Autoenkóderen alapuló jellemzőreprezentáció mély neuronhálós, ultrahang-alapú némabeszéd-interfészekben

Pintér Ádám[1], Gosztolya Gábor[1,2], Tóth László[1], Grósz Tamás[1], Csapó Tamás Gábor[3,5], Markó Alexandra[4,5]

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[3]Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék
[4]Eötvös Loránd Tudományegyetem, Fonetikai Tanszék
[5]MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoport
{ ggabor, tothl, groszt } @ inf.u-szeged.hu
csapot @ tmit.bme.hu, marko.alexandra @ btk.elte.hu

Kivonat A neurális hálón alapuló némabeszéd-interfészek általában a teljes ultrahangkép alapján becslik meg a spektrális paramétereket, melyekből a vokóder aztán beszédet generál. Habár ez a megközelítés igen kézenfekvő, és tapasztalataink szerint érthető beszédet képes generálni, több hátránya is van: egyrészt nehezen ragadja meg az egymáshoz közel eső területek (gyakorlatilag a pixelek) közötti összefüggéseket, másrészt igen pazarló. Könnyen belátható, hogy a képpontok egy jelentős része irreleváns a spektrális paraméterek becslése szempontjából, a szomszédos képpontok által tárolt információ nagyon rendundáns, a mély háló mérete pedig nagy a sok jellemző miatt. Jelen cikkünkben ezen problémák kezelésére egy autoenkóder neurális hálót tanítunk az ultrahangképre, és a szintézishez szükséges spektrális paraméterek becslését az autoenkóder háló rejtett bottleneck rétegében található neuronok aktivációi alapján végezzük egy második mély hálóval. Kísérleti eredményeink alapján a javasolt eljárás hatékonyabb, mint a hagyományos megközelítés: a kapott átlagos négyzetes hibák minden esetben alacsonyabbak, a korrelációértékek pedig magasabbak voltak, mint a standard technikával kapottak. További előnye az eljárásnak, hogy, a bottleneck réteg (relatíve) alacsony neuronszáma miatt több szomszédos kép felhasználása a becslés során nem jár a paraméterszám lényeges növekedésével, miközben szignifikánsan javítja a paraméterbecslés pontosságát.
Kulcsszavak: némabeszéd-interfész, mély neuronháló, autoenkóder

1. Bevezetés

Az utóbbi évtizedben megnőtt az érdeklődés a beszédjel artikulációs jellemzőkből való helyreállítása iránt, ami az ún némabeszéd-interfészek (Silent Speech Interface, SSI) alapját képezi [1]. Ezen a területen a feladat a beszédjel rekonstruálása az artikulációs szervek (pl. nyelv vagy ajkak) mozgásából anélkül, hogy az

alany valóban beszédjelet produkálna. A némabeszéd-interfészeknek kézenfekvő alkalmazási területeik lehetnek a beszédképzésben sérültek (pl. gégeeltávolításon átesett betegek) életminőségének javításában, illetve a beszéd továbbításában extrémen zajos környezetben (pl. katonai alkalmazásokban). Az artikulációs adatok rögzítése történhet ultrahangos képalkotással (ultrasound tongue imaging, UTI) [2,3,4,5,6], elektromágneses artikulográffal (electromagnetic articulography, EMA) [7,8], állandó mágneses artikulográffal (permanent magnetic articulography, PMA) [9], elektromiográfiával (electromyography, EMG) [10], avagy a fentieket keverő multimodális megoldásokkal [11].

A jelenlegi legkorszerűbb SSI rendszerek a „közvetlen szintézis" alapelvét alkalmazzák, vagyis a beszédjelet közbeeső átalakítások (pl. beszédhangok felismerése) nélkül, közvetlenül az artikulációs szervek mozgásából kinyert jellemzőkből állítják elő, vokóder használatával [3,4,5,8,9]. Ebben a folyamatban egy hangsúlyos gépi tanulási lépés az artikulációs jellemzők (pl. ultrahangképből nyert vektorok) alapján a vokóder (spektrális) paramétereinek becslése, melyre általában mély neurális hálót (Deep Neural Network, DNN, pl. [6,8,9]) vagy Gauss keverékmodellt (Gaussian Mixture Model, GMM, pl. [12,13]) szokás használni.

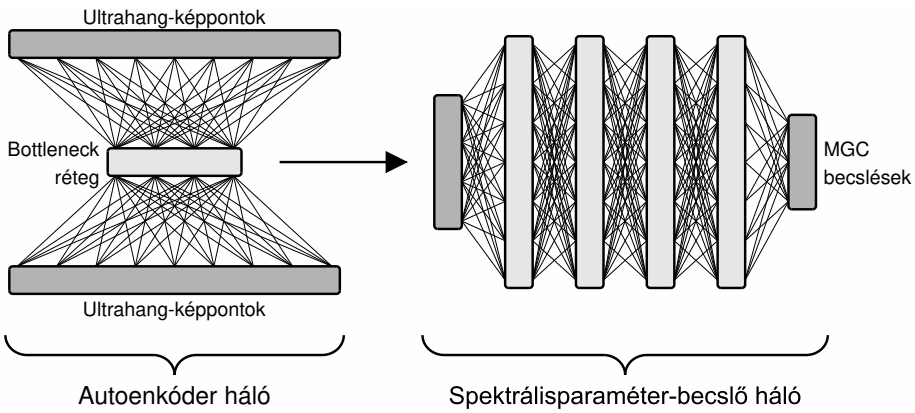
Az ultrahangkép-alapú SSI esetében a gépi tanuló eljárás bemenetét egy képkocka pixelei jelentik. Könnyen látható, hogy ez a megközelítés, bár kézenfekvő és korábbi tapasztalataink (ld. pl. [6,14,15,16]) alapján érthető beszéd szintetizálását teszi lehetővé, több tekintetben is szuboptimális. A bemenetként használt, képenként több ezer képpont (pl. a teljes nyers képkocka 64×842 méretű, azaz 53 888 képpontból áll) nagymértékben redundáns, valamint sok irreleváns jellemzőt is tartalmaz (bár ezen jellemzőkiválasztással lehet segíteni [14]). A túl sok jellemző az alkalmazott mély háló hatékonyságára (tanítási és kiértékelési idők, tárolt súlyok száma) egyértelműen negatív hatással van, és a spektrális paraméterek becslését is ronthatja. Egy hatékony tömörítési eljárással mindkét területen javíthatunk.

Jelen cikkünkben a bemenetként használt ultrahangképet egy autoenkóder hálózat segítségével tömörítjük, és a beszédszintézis spektrális paramétereit a bottleneck réteg aktivációit mint jellemzőket használva becsüljük egy második mély neurális hálóval. Kísérleti eredményeink alapján a javasolt megközelítés pontosabb paraméterbecslést tesz lehetővé, miközben a DNN mérete jelentősen csökken.

2. Némabeszéd-interfész spektrális paramétereinek becslése autoenkóder hálók használatával

2.1. Autoenkóder neurális hálók

Az autoenkóder neurális hálózat tanítására egy olyan felügyelet nélküli gépi tanulási eljárást alkalmazunk, melynek eredményeképpen a háló a rejtett rétegeiben az eredeti információ egy tömörebb változatát állítja elő, majd ezt a kimeneti rétegig visszafejti [17]. Célja, hogy bejövő paraméterekből egy identitásfüggvényhez hasonló leképezést tanuljon meg egy kompaktabb reprezentáción keresztül.



1. ábra: A javasolt kétlépéses DNN-alapú MGC-paraméterbecslő eljárás működési sémája.

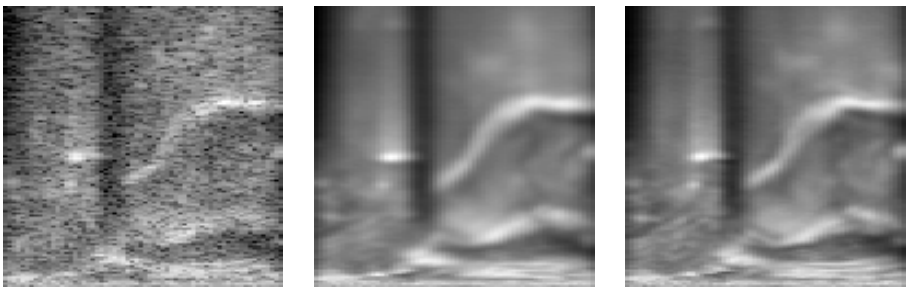
Technikailag általában egy olyan neurális hálóval valósítják meg, melynek a tanítás során elvárt kimenete megegyezik a bemenettel. Tömörítéskor az egyik rejtett rétegnek a bemenő jellemzők számánál lényegesen kevesebb neuronból kell állnia (*bottleneck* réteg). Korábbi kísérletek megmutatták, hogy ez a módszer alkalmas az egyes bemenetek közötti kapcsolatok feltárására [18], zajszűrésre [19], tömörítésre [20] vagy éppen új példák generálására a korábbi adatok alapján [21]. Az autoenkóder hálókat használják többek között képfeldolgozási [20,22], hangfeldolgozási [18] és természetes nyelvi feldolgozási [23] területeken.

Egy autoenkóder háló struktúráját tekintve két fő részből áll: az enkóder rész felelős a tömör reprezentáció előállításáért, a dekóder pedig a tömör információ alapján a bemenet visszaállításáért. A korábban említett bottleneck réteg a két rész metszetében található, ebben a rétegben számítódik/alakul ki a bemenet kódolt változata.

2.2. A spektrális paraméterek becslése autoenkóder hálók használatával

Jelen dolgozatunkban a beszédszintézis spektrális paramétereinek becslésére egy kétlépéses eljárást javaslunk, mindkét lépésben valamilyen mély neurális hálót alkalmazva. Az első lépésben egy autoenkóder hálót tanítunk egy-egy ultrahangkép pixeleinek rekonstruálására. A második lépésben egy újabb mély neurális hálót tanítunk, az autoenkóder háló bottleneck rétegében található neuronok aktivációit használva jellemzőként. Ennek a második hálónak a feladata már a beszédszintézis lépés paramétereinek predikciója (ld. 1. ábra).

Véleményünk szerint ennek a megközelítésnek több előnye is van. Az egyik pozitívum, hogy az autoenkóder háló észleli a szomszédos képpontok redundanciáját és képes az egymástól távolabb eső pixelek közti kapcsolatok felfedezésére is. Egy másik lehetséges előnye a javasolt megoldásnak azzal van kapcsolatban,



2. ábra: Egy szájüreg-ultrahangkép eredeti felvétele (balra), valamint az autoenkóder hálóval visszaállítva $N = 64$ (középen) és $N = 512$ (jobbra) neuront használva a bottleneck rétegben.

hogy az ultrahangkép természeténél fogva zajos. Reményeink szerint az autoenkóder háló azzal, hogy csak a tendenciaszerű változásokat kódolja a bottleneck rétegében, automatikusan elvégez egy zajszűrési lépést is. A harmadik előny, mellyel megközelítésünk rendelkezik, a tömörítéssel kapcsolatos. Egy általunk használt, standard felépítésű háló súlyainak számát nagymértékben határozza meg a bemeneti jellemzők száma; például a teljes, bár 64×128-ra átméretezett ultrahangkép pixeleinek megfelelő 8 192 bemeneti neuron és az első rejtett réteg 1 024 neuronja között kb. 8,4 millió kapcsolat van. Mivel a bottleneck réteg természetszerűleg (relatíve) kevés neuronból áll, ennek aktivációit használva bemenetként a végső hálónk jóval kevesebb kapcsolatból, így kevesebb súlyból állhat, amely mind tárolási szempontból, mind a predikció időigénye szempontjából előnyös. Amennyiben pedig, korábbi kísérleteinket követve (ld. pl. [6,14,15]), a szomszédos ultrahangképeket is felhasználjuk az aktuális keret MGC értékeinek megbecslésére, lehetőségünk nyílik lényegesen több szomszédos „kép” használatára úgy, hogy a háló súlyainak száma nem lesz nagyobb, mint az eredeti hálóé.

A 2. ábrán egy eredeti szájüreg-ultrahang kép látható (bal oldal), valamint ennek autoenkóder háló által visszaállított két változata; a középső kép esetén a bottleneck réteg 64 neuronból állt, míg a jobb oldali képnél 512 neuront tartalmazott. Látható, hogy az eredeti kép igen zajos, míg a visszaállított képek sokkal simábbak. A több rejtett neuront tartalmazó háló láthatólag több apró részletet őrzött meg az eredeti ultrahang-felvételből, mint a csupán 64 rejtett neuronnal rendelkező: utóbbi esetben a kép sokkal homályosabb, ugyanakkor a nyelv kontúrja itt is jól kivehető. Természetesen nem egyértelmű, hogy a konkrét feladat esetén legalább hány neuron szükséges optimális vagy közel optimális teljesítményhez.

3. Kísérletek

A következőkben bemutatjuk az elvégzett kísérletek technikai körülményeit: az alkalmazott adatbázist, a neurális hálók paramétereit és a kiértékeléskor használt metrikákat.

3.1. A felvételek rögzítése

A kísérletekhez használt felvételeket egy (42 éves) magyar anyanyelvű, beszédképzési problémával nem rendelkező nő segítségével rögzítettük, aki összesen 438 mondatot olvasott fel. Eközben a nyelv mozgását az Articulate Instruments Ltd. által gyártott „Micro" típusú ultrahang-berendezéssel rögzítettük 82 kép/másodperc sebességgel. Ezzel párhuzamosan a beszédjelet is felvettük egy Audio-Technica – ATR 3350 típusú kondenzátormikrofonnal (további részletekért lásd [6,14]). A továbbiakban ismertetett kísérletek inputját a nyers ultrahangfelvételek képezték. A 438 felvételt szétosztottuk tanító (310 felvétel), fejlesztési (41 felvétel) és teszthalmazra (87 felvétel).

3.2. Előfeldolgozás és szintetizálás

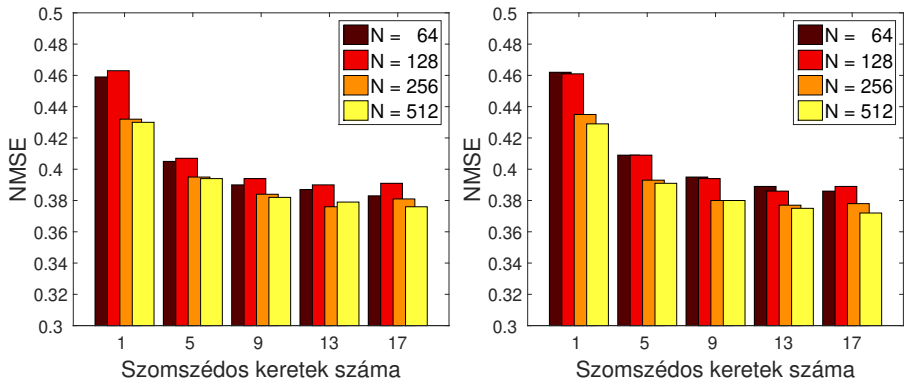
Az ultrahangképeket feldolgozás előtt az eredeti 64×946 felbontásról 64×128 pixelre méreteztük át. Az eredetileg $[0, 255]$ skálát használó pixelértékeket a képfeldolgozásban megszokott módon (ld. pl. [24]) elosztottuk 255-tel, így $[0, 1]$ skálára konvertálva azokat. A beszédjel elemzésére és szintetizálására a nyílt forrású SPTK eszköztár egyik vokóderét használtuk (http://sp-tk.sourceforge.net). A beszédjelet újramintavételeztük 22 050 Hz-en. A spektrális burkológörbét 24 MGC-LSP együtthatóval, valamint az energiaértékkel reprezentáltuk, ami összességében egy 25-dimenziós vektort eredményezett. A paramétereket az ultrahangképekkel szinkronban, 12 ms kereteltolással nyertük ki. A mély neuronhálók tanítása során az előbbi vektor standardizált változata képezte a megtanulandó célvektort.

3.3. A neurális háló paraméterei

A neurális hálók megvalósításához a Tensorflow [25] keretrendszert használtuk; a rejtett rétegekben minden esetben Swish aktivációs függvényt alkalmazó neuronokat alkalmaztunk [26], míg a beszédszintézis-paraméterek becslését szolgáltató 25 neuronnál lineáris aktivációt használtunk. A Swish neuronok α paraméterét 1.0 értéken rögzítettük.

A viszonyítási alapként szolgáló mély háló esetében a bemeneti réteg megfelelt az ultrahangkép képpontjainak, így 8 192 neuront tartalmazott, míg az öt rejtett réteg 1 024-1 024 neuronból állt. A súlyok kordában tartása érdekében L2 regularizációt alkalmaztunk. Korábbi kísérleteink (ld. pl. [6,14]) alapján tudtuk, hogy a szomszédos ultrahangképek használata segíthet az MGC paraméterek becslésében, így egy olyan hálót is tanítottunk, amely öt egymás utáni ultrahangképet kapott bemenetként (így ennek bemeneti rétege 40 960 neuronból állt). A tanítási célértékek a középső képkockához tartozó MGC paraméterek voltak. A két háló paramétereinek száma 12,6 millió (egy ultrahangkép esetén), illetve 46,2 millió (öt szomszédos ultrahangkép használata esetén) volt.

Az autoenkóder háló bottleneck rétegében $N = 64, 128, 256$ és 512 neuronnal kísérleteztünk, melyek közvetlenül (tehát további rejtett rétegek nélkül) voltak összekötve a bemeneti és kimeneti rétegekkel. (Ezek egy utrahangképnek voltak



3. ábra: A fejlesztési halmazon (balra) és a teszthalmazon (jobbra) mért átlagos normalizált hibaértékek az autoenkóder háló bottleneck rétegének neuronszáma (N) és a használt szomszédos keretek számának függvényében.

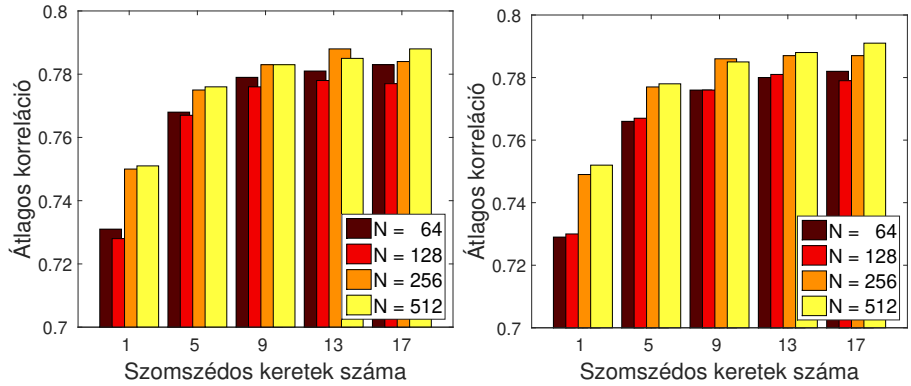
megfeleltetve, tehát 8 192 neuronból álltak.) A bottleneck réteg aktivációira tanított, MGC paraméterbecslő mély háló az előzőekhez hasonlóan egy-egy öt rejtett rétegből álló, mindegyikben 1 024 Swish neuront tartalmazó DNN volt. A teljes képhez viszonyítva lényegesen alacsonyabb jellemzőszám azt is lehetővé tette, hogy még több szomszédos „ultrahangképet” használjunk, így ebben az esetben kísérleteinket (összesen) $m = 1$, 5, 9, 13 és 17 szomszédos keret felhasználásával végeztük.

3.4. Kiértékelés

Mivel az MGC spektrális paraméterek becslése egy regressziós probléma, az egyes modellek kiértékelésére standard regressziós metrikákat alkalmaztunk. Az egyik lehetőség a négyzetes hiba használata; mivel 25 paramétert becsültünk, így kézenfekvő megközelítés az egyes spektrális paraméterekre kapott négyzetes hiba kiátlagolása. Ugyanakkor azt is érdemes figyelembe vennünk, hogy az egyes kimeneti értékek eltérő skálán mozoghatnak; ennek orvoslására inkább a normalizált négyzetes hibát használtuk. Egy másik lehetséges metrika az eredeti és a becsült értékek korrelációjának kiszámítása; a 25 korreláció-értéket egyszerű átlagszámítással összegeztük.

4. Eredmények

A 3. ábra bal oldala mutatja a mért átlagos normalizált négyzetes hibaértékeket a fejlesztési halmazon a különböző, autoenkóder-alapú konfigurációk esetén. Látható, hogy $m = 1$, illetve $m = 5$ (2-2) szomszédos keretet használva a becslések még lényegesen pontatlanabbak, mint akár $m = 9$ keret esetében; efölött viszont a javulás csak minimális, vagy egyenesen nincs is. A bottleneck réteg neuronszámát vizsgálva azt találtuk, hogy az $N = 64$ és $N = 128$ méretű hálók



4. ábra: A fejlesztési halmazon (balra) és a teszthalmazon (jobbra) mért átlagos korrelációértékek az autoenkóder háló bottleneck rétegének neuronszáma (N) és a használt szomszédos keretek számának függvényében.

valamivel pontatlanabb paraméterbecslést adtak, mint az $N = 256$ és $N = 512$ variációk, ugyanakkor a különbség csak akkor volt számottevő, mikor egyáltalán nem használtunk szomszédos kereteket ($m = 1$ eset). A teszthalmazon mért átlagos normalizált négyzetes hibaértékek (ld. 3. ábra jobb oldala) tendenciái szinte tökéletesen megegyeznek a fejlesztési halmazon tapasztaltakkal.

Az átlagos korrelációértékek a fejlesztési és a teszthalmazon (ld. 4. ábra) is nagyon hasonlóan alakultak: $m = 9$ szomszédos jellemzővektort használva optimális vagy aközeli értékeket kaptunk. Az autoenkóder háló bottleneck rétegében, tapasztalataink szerint, érdemes volt legalább 256 neuront használni, habár a különbség általában nem volt jelentős az egyes modellek teljesítménye között (legalább 9 szomszédos képet használva).

A konkrét értékeket (ld. 1. táblázat) megvizsgálva szembeszökő, hogy a teljes képet használva a szomszédos ultrahangképek használata, valamilyen oknál fogva, most nem javított a predikción. Az autoenkóder-alapú modellek esetén a legjobb teljesítményt az $N = 256$ eset hozta 13 (6-6) szomszédot használva mindkét metrika szerint és mindkét halmazon, de az is látható, hogy 9 szomszédot használva is csak kevéssel maradnak el az eredmények ettől a szinttől. A teszthalmazon mért 0,376-0,394 átlagos normalizált négyzetes hibaértékek 25-29%-os relatív hibacsökkenésnek felelnek meg, míg a 0,680-as átlagos korrelációértékekhez viszonyított 0,776-0,787-es értékek 30-33%-os hibacsökkentést jelentenek, melyeket bízvást nevezhetünk szignifikánsnak.

A táblázatban feltüntettük az egyes DNN-alapú modellek méretét (azaz a hálók összes súlyának számát) is. Mivel az autoenkóder-alapú konfigurációk esetében első lépésként az ultrahangkép kódolását kell elvégezni, ezekben az esetekben a feltüntetett értékek tartalmazzák az autoenkóder háló kódolásért felelős részének súlyszámait is. (Ezek 0,5 milliónak ($N = 64$), 1,0 milliónak ($N = 128$), 2,1 milliónak ($N = 256$) és 4,2 milliónak ($N = 512$) adódtak.) Látható, hogy az autoenkóder-alapú konfigurációk összesített súlyszáma csak néhány esetben

Megközelítés	Szomsz. száma	Param. száma	NMSE		Korreláció	
			Fejl.	Teszt	Fejl.	Teszt
Standard	1	12,6M	0,529	0,534	0,680	0,676
	5	46,2M	0,523	0,530	0,684	0,680
Autoenkóder, N = 64	1	4,8M	0,459	0,462	0,731	0,729
	9	5,3M	0,390	0,395	0,779	0,776
Autoenkóder, N = 256	1	6,6M	0,432	0,435	0,750	0,749
	9	8,7M	0,384	0,380	0,783	0,786
	13	9,7M	0,376	0,377	0,788	0,787
Autoenkóder, N = 512	1	8,9M	0,430	0,429	0,751	0,752
	5	11,0M	0,394	0,391	0,776	0,778
	9	13,1M	0,382	0,380	0,783	0,785

1. táblázat. A fejlesztési és a teszthalmazon mért átlagos normalizált négyzetes hibaértékek (NMSE) és átlagos korrelációértékek, valamint az egyes hálók súlyainak száma

haladta meg a viszonyítási alapként szolgáló, közvetlenül a teljes képet feldolgozó hálóét, azonban az öt egymást követő ultrahangképre tanított DNN méretétől jelentős mértékben elmaradtak. Ezen értékek alapján kijelenthetjük, hogy a javasolt, autoenkóder-alapú eljárás nemcsak pontosabb szintézisparaméter-becslésekhez vezet, hanem még számításilag is kedvezőbb.

5. Összegzés

Jelen cikkünkben az ultrahang-alapú némabeszéd-interfészek területén vizsgáltuk az autoenkóder neurális hálók alkalmazhatóságát. Megközelítésünkben a teljes szájüreg-ultrahangképre tanított autoenkóder háló bottleneck rétegének aktivációit mint jellemzőket használtuk, és a beszédszintézis spektrális paramétereit egy második mély hálóval becsültük. Kísérleti eredményeink alapján a javasolt eljárás a viszonyítási alapként szolgáló, pixelalapú megoldásnál hatékonyabbnak bizonyult: a becslések minden esetben pontosabbnak adódtak, és a háló súlyainak száma is csökkent. Véleményünk szerint ez több dolognak tudható be: az autoenkóder háló zajszűrési képességén kívül azt is ki tudtuk használni, hogy így az eredeti kép egy sokkal tömörebb reprezentációját állítottuk elő.

Az elvégzett kísérletek folytatására több kézenfekvő lehetőség is adódik. Az autoenkóder hálót kombinálhatjuk konvolúció alkalmazásával, mely remélhetőleg tovább növeli az eljárás hatékonyságát. Az autoenkóder-alapú reprezentációnak várhatóan nagyobb a robosztussága az ultrahang-készülék esetleges elmozdulásával szemben is, mint annak, amelyben minden képpontot a többi pixeltől független jellemzőként kezelünk. Emiatt megközelítésünk akár még a némabeszéd-interfészek beszélőfüggetlen működésének elérésében is segíthet. A közeljövőben tervezzük ilyen kísérletek elvégzését is.

Köszönetnyilvánítás

A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta (FK 124584). Tóth László munkáját az MTA Bolyai János Kutatási Ösztöndíja, valamint az Emberi Erőforrások Minisztériuma ÚNKP-18-4 kódszámú Új Nemzeti Kiválóság Programja támogatta. Grósz Tamás munkáját a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatta a 2018-1.2.1-NKP-2018-00008 azonosítójú projekt keretében. A cikk elkészítéséhez használt Titan-X grafikus kártyát az NVIDIA Corporation adományozta.

Hivatkozások

1. Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Communication **52**(4) (2010) 270–287
2. Denby, B., Stone, M.: Speech synthesis from real time ultrasound images of the tongue. In: ICASSP, Montreal, Kanada (2004) 685–688
3. Hueber, T., Benaroya, E.l., Denby, B., Chollet, G.: Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface. In: Interspeech, Florence, Olaszország (2011) 593–596
4. Hueber, T., Bailly, G., Denby, B.: Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface. In: Interspeech, Portland, USA (2012) 723–726
5. Jaumard-Hakoun, A., Xu, K., Leboullenger, C., Roussel-Ragot, P., Denby, B.: An articulatory-based singing voice synthesis using tongue and lips imaging. In: Interspeech, San Francisco, USA (2016) 1467–1471
6. Csapó, T.G., Grósz, T., Tóth, L., Markó, A.: Beszédszintézis ultrahangos artikulációs felvételekből mély neuronhálók segítségével. In: MSZNY 2017, Szeged (2017) 181–192
7. Wang, J., Samal, A., Green, J.: Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph. In: SPLAT, Baltimore, USA (2014) 38–45
8. Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., Yvert, B.: Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. PLOS Computational Biology **12**(11) (2016) e1005119
9. Gonzalez, J.A., Cheah, L.A., Green, P.D., Gilbert, J.M., Ell, S.R., Moore, R.K., Holdsworth, E.: Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary. In: Interspeech, Stockholm, Svédország (2017) 3986–3990
10. Nakamura, K., Janke, M., Wand, M., Schultz, T.: Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0. In: ICASSP, Prága, Csehország (2011) 573–576
11. Freitas, J., Ferreira, A.J., Figueiredo, M.A.T., Teixeira, A.J.S., Dias, M.S.: Enhancing multimodal silent speech interfaces with feature selection. In: Interspeech, Szingapúr (2014) 1169–1173
12. Janke, M., Wand, M., Nakamura, K., Schultz, T.: Further investigations on EMG-to-speech conversion. In: ICASSP, Kiotó, Japán (2012) 365–368

13. Gonzalez, J.A., Cheah, L.A., Gomez, A.M., Green, P.D., Gilbert, J.M., Ell, S.R., Moore, R.K., Holdsworth, E.: Direct speech reconstruction from articulatory sensor data by machine learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(12) (2017) 2362–2374
14. Csapó, T.G., Grósz, T., Gosztolya, G., Tóth, L., Markó, A.: DNN-based ultrasound-to-speech conversion for a silent speech interface. In: Interspeech, Stockholm, Svédország (2017) 3672–3676
15. Grósz, T., Tóth, L., Gosztolya, G., Csapó, T.G., Markó, A.: Kísérletek az alapfrekvencia becslésére mély neuronhálós, ultrahang-alapú némabeszéd-interfészekben (in Hungarian). In: MSZNY, Szeged (2018) 196–205
16. Tóth, L., Gosztolya, G., Grósz, T., Markó, A., Csapó, T.G.: Multi-task learning of speech recognition and speech synthesis parameters for ultrasound-based silent speech interfaces. In: Interspeech, Hyderabad, India (2018) 3172–3176
17. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations. MIT Press, Cambridge, MA (1986) 318–362
18. Lattner, S., Grachten, M., Widmer, G.: Learning transformations of musical material using Gated Autoencoders. In: CSMC, Milton Keynes, Nagy-Britannia (2017) 1–16
19. Geras, K.J., Sutton, C.: Scheduled denoising autoencoders. In: ICLR, San Diego, USA (2015) 365–368
20. Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Deep convolutional autoencoder-based lossy image compression. In: PCS, San Francisco, USA (2018) 253–257
21. Zhao, S., Song, J., Ermon, S.: Learning hierarchical features from generative models. In: ICML, Sydney, Ausztrália (2017) 4091–4099
22. Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: StyleBank: An explicit representation for neural image style transfer. In: CVPR, Honolulu, Hawaii (2017)
23. Andrews, M.: Compressing word embeddings. In: ICONIP, Kiotó, Japán (2016) 413–422
24. Varga, L.: Information Content of Projections and Reconstruction of Objects in Discrete Tomography. PhD thesis, Doctoral School of Computer Science, University of Szeged (2013)
25. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015) Software available from tensorflow.org.
26. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions (2018)