Kísérletek az alapfrekvencia becslésére mély neuronhálós, ultrahang-alapú némabeszéd-interfészekben

Grósz Tamás[1,2], Tóth László[1], Gosztolya Gábor[2],
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
{ tothl, groszt, ggabor } @ inf.u-szeged.hu
csapot @ tmit.bme.hu, marko.alexandra @ btk.elte.hu

Kivonat A legkorszerűbb némabeszéd-interfészek vokódert használva közvetlenül az artikulációs jellemzőkből generálnak beszédet. A legtöbb ilyen jellegű kutatásban csak a beszéd spektrális paramétereinek pontos becslésére koncentrálnak, a szintézishez szükséges alapfrekvencia-paraméterként konstans F0-értéket, az eredeti F0-görbét, vagy esetleg fehérzajos gerjesztést használnak. E megoldás mögött az a feltételezés húzódik meg, miszerint az F0-görbe nem becsülhető olyan artikulációs jellemzőkészletből, amely nem tartalmaz a hangszalagok rezgésére vonatkozó közvetlen mérési adatot. Jelen cikkünkben ilyen jellemzőkből, konkrétan ultrahang-felvételekből kíséreljük meg a beszédjel helyreállítását, kiemelt hangsúlyt helyezve a zöngésség, illetve az alapfrekvencia becslésére. Ellentétben a feltevéssel, miszerint az alapfrekvencia ilyen adatokból nem rekonstruálható, a becsült és az eredeti F0-görbe között 0,74-es korrelációs értéket mértünk. Ami még fontosabb, meghallgatásos kísérleteink alanyai nem tudtak különbséget tenni az eredeti és a becsült F0-görbével szintetizált mondatok között, azokat egyforma minőségűnek értékelték.
Kulcsszavak: némabeszéd-interfész, mély neuronháló, alapfrekvencia

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]. A probléma lényege, hogy valamilyen eszközzel rögzítjük az artikulációs szervek (pl. nyelv és/vagy ajkak) mozgását, majd ezekből az adatokból rekonstruáljuk a beszédjelet anélkül, hogy az alany valóban beszédjelet produkálna. A némabeszéd-interfészek hasznosak lehetnek a beszédképzésben sérültek (pl. gégeeltávolításon átesett betegek) számára, vagy olyan alkalmazásokban, ahol a beszédjel átvitele nem lehetséges, például extrém módon zajos

környezetben (lásd katonai alkalmazások). Az artikulációs adatok rögzítése történhet ultrahangos képalkotással (ultrasound tongue imaging, UTI) [2,3,4,5,6,7], elektromágneses artikulográffal (electromagnetic articulography, EMA) [8,9], állandó mágneses artikulográffal (permanent magnetic articulography, PMA) [10], elektromiográfiával (electromyography, EMG) [11] avagy a fentieket keverő multimodális megoldásokkal [12].

A jelenlegi legkorszerűbb SSI rendszerek a „közvetlen szintézis" alapelvét alkalmazzák, vagyis a beszédjelet közbeeső átalakítások nélkül (pl. beszédhangok felismerése) közvetlenül az artikulációs jellemzőkből állítják elő, vokóder használatával [3,4,5,9,10]. Az ilyen jellegű kísérleteket végző kutatók többsége a szintézishez szükséges spektrális jellemzők becslésére fókuszál (ilyen jellemzőkészlet pl. a Mel-Generalized Cepstrum, MGC). Ennek oka, hogy míg a spektrális burkológörbe nyilvánvaló módon a nyelv és az ajkak mozgásával korrelál, az alapfrekvencia (F0) paramétert a hangszalagok rezgése befolyásolja, ami viszont közvetlen módon nem függ a nyelv, az arc vagy az ajkak konfigurációjától [13]. Ennek ellenére vannak arra utaló kutatások, hogy a nyelv konfigurációja némiképp eltér zöngés és zöngétlen hangok esetén, például a hangszalag rezgése lelassul a mássalhangzók artikulációja során [14]. Egyéb tényezők mellett ezek a változások is korrelálnak az obstruens hangok artikulációs konfigurációjával, azaz a hangrés és a zár közötti térfogattal [15]. Mindezen kutatási eredmények ellenére az SSI rendszerek fejlesztői az F0-görbe becslését legtöbbször reménytelen feladatként kezelik, ezért az egyszerűség kedvéért konstans F0-értéket, az eredeti F0-görbét, vagy esetleg fehérzaj-gerjesztést használnak a szintézis során.

Van azonban néhány szerző, aki próbálkozott a zöngésség, illetve az F0-görbe helyreállításával. Nakamura és tsai EMG-felvételekkel dolgoztak, és a feladatot két lépésre bontva igyekeztek megoldani. Egyrészt egy SVM-et használtak a zöngés/zöngétlen (voiced/unvoiced, V/U) jelszakaszok elkülönítésére, majd második lépésben egy GMM modellel becsülték a zöngés szakaszok konkrét F0-értékét. Ezzel a módszerrel 0,5 körüli korrelációs értéket értek el az eredeti és a becsült F0-görbe között, maga a V/U döntés pontossága pedig 84% volt [11]. Hueber és tsai ultrahangos adatok és ajakvideók kombinálásával tettek kísérletet arra, hogy a spektrális paraméterek mellett a V/U döntést is megbecsüljék. A célra egy előrecsatolt (feed-forward) mély neuronhálót (deep neural network, DNN) használtak, és a V/U becslés pontosságára 82%-ot kaptak, ami nagyon hasonló Nakamura és tsai eredményéhez. Mivel a mért adatok nem tartalmaztak közvetlenül a hangszalagok működését reprezentáló jellemzőt, a relatíve magas pontosságot indirekt összefüggésekkel magyarázták, például azzal, hogy a stabil artikulációs konfigurációk általában magánhangzóknak, azaz zöngés jelszakaszoknak felelnek meg.

Két egészen friss publikáció EMA felvételekből próbálta az F0-görbét becsülni. Liu és tsai különféle típusú mély hálók (DNN, RNN és LSTM) teljesítményét hasonlították össze a zöngésség becslésére. Azt kapták, hogy a kétlépéses becslés, azaz először a spektrális jellemzők becslése, majd a becsléssel kapott értékek eredeti jellemzőkkel való összefűzése minden esetben javít az eredményeken [16]. Zhao és tsai pedig úgy találták, hogy az EMA adatok változási sebessége és

gyorsulása nagyon hasznosak az F0-görbe becslése során, és hogy az ún. LSTM neuronhálók jobban teljesítenek e téren, mint a standard DNN-ek. Objektív kiértékelési kritériumok szerint biztató F0-becslési eredményeket kaptak, meghallgatásos kiértékelést viszont nem végeztek [17].

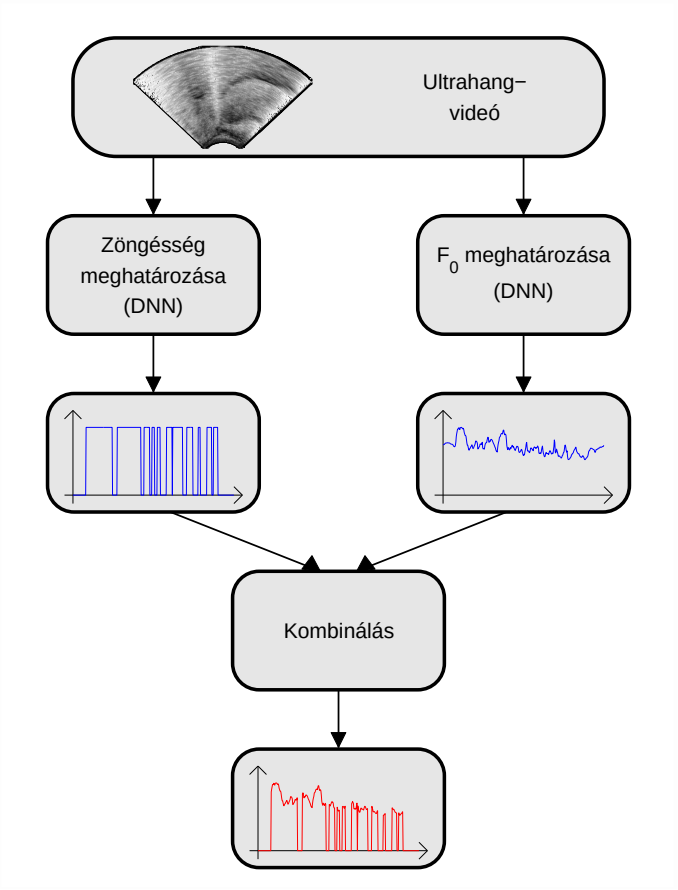
Habár a fenti rövid összegzés azt mutatja, hogy néhány kísérlet ugyan történt az F0-görbe becslésére, konkrétan olyan kutatást nem találtunk, amely a legmodernebb mély hálós technológiát használná ultrahangos kiindulási adatokon. Csapó és tsai végeztek ugyan ilyen kísérleteket, de az F0 becslésével nem próbálkoztak, másokhoz hasonlóan az eredeti F0-görbét használták a szintézis során [6,7]. Az azonban nyilvánvaló, hogy az SSI-rendszerek későbbi gyakorlati használhatóságához az F0 becslését is meg kell oldani. Jelen cikkben ezzel kísérletezünk, Nakamura és tsai megközelítéséhez hasonlóan egy kétlépcsős modellel, amelyben az egyik gépi tanulási komponens a zöngésségi jegyet, míg a másik a zöngés jelszakaszok konkrét F0-értékét igyekszik megbecsülni. Abban viszont eltérünk a Nakamura-féle tanulmánytól, hogy mindkét részfeladatra mély neuronhálókat alkalmazunk [11]. További eltérés, hogy míg ők EMG-felvételekből indultak ki, esetünkben ultrahang-felvétel az input, amely direkt módon nem rögzíti a hangszalagok rezgését. Módszereink kiértékelését egy női beszélőtől származó felvételeken fogjuk végezni.

2. Kísérleti beállítások

A kísérletekhez használt felvételeket egy (42 éves) magyar anyanyelvű, beszédképzési problémával nem rendelkező nő segítségével rögzítettük, aki összesen 473 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. A továbbiakban ismertetett kísérletek inputját a nyers ultrahang-felvételek képezték, a képek mérete 64×119 pixel volt (további részletekért lásd [6,7]).

2.1. Előfeldolgozás és szintetizálás

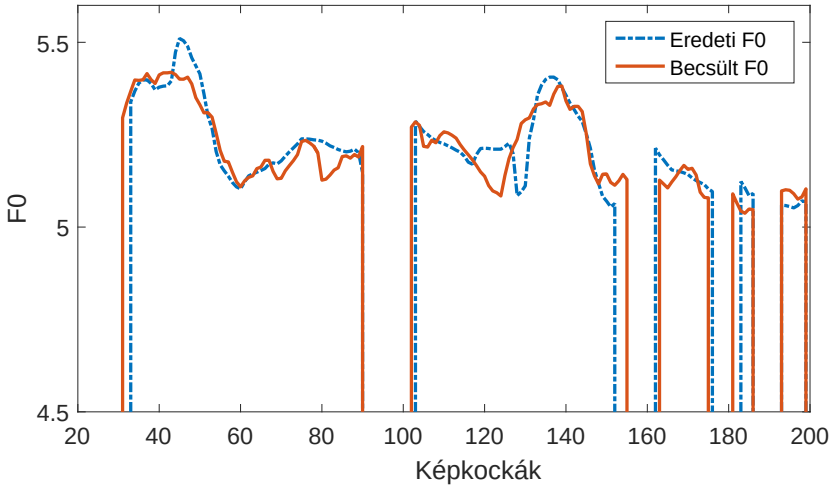
A beszédjel elemzésére és szintetizálására a nyílt forrású SPTK eszköztár egyik vokóderét használtuk (http://sp-tk.sourceforge.net). A beszédjelet újramintavételeztük 11 050 Hz-en. Az F0-görbét a SWIPE algoritmussal nyertük ki [18]. A spektrális burkológörbét 12 MGC-LSP együtthatóval, valamint az energiaértékkel reprezentáltuk, ami összességében egy 13-dimenziós jellemzővektort eredményezett. A paramétereket az ultrahang képekkel szinkronban, 12 ms kereteltolással nyertük ki. A mély neuronhálók tanítása során az előbbi a vektor képezte a megtanulandó célvektort. A szintézis során az eredeti paramétervektor helyett a DNN által az ultrahang-felvételekből becsült értékeket használtuk. Ehhez a vokóder az F0 paraméter segítségével impulzus-gerjesztőjelet produkált, majd ezt átengedte az MGC-LSP paraméterekből képzett ún. Mel-Generalized Log Spectral Approximation (MGLSA) szűrőn [19], így végezve el a beszédjel rekonstrukcióját.



1. ábra. Az F0 becslésére javasolt eljárás.

2.2. Az alapfrekvencia becslése mély neuronhálóval

Az alapfrekvencia becslése pusztán ultrahang-felvételek alapján nem egyszerű feladat. Nakamura és tsai eljáráshoz hasonlóan mi is kétlépcsős gépi tanulási megközelítést alkalmaztunk [11], ahol az egyik gépi tanuló a zöngésség jelenlétét, a másik pedig a zöngés keretek F0-értékét volt hivatva megbecsülni. Nakamura és társaival szemben mi mindkét feladatra mély neuronhálót alkalmaztunk. Mivel a V/U döntés kimenete bináris, ezt osztályozási feladatként kezeltük, a konkrét F0 érték becslését viszont regressziós feladatnak tekintettük, és ez utóbbi neuronhálót csak a zöngés időkeretek jellemzővektorain tanítottuk. A kiértékelési lépés, azaz a szintézis során a két háló kimenetét kombináltuk, azaz a zöngésnek ítélt adatkeretekhez a F0-becslő háló kimenetét rendeltük, míg a zöngétlen pillanatokban egy megfelelő (a vokóder által definiált) konstans értéket adtunk vissza. (A becslés menetéről lásd az 1. ábrát.)



2. ábra. Egy példa a rendszerünk alapfrekvencia-becslési kimenetére, az eredeti F0-görbéhez képest.

Megjegyezzük, hogy az utóbbi években a képi felismerési feladatokban a konvolúciós neuronhálók használata vált dominánssá. Mi azonban egyszerű teljesen kapcsolt hálót használtunk az alábbi megfontolások miatt. Egyrészt a konvolúciós hálók olyan feladatokban bizonyulnak sokkal jobbnak a standard hálóknál, amikor a kép sok apró részletből, hierarchikusan épül föl. Esetünkben azonban a kép alacsony felbontású és csupán pár építőelemből áll (a nyelv, az állkapocs valamint a nyelvcsont árnyéka). Másrészt a konvolúciós hálók a hierarchikus felbontás során elveszítik az egyes komponensek pontos pozícióját. Általános esetben ez még előnyös is, de úgy éreztük, hogy esetünkben lényeges információt veszítenénk. A fenti indokok miatt egy egyszerű, 5 rejtett réteget tartalmazó előrecsatolt hálóval dolgoztunk, rétegenként 1000 ReLU neuronnal. A háló feladata a 12 LSP paraméter, az energia és az F0 érték becslése volt; előzetes kísérleteinkben úgy találtuk, hogy hatékonyabb ezeket együtt tanulni, mint külön-külön. Ezért a paraméterek együttes tanulására (regressziójára) egy 14 lineáris kimeneti neuront tartalmazó DNN-t használtunk. A zöngésség eldöntését tanuló neuronhálónk ugyanilyen felépítésű volt, de egy kétosztályos softmax kimeneti réteggel.

2.3. Az input reprezentálása

Csapó és tsai korábbi cikkükben számos kísérletet végeztek az optimális input-reprezentáció megtalálására. Ezért mi itt csak azokat a megoldásokat vizsgáltuk, amelyek ott a legjobbnak bizonyultak [6,7]. Mindkét neuronháló ugyanazon az input adaton tanult. Összehasonlítási alapként (a továbbiakban *alaprendszer*), a legegyszerűbb megoldás szolgált, amikor a DNN inputja egyetlen ultrahang képkocka. Ezután az inputot kibővítettük, hogy 5 szomszédos képkockából álló adatvektort tartalmazzon (a továbbiakban *5-keret*). Ezzel az egyszerű megoldással

Input	Development			Teszt		
	Zönge. becslés	F_0		Zönge. becslés	F_0	
		NMSE	Korr.		NMSE	Korr.
Alaprendszer (1 keret)	86,1%	0,562	0,719	85,7%	0,577	0,711
5 keret	88,2%	0,476	0,760	87,2%	0,516	0,742
17 keret + jellemzőkiv.	87,4%	0,506	0,747	86,9%	0,526	0,736

1. táblázat. A DNN zöngésségi és F0-becslési pontossága különféle input-jellemzőkészletek esetén

azonban jelentősen megnő a jellemzővektor mérete, ami gépi tanulási szempontból nem előnyös. Emiatt egy jellemzőkiválsztási módszert alkalmaztunk az egyes képek méretének 20%-ra való redukálására [6,7]. Ily módon 8-8 bal- és jobboldali szomszédos képkockát is figyelembe tudtunk venni az inputvektor méretének jelentős növekedése nélkül. Ezt a rendszert *17-keret+jk* néven fogjuk hivatkozni. További részletek a jellemzőkinyerésről Csapó és tsai cikkében találhatók [7].

A legjobbnak bizonyuló F0-becslő modell kiértékelésére szubjektív meghallgatásos teszteket is végeztünk, amelyekben ugyanazt a mondatot különböző F0-görbékkel is szintetizáltuk. Összehasonlítási alapként konstans F0-t alkalmaztunk, ahol a DNN csupán a V/U döntést becsülte. Ez a modell *F0-konstans* címkével szerepel az ábrákon. Az összehasonlítás másik végpontjaként a mondatot az eredeti F0-görbével szintetizáltuk (*F0-eredeti* az ábrákon). Végezetül, a becsült F0-görbét használó modell *F0-DNN* néven szerepel a továbbiakban.

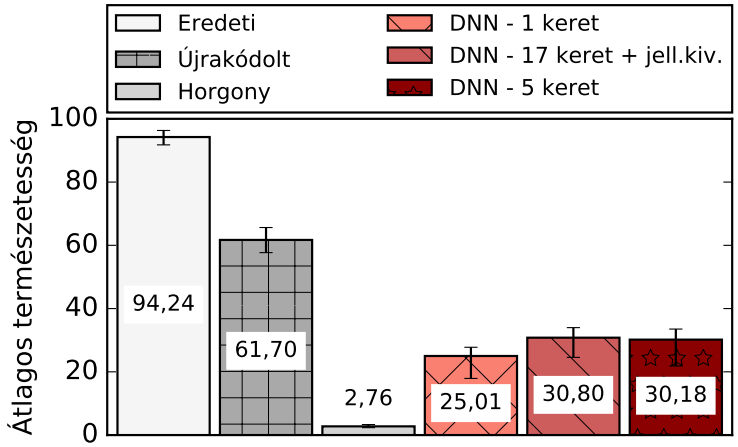
3. Eredmények és diszkusszió

3.1. Objektív kiértékelés

A DNN-alapú F0-becslés pontosságának számszerű kiértékelésére a normalizált átlagos négyzetes hibát (Normalized Mean Square Error, NMSE) és a Pearson korrelációs együtthatót használtuk [7]. Az 1. táblázat összegzi a különféle, az előző fejezetben bemutatott inputreprezentálási módszerek esetén kapott eredményeket. Mindegyik megoldással elég jó V/U döntési pontosságot értünk el, a legjobb F0-becslő modell pedig 0,74 korrelációs értéket produkált. A 2. ábra egy példán érzékelteti az eredeti és a becsült F0-görbe eltérését. Az eredmények alapján az alapfrekvencia ultrahang-felvételekből való becslése egyáltalán nem reménytelen.

3.2. Szubjektív meghallgatásos tesztek

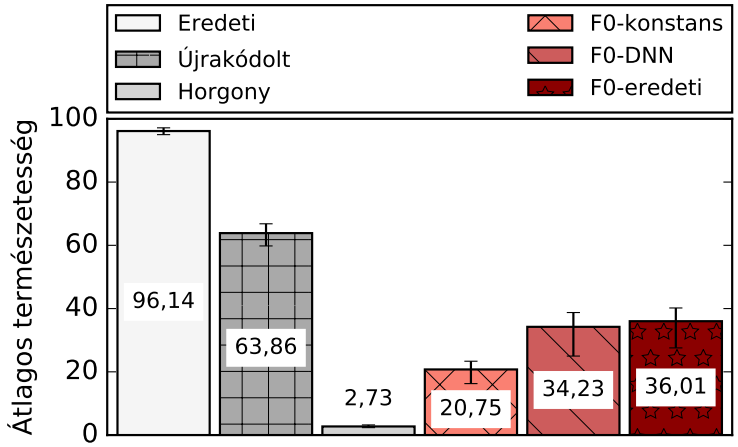
Az objektív hibamértékek sajnos nem sokat mondanak arról, hogy vajon a becslési hiba mennyire befolyásolja a szintetizált mondatok minőségét. Ezért a hangminőség szubjektív kiértékelésére online MUSHRA (MUlti-Stimulus test with



3. ábra. A különböző reprezentációs módszerek összehasonlítására végzett meghallgatásos tesztek eredménye.

Hidden Reference and Anchor) meghallgatásos teszteket végeztünk [20]. A tesztek célja az eredeti, a szintetizált, és a DNN-nel becsült F0-görbével szintetizált hangminták minőségének összevetése volt. Az újraszintetizált mondatokat a vokóder segítségével, a vokódert az eredeti mondatokból kinyert F0 és MGC paraméterekkel futtatva kaptuk. A minőségi skála alsó pontjának belövését segítendő a tesztek mindig tartalmaztak egy rossz minőségű ún. "horgony" mondatot, ezt konstans F0-érték és torzított MGC-paraméterek használatával állítottuk elő. A tesztekben 10 olyan mondat szerepelt, amelyeket a gépi tanuló algoritmus nem látott. Kétféle tesztet végeztünk. Az első célja a különféle input-reprezentálási módozatok összevetése volt, míg a második teszt az F0-becslés pontosságának a hangminőségre való hatását volt hivatva felmérni. Az 1. teszt az *Alaprendszert*, a *17-keret+jk*, és a *5-keret* elnevezésű reprezentálási stratégiákat hasonlította össze, míg a 2. teszt az *F0-konstans*, *F0-DNN*, és *F0-eredeti* nevű F0-becslési módozatokat vetette össze. Mind a 10 mondat, illetve az összes mondatvariáns (6 változat mondatonként) véletlenszerű (és minden tesztalany számára eltérő) sorrendben szerepelt a tesztekben. A tesztalanyok feladata az egyes mondatváltozatok minőségének pontozása volt egy 0-tól (nem természetes) 100-ig (természetes) terjedő skálán, ahol a 100 pontos minőség referenciájaként az eredeti mondatot adtuk meg.

A meghallgatásos tesztek eredménye Összesen 24 személy vett részt a tesztelésben (13 nő és 11 férfi). Mindannyian magyar anyanyelvűek voltak, négyük rendelkezett beszédtechnológiai ismeretekkel. Életkoruk 18–74 év közé esett, az átlag 28 év volt. A teszt elvégzése átlagosan 18 percet igényelt. A MUSHRA pontértékeket a 3. ábra mutatja az első, a 4. ábra a második teszt esetén (a 95%-os konfidenciaintervallumokat is feltüntetve). Mint látható, az eredeti mondatok közel 100%-ot értek el a természetességi skálán. Az újraszintetizált mondatok



4. ábra. A különböző F0-predikciós módszerek összehasonlítására végzett meghallgatásos tesztek eredménye.

ehhez képest csupán 60%-ot kaptak, míg a skála alsó pontját mutató „horgony" mondatok valóban messze a legalacsonyabb pontszámot kapták. A három DNN-alapú, eltérő jellemzőreprezentálást alkalmazó mondatokat a hallgatók 20–36% közé tették, ami azt mutatja, hogy a természetességük az újraszintetizált és a horgony mondatok között körülbelül félútra esik.

A meghallgatásos tesztben előállt rangsort Mann-Whitney-Wilcoxon teszttel is összevetettük, 95%-os konfidenciaszintet használva. Ennek alapján a DNN segítségével szintetizált mondatok szignifikánsan különböző minőségűnek bizonyultak az eredeti, a horgony és az újraszintetizált mondatok mindegyikéhez képest. Az első tesztben a *17-keret+jk* és az *5-keret* reprezentációs stratégiák szignifikánsan jobbak voltak, mint az *alaprendszer*, de kettőjük között szignifikáns különbség nem mutatkozott. A második tesztben a *F0-konstans* becslést használó modell szignifikánsan rosszabb eredményt ért el, mint az *F0-DNN* és az *F0-eredeti* rendszer, és ami a legfontosabb, az utóbbi kettő között az eltérés nem volt szignifikáns. Ez azt jelenti, hogy a tesztalanyok nem tudtak különbséget tenni a DNN-predikált és az eredeti F0-görbével szintezitált mondatok között, azokat egyforma minőségűnek ítélték.

4. Konklúzió

Cikkünkben kísérleteket végeztünk az F0-görbe becslésére ultrahang-felvételekből történő beszédjel-helyreállítás során. E célra két mély neuronhálót tanítottunk párhuzamosan, egyet a zöngésség, egyet pedig a konkrét F0-érték és a spektrális paraméterek becslésére. Objektív kiértékelési mértékeket használva azt találtuk, hogy a becsült és az eredeti alapfrekvencia-görbe korrelációs együtthatója 0,74. Szubjektív meghallgatási tesztjeink során a tesztalanyok az eredeti, illetve a predikált F0-görbével szintetizált mondatokat közel egyformán természetesnek

minősítették. Eredményeink alapján elmondhatjuk, hogy a közkeletű feltevéssel szemben az alapfrekvencia becslése nem reménytelen még akkor sem, ha a mérési adatok (esetünkben ultrahang-felvételek) a hangszalagok rezgését közvetlen módon nem reprezentálják. Mindamellett megjegyezzük, hogy kísérleteinkben egyetlen beszélő anyagával dolgoztunk, és ez a tény feltehetőleg nagy mértékben hozzájárult a jó eredményekhez. Azt gondoljuk, hogy a beszélőfüggőség nem jelent hátrányt olyan értelemben, hogy a jövőbeni némabeszéd-interfészek minden bizonnyal személyfüggőek lesznek. Ennek ellenére tervezzük kísérleteinket megismételni több beszélőtől (férfiak és nők vegyesen) származó felvételekkel. További megjegyzésünk, hogy éles alkalmazásokban a rendszernek majd a hangszalagok rezgése nélkül is működnie kell, mi viszont egy egészséges tesztalannyal dolgoztunk, aki normál beszédet produkált. Ebből következően szükséges lesz a kísérletek valódi néma beszéddel való megismétlése is.

Köszönetnyilvánítás

A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta (FK 124584). Tóth László munkáját az MTA Bolyai János Kutatási Ösztöndíja támogatta. Grósz Tamást az Emberi Erőforrások Minisztériuma ÚNKP-17-3 kódszámú Új Nemzeti Kiválóság Programja támogatta. 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. (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. 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
8. 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

9. 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
10. 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
11. 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
12. 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
13. Jiang, J., Alwan, A., Bernstein, L.E., Keating, P., Auer Jr., E.: On the correlation between facial movements, tongue movements, and speech acoustics. EURASIP Journal on Applied Signal Processing (11) (2002) 174–1188
14. Bickley, C.A., Stevens, K.N.: Effects of a vocal tract constriction on the glottal source: experimental and modeling studies. Journal of Phonetics **14** (1986) 373–382
15. Westbury, J.R., Keating, P.A.: On the naturalness of stop consonant voicing. Journal of Linguistics **22** (1986) 145–166
16. Liu, Z.C., Ling, Z.H., Dai, L.R.: Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks. In: Interspeech. (2016) 1502–1506
17. Zhao, C., Wang, L., Dang, J., Yu, R.: Prediction of F0 based on articulatory features using DNN. In: ISSP, Tienjin, Kína (2017)
18. Camacho, A., Harris, J.G.: A sawtooth waveform inspired pitch estimator for speech and music. The Journal of the Acoustical Society of America **124**(3) (2008) 1638–1652
19. Imai, S., Sumita, K., Furuichi, C.: Mel log spectrum approximation (MLSA) filter for speech synthesis. Electronics and Communications in Japan (Part I: Communications) **66**(2) (1983) 10–18
20. : ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality (2001)