Mély neuronhálós akusztikus modellek súlyinicializálásának vizsgálata

Pintér Ádám[1], Tóth László[1], Gosztolya Gábor[1,2]

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

Kivonat Az automatikus beszédfelismerés területén az akusztikus modellezésben gyakorlatilag egyeduralkodókká váltak a mély neurális hálók. Az irodalomban számos megoldást találunk arra, hogy hogyan érdemes beállítani a különböző paramétereket a DNN akusztikus modellek tanítása során, azonban általában kevés figyelmet szentelnek annak, hogy a hálók súlyait hogyan érdemes inicializálni. Eközben a gépi tanulási irodalomban ez egy igen aktív terület; a közelmúltban több stratégia is napvilágot látott a DNN kezdősúlyainak beállítására. Jelen munkánkban három ilyen eljárást tesztelünk mély neurális hálós akusztikus modellekben, három különböző aktivációs függvényt (szigmoid, ReLU és szoftplusz) használva. Eredményeink alapján mindenképp érdemes valamilyen speciális súlyinicializálási eljárást alkalmaznunk, ugyanakkor a három vizsgált stratégia (Glorot, He és Edge of Chaos) használatával elért fonémaszintű hibaarányok között nem találtunk szignifikáns különbséget.
Kulcsszavak: beszédfelismerés, mély neurális hálók, súlyinicializálás, Glorot inicializálás, He inicializálás, Edge of Chaos

1. Bevezetés

Az elmúlt évtizedben a mély neurális hálók (Deep Neural Networks, DNN) nagyon gyorsan elterjedtek a gépi tanulás szinte minden területén. Az automatikus beszédfelismerésben is gyakorlatilag egyeduralkodókká váltak az akusztikus modellezés részfeladatán, mely elsősorban az általuk elérhetővé váló alacsony hibaarányoknak köszönhető. A beszédfelismerési feladatban ugyanakkor számos olyan részprobléma található, melyre valamilyen speciális algoritmus használata terjedt el (pl. kapcsolt állapotok létrehozása, vagy az akusztikus modell felvételszintű annotációira optimalizáló tanítási eljárások), és ezek neurális hálókra adaptálása folyamatosan zajlik (Grósz és mtsai, 2015; Zhu és mtsai, 2015; Grósz és mtsai, 2017). Emellett a hálók tanítása számos új hiperparaméter behangolását és az akusztikumra fókuszáló speciális tanítási technikák vagy metódusok kifejlesztését is magával vonta (ilyen pl. a Connectionist Temporal Classification (Graves és mtsai, 2006)).

Jelen cikkünkben is a DNN-tanítás egy „hiperparaméterére” fókuszálunk: azt vizsgáljuk meg, hogy a mély neurális hálók mennyire érzékenyek a súlyok kezdeti

értékeire. Habár a súlyokat mindig valamely valószínűségi eloszlást követve választjuk véletlenszerűen, az ezen eloszlást meghatározó paraméterek (jellemzően a szórás) kiválasztására számos stratégiát mutattak be az elmúlt években, és általánosságban is igen aktívan kutatott terület (ld. pl. (He és mtsai, 2015; Poole és mtsai, 2016; Schoenholz és mtsai, 2017; Pennington és mtsai, 2017; Hanin és Rolnick, 2018; Pretorius és mtsai, 2018)). Tudomásunk szerint nem született még olyan tanulmány, amely különböző súlyinicializálási eljárások fonéma- vagy szószintű hibaarányait vizsgálta volna az automatikus beszédfelismerés problémakörében. Vizsgálatunk aktualitását növeli, hogy a közelmúltban jelent meg az Edge of Chaos (röviden EOC, (Hayou és mtsai, 2019)) súlyinicializálási eljárás, mely kifejlesztői szerint lehetővé teszi extrém mély neurális hálók tanítását is. Bár jelen tanulmányunkban nem kísérlünk meg ilyen extrém struktúrájú DNN-alapú akusztikus modellt tanítani, egy ilyen súlyinicializáló eljárás akár alacsonyabb hibaarányokhoz is vezethet.

2. Mély neurális hálók súlyinicializálási stratégiái

A neurális hálókat jellemzően egy iteratív hibavisszaterjesztési (backpropagation) eljárással szokás tanítani. Az eljárásról ugyanakkor ismert, hogy nem garantálja a globális optimumot, hanem lokális optimumhoz vezet. Több rejtett réteg esetén (és a hagyományos szigmoid vagy tanh aktivációs függvényeket alkalmazva) ráadásul föllép a „megszűnő gradiens” (vanishing gradient, (Hochreiter és mtsai, 2001)) néven ismert effektus, mely azt eredményezi, hogy a túl nagy értékű súlyok miatt a mélyebben elhelyezkedő rétegek súlyai nem változnak érdemben (azaz a háló nem tanul). Túl kis súlyok esetén pedig, mivel a tanh és szigmoid függvények 0 körül gyakorlatilag lineárisak, elveszítjük a modell nemlinearitását, valamint a gradiensek „elszabadulhatnak” (exploding gradient). Emiatt létfontosságú, hogy a súlyokat a megfelelő intervallumban tartsuk, illetve onnan is indítsuk.

A következőkben részletesebben ismertetünk három eljárást a kezdősúlyok meghatározására. Viszonyítási alapnak azt tekintettük, hogy a súlyokat egy normális vagy egyenletes eloszlásból vettük, 0 átlaggal. Az értékek szórását ekkor előzetes tesztekkel $0,001$-ben határoztuk meg.

2.1. Glorot súlyinicializálási eljárása

A bevezetőjéről elnevezett Glorot-féle (vagy Xavier-féle) stratégia alapötlete, hogy az egyes rétegek kimeneteinek varianciáját azonos értéken tartsa, hogy az ne csökkenjen, ahogy a hiba visszaterjesztése a háló mélyebben fekvő rétegei felé halad (Glorot és Bengio, 2010). Mivel egy teljes kapcsolású (fully connected) hálóban minden neuron kapcsolatban áll az előző és a következő réteg összes neuronjával, a metódus szerint a súlyok szórása a következőképpen alakul:

$$\sigma = \sqrt{\frac{2}{n_{bemenet} + n_{kimenet}}} \tag{1}$$

míg az átlag 0. Amennyiben a súlyokat normális helyett egyenletes eloszlás szerint választjuk, azoknak praktikusan az

$$U = \Big[-\frac{\sqrt{6}}{\sqrt{n_{bemenet} + n_{kimenet}}}; \frac{\sqrt{6}}{\sqrt{n_{bemenet} + n_{kimenet}}} \Big] \tag{2}$$

intervallumból kell jönniük.

2.2. He súlyinicializálási eljárása

Glorot és Bengio cikke idején az elterjedt aktivációs függvények a szigmoid és a tanh függvények voltak, melyek nulla körül szimmetrikusak és deriváltjuk megközelítőleg egy (azaz a függvény lineáris). Ezt kihasználva elhanyagolhatták a levezetésből az aktivációs függvény alkalmazását. Ez a lépés azonban a később elterjedt függvények (pl. ReLU) esetén nyilvánvalóan nem megalapozott. Az előző számítások adaptálását végezték el He és munkatársai (He és mtsai, 2015). Eredményeik alapján normális eloszlás használata esetén 0 átlaggal és az alábbi szórással kell kiválasztanunk a súlyokat:

$$\sigma = \sqrt{\frac{2}{n_{bemenet}}}. \tag{3}$$

Egyenletes eloszlást használva a kezdősúlyok intervalluma a következő lesz:

$$U = \Big[-\sqrt{\frac{6}{n_{bemenet}}}; \sqrt{\frac{6}{n_{bemenet}}} \Big]. \tag{4}$$

2.3. Edge of Chaos

A „Káosz határa” (Edge of Chaos, EOC) inicializálási stratégia más megközelítésen alapszik. Az alapötlet az, hogy egy csupa véletlenül inicializált súlyt tartalmazó, teljesen kapcsolt mély neurális hálón különböző bemeneti értékekre megvizsgálva az előálló kimeneteket elvárjuk, hogy a bemenő információ *valamilyen mértékben* megjelenjen a kimenetekben. Ehhez a szerzők azt vizsgálták, hogy a bemeneti vektorok *párjai*, valamint a hozzájuk tartozó kimeneti értékek mennyire korrelálhatnak. A súlyok bizonyos eloszlásai a „rend fázisába” tartoznak, melyekre igaz, hogy minden bemeneti párhoz tartozó kimenetek aszimptotikusan korreláltak, és így ezek eltűnő gradienshez vezethetnek. Ezzel szemben más súlyeloszlások a „kaotikus fázisba” sorolódnak (ahol a megfelelő kimenetek aszimptotikusan dekorreláltak, és fölrobbanó gradienshez vezethetnek) (Poole és mtsai, 2016). A két eloszláshalmazt elválasztó határ a „káosz határa”, és kívánatos a kezdősúlyainkat egy ilyen eloszlás szerint választanunk (Schoenholz és mtsai, 2017).

A föntieket vitte tovább Hayou és munkatársai cikke (Hayou és mtsai, 2019), melyben elsősorban a különlegesen mély hálókra (10-200 rejtett réteg) koncentráltak. Megmutatták, hogy a korábbi inicializálási eljárások ilyen mélységben már nem vezetnek konvergenciához. Javaslatuk az volt, hogy a súlyok eloszlását

úgy kell megválasztani, hogy azok a „káosz határára" essenek, ami azt is jelenti, hogy (továbbra is 0 átlagon tartva azokat) a szórás értékét minden aktivációs függvényhez egyedileg (valamint a biasok szórásához igazodva) kell meghatározni. Glorot és He módszereivel összhangban a kapott értékeket továbbra is el kell osztani az adott réteg bemeneteinek számának négyzetgyökével. Kísérleti eredményeik alapján ez az eljárás lehetővé tette akár 200 rejtett réteget tartalmazó háló tanítását is tanh és ReLU aktivációs függvényekkel (Hayou és mtsai, 2019), ugyanakkor a módszer könnyűszerrel alkalmazható más függvényekre is.

3. A kísérletek technikai paraméterei

3.1. A tesztelt aktivációs függvények

Kísérleteinkben három aktivációs függvényt alkalmaztunk. Az első a hagyományos **szigmoid** függvény volt, melynek képlete

$$sig(x) = \frac{1}{1 + e^x}. \tag{5}$$

A következő aktivációs függvény, mely szintén igen elterjedt mind a beszédtechnológia, mind általában a gépi tanulás területén, a rectifier (vagy **ReLU**) függvény:

$$ReLU(x) = \max(x, 0). \tag{6}$$

Végül teszteltük a **szoftplusz** aktivációs függvényt is, melyet szokás a ReLU függvény folytonosan deriválható közelítésének is tartani (Dugas és mtsai, 2001):

$$softplus(x) = \log(1 + e^x). \tag{7}$$

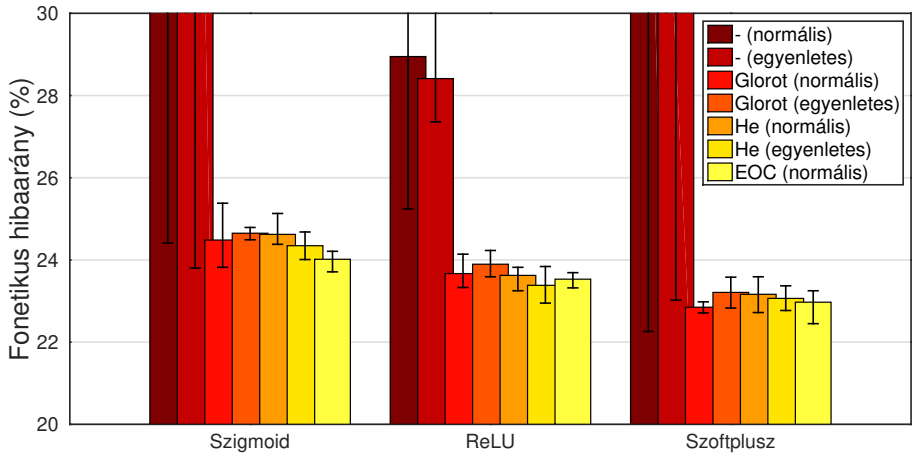
A kimeneti rétegben minden esetben a szoftmax függvényt alkalmaztuk.

3.2. A TIMIT adatbázis

Kísérleteinket az angol nyelvű TIMIT beszédadatbázison végeztük (Lamel és mtsai, 1986), mely relatíve kis mérete (kb. 3 óra) ellenére még mindig gyakran használt. A súlyokat a 3696 felvételből álló tanítóhalmaz közelítőleg 90%-án (3342 felvételen) tanítottuk, a fennmaradó 354 felvétel pedig a tanítási ráta vezérlésében kapott szerepet (*learn rate scheduling*). Mivel nem volt hangolandó hiperparaméterünk, a kiértékelést közvetlenül a 192 felvételből álló „mag" (core) teszthalmazon végeztük. Kiértékelés előtt a fonémacímkéket a bevett gyakorlatnak megfelelően 39 kategóriába vontuk össze (Lee és Hon, 1989).

3.3. DNN-paraméterek

Akusztikus neurális hálóink 5 rejtett réteget tartalmaztak, minden rejtett rétegben 1024 neuronnal. Bemenetként keretszinten egy 40 sávos mel-szűrőkészlet



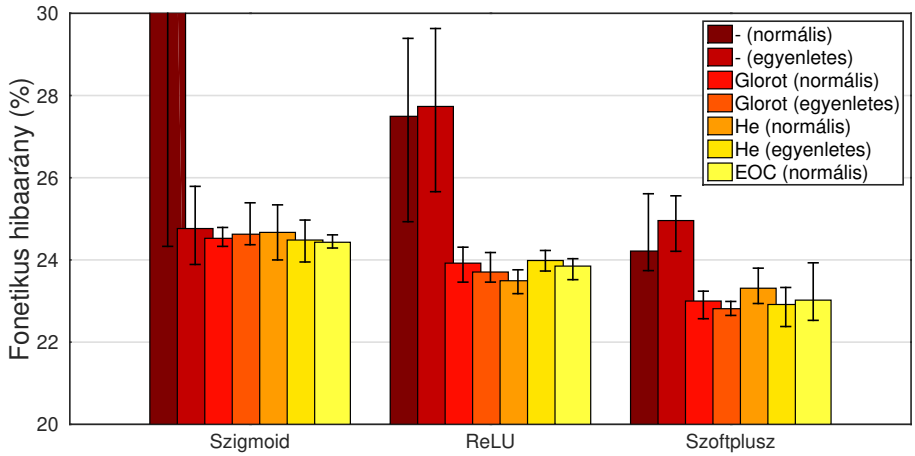
1. ábra: A különböző súlyinicializáló megközelítésekkel elért átlagos fonetikai hibaarányok **regularizáció nélkül** a TIMIT adatbázis „mag" teszthalmazán.

energiakimeneteit használtuk, a szokásos első és második derivált értékeivel kiegészítve; minden kerethez felhasználtuk a mindkét oldali szomszédos 8-8 keret jellemzővektorait is, így a hálók 2091 bemeneti neuronnal rendelkeztek. Kimenetként 858 kontextusfüggő kapcsolt állapotot használtunk, ennek megfelelő számú kimeneti neuronnal. A keresési lépést a HTK programcsomag (Young és mtsai, 2006) egy módosított változatával végeztük el.

Mivel kíváncsiak voltunk arra is, hogy az egyes inicializálási eljárások mennyire igénylik a tanítás során valamilyen regularizációs technika alkalmazását, minden kísérletünket megismételtük L2 regularizációval is. Mivel kísérleteink tárgya alapvetően a *véletlen* súlyinicializálási stratégiák hatékonysága volt, releváns volt a kapott eredmények stabilitása is, ezért minden konfigurációra öt különböző modellt tanítottunk (eltérő random seedek használatával).

4. Eredmények

A **regularizáció nélkül** elért átlagos fonémaszintű hibaarányok az 1. ábrán láthatók; a képen feltüntettük az öt tanított modell közül a legjobb és a legrosszabb eredményét is. Látható, hogy amennyiben standard súlyinicializálást használunk, az eredmények elég rosszak: az öt tanított modellből a szigmoid aktivációs függvényt alkalmazva csak két-két (normális és egyenletes eloszlású kezdősúlyok), míg a szoftplusz függvény esetén csak három és két modell tanult egyáltalán, így adódott az átlagos hiba ilyen magasnak. A ReLU függvény esetében ennél kedvezőbb volt a helyzet, de kompetitívnek ekkor sem tekinthetjük: normális eloszlású kezdősúyok esetén négy modell hibája $28,8 - 30,7\%$ közé esett, és csak egy esetben kaptunk elfogadható teljesítményt ($25,2\%$-os fonéma-hibaarány), míg egyenletes eloszlású súlyoknál mind az öt modell $27,3\%$ és $31,2\%$ közé eső fonetikai hibaarányokhoz vezetett.



2. ábra: A különböző súlyinicializáló megközelítésekkel elért átlagos fonetikai hibaarányok **L2 regularizációval** a TIMIT adatbázis „mag” teszthalmazán.

A fönnmaradó három tesztelt súlyinicializálási eljárás (Glorot, He és EOC) esetében azt látjuk, hogy nem szükséges a súlyok L2 regularizációja ahhoz, hogy használható fonémafelismerési teljesítményt kapjunk: minden esetben $23-24\%$ körüli átlagos fonetikai hibaarányokat tapasztaltunk. Az Edge of Chaos eljárás ugyan stabilan a legjobb két modell között volt, de a különbség nyilvánvalóan nem szignifikáns. A módszer előnye lehet ugyanakkor, hogy a szigmoid és a ReLU függvények esetében az öt tanított modell egymáshoz nagyon hasonló teljesítményhez vezetett. Ez azonban igen korlátozott előnynek tűnik, egyrészt mert ez pont az egyébként legalacsonyabb hibarátákhoz vezető szoftplusz aktivációs függvény esetében nem teljesült, másrészt mert a tanított modellek teljesítménye közötti különbség nagyobb tanítóadatbázis használata esetén eltűnhet.

Különbségek inkább az egyes aktivációs függvények esetében adódtak: látható, hogy a szigmoid függvények helyett érdemes a ReLU, de még inkább a szoftplusz függvényt alkalmazni. Természetesen az, hogy L2 vagy más regularizáció (pl. dropout) használata nélkül is lehetségesnek bizonyult egy öt rejtett rétegből álló neurális háló tanítása, már önmagában is érdekes tapasztalat (habár a javasolt súlyinicializáló eljárások motivációja éppen ez volt).

Az **L2 regularizáció használatával** elért átlagos fonémaszintű hibaarányokat a 2. ábrán tüntettük föl (ismét a legjobb és legrosszabb modellek teljesítményével együtt). Látható, hogy a regularizáció használata lehetővé tette a standard súlyinicializáló eljárás használatát a szoftplusz aktivációs függvény esetében is; a másik két függvény esetén azonban a helyzet nem változott (azaz a szigmoidnál még mindig használhatatlan, a ReLU esetében pedig még mindig egyszerűen csak rossz eredményeket kaptunk). Más tekintetben nagyon hasonlóak az eredmények a súlyregularizáció nélkül elértekhez. Véleményünk szerint ez azt jelenti, hogy (legalábbis a DNN akusztikus modellek esetén megszokott méretű hálók esetén) a súlyok megfelelően megválasztott kezdőértékei mellett a

gyakorlatban nincs szükség a tanítás során további regularizációra sem a vanishing gradient, sem az exploding gradient effektus elkerüléséhez. Természetesen a későbbiekben tervezzük ezt a következtetésünket mind nagyobb adatbázisokon, mind mélyebb hálók használata esetén ellenőrizni.

Inicializálás módja		**Szigmoid**		**ReLU**		**Szoftplusz**	
		—	L2	—	L2	—	L2
—	Normális	67.0%	66.6%	28.9%	27.5%	52.3%	24.2%
	Egyenletes	66.7%	24.8%	28.4%	27.7%	37.9%	25.0%
Glorot	Normális	24.5%	24.5%	23.7%	23.9%	22.8%	23.0%
	Egyenletes	24.7%	24.6%	23.9%	23.7%	23.2%	22.8%
He	Normális	24.6%	24.7%	23.6%	23.5%	23.2%	23.3%
	Egyenletes	24.3%	24.5%	23.4%	24.0%	23.1%	22.9%
EOC	Normális	24.0%	24.4%	23.5%	23.8%	23.0%	23.0%

1. táblázat. *A különböző megközelítések által elért átlagos fonetikai hibaarányok a TIMIT adatbázis „mag” teszthalmazán.*

Az átlagos fonetikai hibaarányokat az 1. táblázatba is kigyűjtöttük. Látható, hogy (a viszonyítási alapként szolgáló inicializálástól eltekintve) a hibaértékeket elsősorban az aktivációs függvény határozza meg: szigmoid esetén a $24,0-24,7\%$, ReLU esetén a $23,4-24,0\%$, szoftplusz esetén pedig a $22,8-23,3\%$ intervallumba estek.

5. Összegzés

Jelen tanulmányunkban különböző kezdősúly-meghatározási stratégiákat hasonlítottunk össze mély neurális hálós akusztikus modellek esetében. Vizsgálatainkban három különböző eljárással határoztuk meg a súlyok véletlen (normális, illetve egyenletes) eloszlásának szórását, míg annak átlagát minden esetben nullára állítottuk. Tesztjeink során három különböző aktivációs függvényt is megvizsgáltunk. Az előálló fonetikai hibaarányok alapján úgy véljük, érdemes valamelyik megvizsgált stratégiát alkalmazni a tanítás előtt, ugyanakkor az egyes eljárások pontossága között nem találtunk markáns különbségeket.

Tanulmányunkban azt is vizsáltuk, hogy a súlyok L2 regularizációja milyen hatással van a tanított DNN-ek teljesítményére. Tapasztalataink szerint amennyiben akár a Glorot, akár a He, akár az Edge of Chaos inicializálást alkalmazzuk, a súlyok tanítás közbeni regularizációja elhagyható. Ugyanakkor fontosnak érezzük megjegyezni, hogy kísérleteink egy kisméretű beszédadatbázison (a TIMIT-en) történtek; reálisnak tartjuk, hogy több tanító adat használatával az azonos paraméterekkel, csak eltérő random seeddel tanított modellek teljesítménye még jobban közelítsen egymáshoz. Egy másik érdekes lehetséges kutatási iránynak tartjuk a jelenleg általánosan használtnál lényegesen mélyebb (10-20,

akár 50-100) rejtett rétegű DNN akusztikus modellek tanítását, ez azonban szintén további vizsgálatokat igényel.

Köszönetnyilvánítás

Jelen kutatás eredményei az „Integrált kutatói utánpótlás-képzési program az informatika és számítástudomány diszciplináris területein" című, EFOP-3.6.3-VEKOP-16-2017-0002 számú projekt támogatásával készültek. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg. A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (FK 124413), részben pedig az Innovációs és Technológiai Minisztérium (ITM 2018-1.2.1-NKP-2018-00008 és TUDFO/47138-1/2019-ITM) is támogatta. Gosztolya Gábor és Tóth László kutatásait az MTA Bolyai János Kutatási ösztöndíja és az Új Nemzeti Kiválóság Program Bolyai+ pályázata (azonosítók: ÚNKP-19-4-SZTE-51) támogatta.

Hivatkozások

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., Garcia, R.: Incorporating second-order functional knowledge for better option pricing. In: Advances in Neural Information Processing Systems (2001)

Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Machine Learning Research. pp. 249–256 (2010)

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: International Conference on Machine Learning. pp. 369–376. Pittsburgh, PA, USA (2006)

Grósz, T., Gosztolya, G., Tóth, L.: Környezetfüggő akusztikai modellek létrehozása Kullback-Leibler-divergencia alapú klaszterezéssel (in Hungarian). In: MSZNY. pp. 174–181. Szeged (2015)

Grósz, T., Gosztolya, G., Tóth, L.: Mély neuronhálós beszédfelismerők gmm-mentes tanítása (in Hungarian). In: MSZNY. pp. 170–180. Szeged (2017)

Hanin, B., Rolnick, D.: How to start training: The effect of initialization and architecture. In: Neural Information Processing Systems. Montréal, Kanada (2018)

Hayou, S., Doucet, A., Rousseau, J.: On the impact of the activation function on deep neural networks training. In: International Conference on Machine Learning (2019)

He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: IEEE International Conference on Computer Vision. pp. 1026–1034 (2015)

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Kremer, S.C., Kolen, J.F. (szerk.) A Field Guide to Dynamical Recurrent Neural Networks (2001)

Lamel, L., Kassel, R., Seneff, S.: Speech database development: Design and analysisof the acoustic-phonetic corpus. In: DARPA Speech Recognition Workshop. pp. 100–109 (1986)

Lee, K., Hon, H.: Speaker-independent phone recognition using Hidden Markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing 37(11), 1641–1648 (1989)

Pennington, J., Schoenholz, S.S., Ganguli, S.: Resurrecting the sigmoid in deep learning throughdynamical isometry: Theory and practice. In: Neural Information Processing Systems. Long Beach, CA, USA (2017)

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., Ganguli, S.: Exponential expressivity in deepneural networks through transient chaos. In: Advances in Neural Information Processing Systems. pp. 3360–3368 (2016)

Pretorius, A., Van Biljon, E., Kroon, S., Kamper, H.: Critical initialisation for deep signal propagation in noisy rectifier neural networks. In: Neural Information Processing Systems. Montréal, Kanada (2018)

Schoenholz, S.S., Gilmer, J., Ganguli, S., Sohl-Dickstein, J.: Deep information propagation. In: International Conference on Learning Representations (2017)

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)

Zhu, L., Kilgour, K., Stüker, S., Waibel, A.: Gaussian free cluster tree construction using Deep Neural Network. In: Interspeech. pp. 3254–3258. Drezda, Németország (Sep 2015)