End-to-end és hibrid mélyneuronháló alapú gépi leiratozás magyar nyelvű telefonos ügyfélszolgálati beszélgetésekre

Mihajlik Péter[1,2], Balog András[2], Tarján Balázs[1,3], Fegyó Tibor[1,3]

[1] Budapest Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék, 1111, Budapest, Műegyetem rkp. 3.
mihajlik@tmit.bme.hu
[2] THINKTech Nonprofit Kft., 2600, Vác, Váczy Pál u. 15.
abalog@thinktech.hu
[3] SpeechTex Kft, 1181, Madách Imre u. 47.
{tarjan, fegyo}@speechtex.com

Kivonat A tisztán mélyneuronhálóra épülő gépi beszédfelismerés alig pár éve került a tudományos köztudatba, de máris az egyik leginkább kutatott szakterületté vált. Magyar nyelvre történő alkalmazása, illetve annak publikációja azonban ez idáig váratott magára. Cikkünkben beszámolunk az első tapasztalatainkról a téren, magyar nyelvű telefonos ügyfélszolgálati beszéd leiratozása témában. A valós idejű működés érdekében nagy számú, egy dimenziós, idő- és csatornatengely szerint szeparált konvolúciós réteget alkalmazunk reziduális kapcsolatokkal és normalizációkkal. Karakter akusztikus modellekkel, szótár és nyelvmodell nélkül is összevethető, bár gyengébb szó- és betűpontossági eredményeket kaptunk a hibrid (rejtett Markov-modell + mélyneuronháló) rendszeréhez képest. Nyelvmodell hozzáadásával és angol nyelven előtanított súlyokkal történő inicializálás alkalmazásával szignifikáns javulást mértünk, meghaladva a hibrid eredményeket. Kutatásunk megerősíti, hogy az end-to-end beszédfelismerési modellezés magyar nyelven is életképes, azonban a teljes potenciál eléréséhez valószínűleg nagyságrendi növekedésre lenne szükség az akusztikus tanítóadatok (hang + leirat) terén.

1 Bevezetés

Történelmi távlatból nézve a beszédfelismerés hatékonysága, használhatósága szempontjából mindig az adat (hullámforma + leirat) jelentette a kulcsot. A dinamikus idővetemítés sikere annak volt köszönhető, hogy intuitív frekvenciaelemzés helyett valós referenciabeszéd-felvételekhez hasonlította a felismerendő bemondást (Sakoe és Chiba, 1978). A beszélőfüggetlenséget a nagy mennyiségű beszédadatot felhasználni képes, már valódi gépi tanulás alapú rejtett Markov-modell (Jelinek és mtsai, 1975) tette lehetővé. A folyamatos beszéd szöveggé alakításában pedig annak felismerése volt a kulcslépés, hogy az egyes szósorozat-hipotézisekhez az a-priori valószínűségeket –

n-gram nyelvi modelleken keresztül – a témához illeszkedő nagy mennyiségű szöveg alapján becsülhetjük (Kunh és De Mori, 1990).

Ebbe a trendbe illeszkedik a mélytanulás: erősen leegyszerűsítve úgy is mondhatjuk, hogy a mély neuronhálók fő erénye a (sokkal) több adat (sokkal) hatékonyabb hasznosítása, melyhez persze némi célzott hardware támogatás is (GPU) szükségeltetik.

Miután mind az akusztikus, mind a nyelvi modellek terén a mélyneuronhálók alkalmazása vált egyeduralkodóvá a state-of-the-art rendszerekben, valamint a graféma alapú akusztikus modellek már korábban is jól teljesítettek (Mihajlik és mtsai, 2007), észszerű irányként merült fel a bemenő akusztikus megfigyelések közvetlenül neuronhálóval történő leképezése szó/karakter szekvenciává. Az end-to-end – azaz elejétől végéig mély neuronháló alapú – megközelítés nem hozott azonnali áttörést (Graves és Jaitly, 2014), inkább letisztultságával tűnt ki, azonban rövid idő alatt rendkívül kutatott és sikeres területté vált. Angol nyelv esetén a mai napig folyamatos versenyben vannak a hibrid, mély tanulást és HMM struktúrát is a használó, ill. a tisztán mélytanulási technikák – a cikk írása idején a Switchboard korpuszon éppen (Tüske és mtsai, 2020) érte el a legjobb eredményt egy viszonylag egyszerű end-to-end struktúrával, míg a LibriSpeechen (Pan és mtsai, 2020) vezet hibrid megközelítéssel.

Így indokoltnak láttuk egy releváns, telefonközpontos beszélgetéseket tartalmazó magyar nyelvű adatbázison megvizsgálni a két beszédfelismerési irányzat egy-egy képviselőjének eredményeit.

2 A beszédfelismerési feladat

A magyar nyelvű, ügyfélszolgálati témakörű telefonos beszéd felismerése kiemelt jelentőséggel bír napjainkban. Így a rendelkezésünkre álló összes, nem publikus, anonimizált telefonközponti leiratozott beszélgetést felhasználva alakítottuk ki a tanító, validáló és kiértékelő halmazokat az 1. táblázat szerint.

1. Táblázat: Telefonos ügyfélszolgálati beszédadatbázis részhalmazai

	Tanító	Validáló	Kiértékelő
Audio (óra)	290	7	12
Text (szószám)	3.4M	46K	66K

A beszélgetések valós, spontán, ügyfélszolgálati beszélgetéseket tartalmaznak, olykor jelentős háttérzajjal terheltek. A tanító, validáló és kiértékelő halmazokra bontásnál alapfeltétel volt, hogy egy beszélgetés csak egy halmazba kerülhet. Míg a validáló és kiértékelő halmazokba csak teljes, vágatlan beszélgetéseket tettünk, a tanító halmazból a beszélgetések egyes részei (pl. a túl zajos szakaszok, vagy a lejegyző által nem értett részek) kikerülhettek.

A beszédfelismerési folyamatra megkötés, hogy csak valós időben alkalmazható technológiák jöhetnek szóba (így pl. két körös beszédfelismerés, vagy BLSTM struktúra sem). További szempont az alacsony késleltetés és a nagy offline áteresztőképesség (egy feldolgozási időegység alatt minél több bejövő beszédidőegység feldolgozása).

2 Vizsgált megközelítések

2.1 TDNN-F – HMM hibrid akusztikus modell

Klasszikus rejtett Markov-modell alapú megközelítés, ahol környezetfüggő (bal difón) fonéma akusztikus modelleket (AM) használunk és az egyes HMM állapotokhoz tartozó valószínűség-sűrűség függvényeket modellezzük mély neuronhálóval. A GMM-es előtanítást követően, annak időszegmentálását felhasználva tanítjuk be a faktorált Time Delay Neural Network (Povey és mtsai, 2018) architektúrát (a továbbiakban: TDNN-F). A tanítás lépései és az alkalmazott módszerek megegyeznek a Kaldi[1] LibriSpeech s5 receptúrájával – azzal a megkötéssel, hogy az általános használhatóság kedvéért a beszélőcímkéket igénylő i-vektorokat nem alkalmazzuk. Bemeneti vektorokként nagy felbontású MFCC-t használtunk, az audio adatok dúsítására (augmentálására) a felvételek sebességét és hangerejét perturbáltuk. A TDNN-F paraméterszáma hozzávetőlegesen 18.5M volt.

2.2 Idő- és csatornatengely mentén szeparált 1D konvolúciósháló-alapú end-to-end akusztikus modell

Connectionist Temporal Classification (CTC) költségfüggvénnyel (Graves és mtsai, 2006) tanított, teljes kiépítésben 78 rétegű 1D konvolúciós neuronháló (Kriman és mtsai, 2020) karakterszintű kimenettel. A paramétertér csökkentése érdekében a konvolúció szétválik csatornánkénti FIR (Finite Impulse Response) szűrésre, majd a szűrt csatornák lineáris kombinációjára ill. a nemlinearitás alkalmazására. Normalizálásnak a batchnorm-ot használtuk, aktivációs függvénynek a ReLU-t. A gradiens visszaterjesztés elősegítésére az 5-ös blokkokra bontott konvolúciós rétegeket átívelő reziduális kapcsolatot alkalmaztunk. A reprodulkálhatóság kedvéért a standard receptúrát (Kriman és mtsai, 2020) követtük itt is. A bemenő akusztikus adatok Mel-skálázott rövid idejű amplitúdó spektrumok voltak. SpecAugment-et (Park és mtsai, 2019) minden esetben, hangerő és beszédsebesség perturbációt opcionálisan alkalmaztunk. A mélyneuronháló össz. paraméterszáma 18.9M.

2.3 Nyelvmodellek

Alapértelmezésben hagyományos, szó alapú back-off 4-gram nyelvmodellt (LM: Language Model) alkalmaztunk a szokásos, módosított Kneser-Ney (Chen és Goodman, 1999) simítással. A magyar nyelvhez jobban illeszkedő morf nyelvi modellt is kipróbáltunk, további részletek a neurális tudástranszfer alapú nyelvmodellezésről a (Tarján és mtsai, 2020)-ban találhatók – mi a jelen kutatásban a kisebb (1 GB) memóriaigényű augmentált morf modellt használtuk.

[1] https://github.com/kaldi-asr/kaldi

3 Kísérleti eredmények

3.1 Kísérleti elrendezés

End-to-end esetben lehetőség van tisztán akusztikai alapon történő leiratozásra. Ennek legegyszerűbb módszere a „greedy" algoritmus: CTC módszerrel tanított neurális modellek kimeneteiből keretszinkron módon a legvalószínűbb karaktert kiválasztjuk, a közvetlenül egymás után ismétlődőket egyetlen karakterrel helyettesítjük, majd a „blank" karaktert kiszedve összeolvassuk az eredményt. Természetesen a szóköz a tanításnál kötelező a szavak között, így a felismerési fázisban a szóhatárok természetes módon visszaállítódnak (helyes felismerés esetén).

HMM-es beszédfelismerésnél a súlyozott kiejtési alternatívákat, a fonemikus környezetfüggőséget és a nyelvmodellt WFST keretrendszerben (Mohri és mtsai, 2002) integráljuk és optimalizáljuk, majd a szokásos beam-search eljárással választjuk ki – szintén keretszinkron módon – a (Viterbi közelítéssel mért) legvalószínűbb szósorozat-hipotézist.

A korrekt összehasonlíthatóság érdekében a nyelvmodell end-to-end akusztikus modellel történő kombinációját a HMM-es elrendezéssel azonos módon, ugyanazon beam-search dekódolóval végezzük. Ekkor a karakter kiejtési szótár formális, csupán a „blank" karakterek beékelődésire kell felkészíteni, ill. nincs fonológiai értelemben vett környezetfüggés.

A dekódolást mindig telítésközeli munkapontban végeztük, így a dekódolási sebesség a valós időnél még mindig kb. 70-szer gyorsabb. Ebbe az akusztikus hasonlóságot számoló neuronhálók „inference" számítását nem értjük bele, ami GTX 1080 TI GPU alkalmazásával a valós időtől kb. 60-szor gyorsabb.

Az optimális nyelvi és akusztikus modell súlyozást a validáló halmazon végeztük.

3.2 Kalibrációs tesztek

Az első lépés a kijelölt megközelítések ellenőrzésére, hogy ismert angol nyelvű beszédfelismerési feladaton lemérjük a pontosságukat. A publikus LibriSpeech adatbázist Panayotov és mtsai, 2015) választottuk, ennek teljes tanítóanyagán (960 óra) tanítottunk mindkét esetben, illetve a standard „test clean" halmazon értékeltük ki az eredményeket.

A TDNN-F modelleket az egyik standard („3-gram ARPA LM, pruned with theshold 1e-7") nyelvi modellel értékeltük ki. A szófelismerési hiba (Word Error Rate) a 2. táblázatban látható, marginálisan – jobb, mint a Kaldi saját közlése (5.3%).

2. Táblázat: Kalibrációs eredmények a LibriSpeech (960 óra) adatbázison

AM	LM	WER (test clean)
TDNN-F	word	5.24%
end-to-end	–	5.20%
end-to-end	word	3.78%

Az end-to-end rendszert a rendelkezésre álló szűkösebb GPU memóriakapacitások miatt (4 x GTX 1080 TI / 11GB) kisebb batch mérettel (64) és kevesebb epoch számmal (200) tanítottuk, mint (Kriman és mtsai, 2020) ajánlják, azonban így is jobb eredményt ért el, mint a hibrid rendszer lexikon és nyelvmodell nélkül is, nyelvmodell hozzáadásával pedig előnye szignifikánsan megnőtt. Ezek ellenére, a (Kriman és mtsai, 2020) által közölt pontosságot (3.9% nyelvmodell nélkül) nem értük el, aminek oka a rövidebb tanítás, a beszédsebesség-perturbáció hiánya, valamint a kevésbé kimerítő hiperparaméter-optimalizálás lehetett. Mindazonáltal, a „kalibrációt" sikeresnek tekintettük, mind a hibrid, mind a tisztán neuronháló alapú megközelítés működik, hozza az elvárt eredményeket.

3.3 Telefonos ügyfélszolgálati beszédfelismerési teszteredmények

A telefonos ügyfélszolgálati beszédadatok akusztikus modelltanításra előkészítése jelentette az első feladatot. Itt különféle előszegmentálások és szűrések után gyakorlatilag kereken 200 óra tanítóanyag maradt közvetlenül felhasználható a tanításra. Mindkét alább vizsgált megközelítésnél ugyanezt a tanítóanyagot használtuk és mindenhol közöljük a szóhibaarány (WER) mellett a betűhibaarányt (LER) is.

A TDNN-F modelleket az előzőek szerint tanítottuk, mindössze azzal a különbséggel, hogy a bemenetünk most 8kHz-es mintavételezésű, így az előfeldolgozást ennek megfelelően a Kaldi Switchboard receptúrája szerint (minimálisan) módosítottuk. A kiértékelő teszthalmazon mért beszédfelismerési eredmények a 3. táblázatban találhatók.

3. Táblázat: Beszédfelismerési eredmények magyar nyelvű telefonos ügyfélszolgálati adatokon

AM	LM	WER	LER
TDNN-F	word	21.40%	9.93%
TDNN-F	morf	18.96%	9.19%
end-to-end (baseline)	–	30.63%	12.52%
end-to-end (augment)	–	29.07%	12.71%
end-to-end (pretrain)	–	28.54%	11.81%
end-to-end (pretrain + la)	–	27.65%	12.36%
end-to-end (pretrain + la + augment)	–	26.07%	11.84%
end-to-end (pretrain + la + augment)	word	18.79%	9.56%
end-to-end (pretrain + la + augment)	morf	17.83%	9.15%

Az end-to-end modellek esetén első körben szintén 8kHz-es mintavételezésre állítottuk a megfelelő alacsonyszintű jelfeldolgozási paramétert és csak a learning-rate értékét optimalizáltuk, valamint a 200 epoch-os tanítási hosszat (egy hetes futásidő az ismertetett hardveren) megtartottuk (baseline). Ezután – a mindig bekapcsolt „on-the-fly" spectral augment mellett – beszédsebesség és hangerő perturbációt alkalmaztunk (augment). Majd 16kHz-es mintavételezést/adatkonvertálást beállítva, az angol nyelvű LibriSpeech-en az előző pontban betanított neurális hálózat súlyaival inicializáltuk a

tanítást (csak az encoder hálózatot). Így drasztikusan csökkentett, 15-ös epoch szám mellett is az előzőeknél jobb eredményt kaptunk (pretrain), amit az alacsony szintű beszédjellemzők betanulásának és az így megvalósult tudásátadásnak (transfer learning) tulajdonítunk. Végül hosszabb, 45-ös epoch esetén („la", mint long adaptation) további javulást tapasztaltunk (pretrain + la), valamint a korábbi augmentáció és hosszabb, angol nyelven előtanított súlyokkal inicializált tanítás esetén még további javulást (pretrain + la + augment).

Ahogy az 3. táblázat mutatja, a (greedy) end-to-end eredmények – főleg szóhibaarány tekintetében – némiképp elmaradnak a TDNN-F rendszeréhez képest. Ugyanakkor, összehasonlítva egy korábbi MSZNY konferencián bemutatott rendszerünk eredményével, ahol előrecsatolt DNN ACM-et alkalmaztunk (Tarján és mtsai, 2019), az end-to-end modellek úgy mutatnak javulást, hogy se kiejtési szótárt, se nyelvmodellt nem használnak.

Nyelvmodell hozzáadásával az end-to-end szóhibaarányok ugrásszerűen javulnak, a betűhibaarányok – a felismerési elvből fakadóan – érthetően kevésbé. Viszont mind WER, mind LER tekintetében sikerült áttörést elérni: az azonos nyelvmodellel mért hibrid és end-to-end eredmények közül minden esetben az utóbbiak bizonyultak jobbak. Ezzel együtt is, az end-to-end beszédfelismerési eredmények a hibrid TDNN-F rendszeréhez viszonyítva kisebb javulást hoztak a vártnál. Ennek elsődleges okaként a „kalibrációs" adathalmazhoz képesti jóval kisebb tanítóadatméretet tudjuk megjelölni.

Megjegyezzük, hogy az itt publikált hibrid mély-neuronhálós eredmények ugyan számszerűen a legalacsonyabbak, de a TDNN-F hibrid rendszer telefonos magyar ügyfélszolgálati nyelvre alkalmazása nem új, csupán a kiértékelésre használt referencialeirat tisztítása okozza a látszólagos javulást (Tarján és mtsai, 2020)-hoz képest.

4 Következtetések

Megvizsgáltuk, hogy a nagy áteresztőképességgel működő mélytanulásra épülő beszédfelismerési módszerek milyen pontosságot érhetnek el magyar nyelvű telefonos ügyfélszolgálati beszéd gépi leiratozásánál. Azt kaptuk, hogy a rendelkezésre álló akusztikus tanítóadatbázis-méret mellett az angol nyelven széles körben alkalmazott end-to-end megközelítés nyelvmodell alkalmazása nélkül is összevethető eredményeket ad a jelenleg legjobb hibrid megközelítéshez képest, a neurális tudástranszferrel készült morf nyelvmodell alkalmazásával pedig felül is múlja azt. Meggyőződésünk, hogy elsősorban az akusztikus tanítóadat mennyiségét szükséges növelni a még jobb, az angol nyelvű eredményekkel összemérhető pontosságértékekért. A nemzetközi eredményekkel összevetésben és jelen vizsgálatok alapján is úgy látjuk, jelenleg nem a HMM struktúra létén/nem létén múlik a pontosság, hanem sokkal inkább az alkalmazott mélytanulási módszereken. A magyar nyelv digitális fenntarthatósága érdekében tehát elsősorban a megfelelő tanítóadatok (hangfelvétel + leirat) nagyságrendi növelése lenne a cél, másodsorban pedig a kiszolgáló számítástechnikai infrastruktúra (korszerű GPU gridek) fejlesztése, például, hogy a hiperparaméterek érdemi optimalizálására is legyen reális lehetőség.

Hivatkozások

Chen, S. F. and Goodman, J.: "An empirical study of smoothing techniques for language modeling," Computer Speech & Language, vol. 13, no. 4, pp. 359–393 (1999)

Graves, A.; Fernández, S.; Gomez, F., Schmidhuber, J.: "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". ICML 2006, pp. 369–376, (2006)

Graves, A. and Jaitly, N.: "Towards End-To-End Speech Recognition with Recurrent Neural Networks." ICML (2014).

Jelinek, F.; Bahl, L.; Mercer, R.: "Design of a linguistic statistical decoder for the recognition of continuous speech". IEEE Transactions on Information Theory. 21 (3), pp. 250. (1975).

Kriman S. et al., "Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, , pp. 6124-6128, (2020)

Kuhn, R. and De Mori, R.: "A cache-based natural language model for speech recognition." IEEE Transactions on pattern analysis and machine intelligence 12.6: 570-583(1990)

Mihajlik, P., Fegyó, T., Tüske Z., and Ircing P.: "A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian," Interspeech'07, Antwerp, Belgium, (2007)

Mohri, M, Pereira, F. and Riley, M.: "Weighted Finite-State Transducers in Speech Recognition", Computer Speech and Language, 16(1), pp. 69–88, (2002)

Pan J., Shapiro J., Wohlwend J., Han K. J., Lei T., and Ma T., "ASAPP-ASR: Multistream CNN and Self-Attentive SRU for SOTA Speech Recognition," in Proc. INTERSPEECH, pp. 16–20. (2020)

Panayotov V., Chen G., Povey D., and Khudanpur S.,"Librispeech: an ASR corpus based on public domain audio books," in ICASSP, pp. 5206–5210 (2015)

Park, D. S. et al., "SpecAugment: A simple data augmentation method for automatic speech recognition", in Proc. Interspeech, (2019)

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S.: Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. Proc. Interspeech, 3743-3747. (2018)

Sakoe, H.; Chiba, S.: "Dynamic programming algorithm optimization for spoken word recognition". In: IEEE Transactions on Acoustics, Speech, and Signal Processing. 26 (1), pp. 43–49. (1978)

Tarján, B, Fegyó, T és Mihajlik, P.: Ügyfélszolgálati beszélgetések nyelvmodellezése rekurrens neurális hálózatokkal,in Proc MSZNY (2019)

Tarján, B, Szaszák G, Fegyó T, Mihajlik P: Improving Real-time Recognition of Morphologically Rich Speech with Transformer Language Model, in Proc 11th IEEE International Conference on Cognitive Infocommunications (2020)

Tüske, Z; Saon, G; Audhkhasi, K; Kingsbury, B.: Single headed attention based sequence-to-sequence model for state-of-the-art results on Switchboard, in Proc Interspeech (2020)