \right)}{\vartheta} \tag{10}$$

Ekkor F szűrő alkalmazása a spektrogramból T időpontban kinyert mintára a következő egyenlet megoldását jelentené:

$$c_T = \sum_{f=1}^{N} \sum_{t=1}^{M} F(f,t) \cdot S_\Delta(f, T+t), \tag{11}$$

ami megegyezik (9) egyenlettel. Tehát ha a célunk egy jellemző Δ együtthatójának kinyerése, azt elérhetjük úgy is, hogy először a spektrogram Δ változatát állítjuk elő, majd abból nyerjük ki a jellemzőt.

6. Kísérleti eredmények

A paraméterek meghatározásához végzett kísérletekkel szemben az alább ismertetett kísérletekben a hálók egymáshoz viszonyított teljesítményén kívül az abszolút teljesítmény is fontos volt. Ezért ezekben a kísérletekben nagyobb hálókat alkalmaztunk. A bottleneck réteg elé további három (egyenként 1000 neuront tartalmazó) konvolúciós réteget helyeztünk el, és az így kapott hálókat két lépésben tanítottuk. Az első lépésben konvolúció nélkül tanítottuk a hálót oly módon, hogy a kimeneti réteget közvetlenül a bottleneck réteg után helyeztük el, majd a következő lépés előtt ezt a kimeneti réteget töröltük, két (egyenként 1000 neuront tartalmazó) réteget és egy új kimeneti réteget vettünk fel, majd egy újabb tanítást indítottunk, ezúttal konvolúció használatával. További módosítás a korábbi kísérletekhez képest, hogy az eredeti beállításaink (F2/T1) esetén a bottleneck réteg 220 neuront tartalmazott (szemben a korábbi 200-al), és a korábban 1000 neuront tartalmazó rétegek neuronszámát 1100-ra növeltük. Ezzel azt biztosítottuk, hogy a különböző jellemzőkinyerési paraméterbeállításokkal dolgozó neuronhálók mérete közel azonos legyen.

6.1. Eredmények a TIMIT beszédadatbázison

A mások által elért eredményekkel való jobb összehasonlítás érdekében ezekben a kísérletekben a tanítást ([6] nyomán) 858 állapot felhasználásával végeztük. A kiértékelés előtt ugyanúgy elvégeztük a 39 kategóriába való összevonást, mint korábban. Az így kapott eredmények leolvashatók a 4. táblázatból. Először vizsgáljuk meg az első két sort, azaz azt a két esetet, amikor a hálók szűrés megvalósításért felelős rétegeit véletlen súlyokkal (1. sor), illetve a korábban bemutatott Gábor szűrők [5] alapján (2. sor) inicializáljuk. Láthatjuk, hogy a súlyok Gábor szűrők alapján történő inicializálása 0,2 százalékpontos hibaarány-csökkenéshez vezet az eredményekben, ami szignifikáns ugyan ($p = 0,025$ értéken), ám minimális. Ezért a további kísérletekben a szűrést megvalósító rétegek súlyait, a háló többi súlyához hasonlóan, véletlen számokkal inicializáltuk. Kiolvasható továbbá a táblázatból, hogy az új paraméterek használatával jelentős javulást értünk el a hibaarányt tekintve (22 százalékos relatív hibacsökkenés), és a Δ valamint $\Delta\Delta$ együtthatók hozzáadásával ezen az eredményen is javítani tudtunk.

4. táblázat. Fonéma szintű hibaarányok (10 függetlenül tanított neuronháló eredményeinek átlaga) a TIMIT „mag" teszthalmazán (a legjobb eredmények, és az azoktól szignifikánsan nem eltérő eredmények vastagon szedve).

Inicializálás	Paraméterek		Δ	PER
	Frekvencia	Idő	$\Delta\Delta$	
Random	F2	T1		24,4%
Gábor	F2	T1		24,2%
Random	F4	T3		18,8%
Random	F4	T3	✓	**18,5%**

5. táblázat. Fonéma szintű hibaarányok (PER) a TIMIT „mag” teszthalmazán (a legjobb eredmények vastagon szedve).

Módszer	PER
Plahl és tsai. [15]	19,1%
Tóth [6]	18,7%
Jelen cikk	*18,5%*
Graves és tsai. [13]	17,7%
Tóth [14]	**16,7%**

Az elért eredményeket összevetve az irodalomban találtakkal (5. táblázat) azt látjuk, hogy a javasolt változtatásokkal a rendszerünk versenyképes eredményeket produkál. Bár az elért fonémafelismerési eredmények elmaradnak Graves és tsai. [13] eredményeitől, ám ők kísérleteikben rekurrens hálókat alkalmaztak. Eredményeink továbbá jelentősen elmaradnak Tóth 2014-es eredményeitől [14], azonban az általa használt hálók az időtartományban *és* a frekvenciatartományban is alkalmaztak konvolúciót, továbbá a dropout módszert is felhasználták. Rendszerünk leginkább ugyanazon szerző egy korábbi cikkében bemutatott rendszerével összehasonlítható [6], melynek eredményein kis mértékben javítani is tudtunk, úgy, hogy az általunk használt hálók méretei csupán negyede az említett cikkben használt háló méretének.

6.2. Eredmények az Aurora-4 beszédadatbázison

Annak érdekében, hogy a neuronhálók teljesítményét különböző zajtípusok (illetve átviteli karakterisztikák) esetén is vizsgálni tudjuk, az elvégzett kísérleteket megismételtük az Aurora-4 szófelismerési feladatára is (a multi-condition tanítóhalmaz felhasználásával). Az eredmények leolvashatók a 6. táblázatból. Ahogy látható, mindkét javasolt módosítás a szófelismerési pontosság javulásához vezetett az Aurora-4 teszthalmazain. A frekvencia- és időparaméterek módosításával 4 százalékos, a Δ valamint $\Delta\Delta$ együtthatók felvételével pedig további 2 százalékos relatív hibaarány-csökkenést értünk el. A különbség szignifikáns mind az első ($p = 0,00005$ értéken) mind pedig a második módosítás esetén ($p = 0,00044$ értéken).

6. táblázat. Szószintű hibaarányok (5 függetlenül tanított neuronháló eredményeinek átlaga) az Aurora-4 teszthalmazán (a legjobb eredmények, es az azoktól szignifikánsan nem eltérő eredmények vastagon szedve).

Inicializálás	Paraméterek		Δ	WER
	Frekvencia	Idő	$\Delta\Delta$	
Random	F2	T1		12,4%
Random	F4	T3		11,9%
Random	F4	T3	✓	**11,6%**

7. táblázat. Szószintű hibaarányok (WER) az Aurora-4 teszthalmazán (a legjobb eredmények vastagon szedve).

Módszer	WER
Chang, Morgan [16]	16,6%
Castro és tsai. [17]	12,3%
D. Baby és tsai. [18]	11,9%
Jelen cikk	***11,6%***

Az eredményeinket az irodalomban talált eredményekkel ismét egy külön táblázatban (7. táblázat) hasonlítjuk össze. Chang és Morgan [16] hozzánk hasonlóan mély konvolúciós hálókat alkalmaztak, melyek alsó rétegébe szűrők együtthatóit építették be, ám velünk ellentétben náluk bemenetként a PNS (Power Normalized Spectrum) szolgált, továbbá ők több és nagyobb szűrőket alkalmaztak, de nem használták a Δ valamint gyorsulási együtthatókat. Castro és tsai. [17] szintén felhasználtak Gábor szűrőket is, ám legjobb eredményeiket az úgynevezett Amplitude Modulation Filter Bank (AMFB) segítségével érték el. Valamint szintén mély neuronhálókat alkalmaztak, ám ők ezt a beépített Kaldi recept alapján, előtanítás használatával tették. További különbség, hogy a mieinknél jelentősen nagyobb (7 rejtett rétegű, rétegenként 2048 neuront tartalmazó) hálókat használtak. D. Baby és tsai. [18] egy a miénktől jelentősen eltérő megközelítést a minta-alapú beszédkiemelés módszerét alkalmazták, hozzánk hasonlóan egy DNN/HMM hibrid architektúrába (ám Castrohoz és társaihoz hasonlóan a mieinknél jelentősen nagyobb – 6, egyenként 2048 rétegből álló – neuronhálót használva). Ahogy a 7. táblázatból látható, az általunk elért legjobb eredmények felülmúlják a három említett cikkben bemutatott eredményeket (30 és 2,5 százalék közötti relatív hibaarány-csökkenéssel).

7. Konklúzió és jövőbeni munka

Cikkünkben két módosítást javasoltunk az általunk használt keretrendszerhez, a TIMIT beszédadatbázison végzett kísérletek, valamint korábbi kísérleteink alapján. A TIMIT fonémafelismerési, valamint az Aurora-4 szófelismerési feladaton végzet kísérletek nyomán azt láttuk, hogy mindkét módosítás az eredmények szignifikáns javulásához vezet. Bár a TIMIT adatbázison elért fonémafelismerési eredményeken azt láttuk, hogy a háló jellemzőkinyerésért felelős rétegeiben található súlyok Gábor-szűrők alapján történő inicializálása szignifikáns javulást eredményezett, ez a javulás minimális volt. A későbbiekben érdemes lehet megvizsgálni a két lépéses tanítás hatását a szűrőkre (beleértve azt az esetet, amikor az első vagy második lépés során a szűrőkhöz kapcsolódó súlyokat változatlanul hagyjuk) valamint a kimenetként kapott szűrők hasznosságára a felismerési eredmények szempontjából.

Hivatkozások

1. Kovács, Gy., Tóth, L.: The joint optimization of spectro-temporal features and neural net classifiers. In: Proc. TSD. (2013) 552–559
2. Kovács, Gy., Tóth, L.: Joint optimization of spectro-temporal features and deep neural nets for robust automatic speech recognition. Acta Cybernetica **22**(1) (2015) 117–134
3. Picone, J.W.: Signal modeling techniques in speech recognition. Proceedings of the IEEE **81**(9) (1993) 1215–1247
4. Pundak, G., Sainath, T.: Lower frame rate neural network acoustic models. In: proc. Interspeech. (2016) 22–26
5. Kovács, Gy., Tóth, L., Van Compernolle, D.: Selection and enhancement of Gabor filters for automatic speech recognition. International Journal of Speech Technology **18**(1) (2015) 1–16
6. Tóth, L.: Convolutional deep rectifier neural nets for phone recognition. In: Proc. Interspeech, IEEE (2013) 1722–1726
7. Veselý, K., Karafiát, M., Grézl, F.: Convolutive bottleneck network features for LVCSR. In: Proc. ASRU. (2011) 42 – 47
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) 100–109
9. Lee, K.F., Hon, H.: Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust., Speech, Signal Processing **37** (1989) 1641–1648
10. Hirsch, H.G., Pearce, D.: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW). (2000) 29–32
11. Kovács, Gy., Tóth, L.: Phone recognition experiments with 2D DCT spectro-temporal features. In: Proc. SACI, IEEE (2011) 143–146
12. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, Cambridge (2005)
13. Graves, A., Mohamed, A., Hinton, G.E.: Speech recognition with deep recurrent neural networks. In: Proc. ICASSP. (2013) 6645–6649
14. Tóth, L.: Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition. In: Proc. ICASSP. (2014) 190–194
15. Plahl, C., Sainath, T.N., Ramabhadran, B., Nahamoo, D.: Improved pre-training of deep belief networks using sparse encoding symmetric machines. In: Proc. ICASSP. (2012) 4165–4168
16. Chang, S.Y., Morgan, N.: Robust CNN-based speech recognition with Gabor filter kernels. In: Proc. Interspeech. (2014) 905–909
17. Martinez, A.M.C., Moritz, N., Meyer, B.T.: Should deep neural nets have ears? the role of auditory features in deep learning approaches. In: Proc. Interspeech. (2014) 2435–2439
18. Baby, D., Gemmeke, J.F., Virtanen, T., Van Hamme, H.: Exemplar-based speech enhancement for deep neural network based automatic speech recognition. In: Proc. ICASSP. (2015) 4485–4489