Beszédszintézis ultrahangos artikulációs felvételekből mély neuronhálók segítségével

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

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék,
[2]MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoport,
[3]Szegedi Tudományegyetem, Informatikai Intézet,
[4]MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
[5]Eötvös Loránd Tudományegyetem, Fonetikai Tanszék,

e-mail: csapot@tmit.bme.hu, groszt@inf.u-szeged.hu,
tothl@inf.u-szeged.hu, marko.alexandra@btk.elte.hu

Kivonat A kutatás célja egy olyan rendszer létrehozása, amely a nyelv ultrahangos felvételeiből beszédet tud szintetizálni. A kutatás során egy női beszélőtől rögzítettünk közel 200 bemondáshoz tartozó szinkronizált akusztikai és artikulációs adatot, azaz nyelvultrahang-felvételt. A beszédből az alapfrekvenciát és spektrális paramétereket nyertük ki. Ezután mély neurális hálón alapuló gépi tanulást alkalmaztunk, melynek bemenete a nyers nyelvultrahang volt, kimenete pedig a beszéd spektrális paraméterei, ún. ,,mel-általánosított kepsztrum" reprezentációban. A tesztelés során egy impulzus-zaj gerjesztésű vokódert alkalmaztunk, mellyel az eredeti beszédből származó F0 paraméterrel és a gépi tanulás által becsült spektrális paraméterekkel mondatokat szintetizáltunk. Az így szintetizált beszédben sok esetben szavak, vagy akár teljes mondatok is érthetőek lettek, így a kezdeti eredményeket biztatónak tartjuk.

Kulcsszavak: gépi tanulás, artikuláció, beszédtechnológia, vokóder

1. Bevezetés

A beszédhangok az artikulációs szervek (hangszalagok, nyelv, ajkak stb.) koordinált mozgásának eredményéből állnak elő. Az artikuláció és a keletkező beszédjel kapcsolata régóta foglalkoztatja a beszédkutatókat. Beszéd közben a nyelv mozgását többféle technológia segítségével is lehet rögzíteni és vizsgálni, például röntgen [1,2,3], ultrahang [4,5], elektromágneses artikulográf (EMA) [6,7], mágnesesrezonancia-képalkotás (MRI) [8,9] és permanens mágneses artikulográf (PMA) [10]. Az ultrahangos technológia előnye, hogy egyszerűen használható, elérhető árú, valamint nagy felbontású (akár 800 x 600 pixel) és nagy sebességű (akár 100 képkocka/s) felvétel készíthető vele. A hátránya viszont az, hogy a hagyományos beszédkutatási kísérletekhez a rögzített képsorozatból ki kell nyerni a nyelv és a többi beszédszerv körvonalát ahhoz, hogy az adatokon további vizsgálatokat lehessen végezni. Ez elvégezhető manuálisan, ami

rendkívül időigényes, illetve automatikus módszerekkel, amelyek viszont ma még nem elég megbízhatóak [11]. Arra is lehetőség van, hogy az ultrahangképekből közvetlenül, a nyelvkontúr kinyerése nélkül állapítsunk meg az artikulációs szerv aktuális pozíciójára utaló információt [12].

Az artikuláció és az akusztikai kimenet kapcsolatát gépi tanulás alapú eszközökkel is vizsgálták már. Az artikuláció-akusztikum konverzió eredményei a szakirodalomban elsősorban az ún. 'Silent Speech Interface' (SSI, magyarul 'némabeszéd-interfész') rendszerek fejlesztéséhez járulnak hozzá [13]. Az SSI lényege, hogy az artikulációs szervek hangtalan mozgását felvéve a gépi rendszer ebből beszédet szintetizál, miközben az eszköz használója valójában nem ad ki hangot. Ez egyrészt a beszédsérült embereknek (pl. gégeeltávolítás után) lehet hasznos, másrészt potenciálisan alkalmazható zajos környezetben történő beszédhang kiadására, kiabálás nélkül. Mivel az SSI közvetlenül az artikulációt rögzíti, ezért a rendszer nem érzékeny a környezeti zajokra. A konverziós feladathoz többnyire EMA-t [14,15,16], ultrahangot [17,18,19,20,21,22] vagy PMA-t [23] használnak inputként, mi azonban csak az ultrahangra koncentrálunk a jelen áttekintésben.

Az egyik első hasonló kísérletben egy egyszerű neurális hálózattal próbálták a nyelvmozgás ultrahangos képének és a beszéd spektrális paramétereinek összefüggését megtalálni [17], de az eredmények ekkor még nem voltak meggyőzőek, mert az alkalmazott neurális hálózat nem volt alkalmas a komplex feladat megoldására. Később az SSI rendszereket „felismerés-majd-szintézis" alapon valósították meg, azaz a cél az volt, hogy az ultrahangalapú artikulációs adatokból először a beszédhangokat kinyerjék egy vizuális felismerő módszerrel, majd ezután egy beszédszintézis-rendszer felolvassa a beszédet [18]. Ezen megoldás hátránya, hogy a komponensek hibája összeadódik, azaz a beszédhang-felismerés esetleges tévesztése nagyon elrontja a beszédszintézis eredményét. A későbbi SSI rendszerekben ezért a „közvetlen szintézis" módszer terjedt el, azaz a köztes beszédhangfelismerés nélkül, az artikulációs adatok alapján próbálják megbecsülni a beszéd valamilyen reprezentációját (tipikusan a spektrális paramétereit) [19,20,21]. Az alkalmazott gépi tanulási módszer ezekben a kísérletekben Gauss-keverékmodell (gaussian mixture model, GMM) [19], illetve rejtett Markov-modell volt [20,21].

A legújabb eredmények szerint a mély neurális hálózatok (például a konvolúciós hálózatok) az emberi teljesítményt megközelítő vagy akár jobb pontosságot értek el olyan feladatokban, mint az objektumfelismerés [24], képek osztályozása [25], él/kontúr-detekció [26] stb. Az ultrahangalapú SSI témakörében eddig egyetlen kutatás alkalmazott mély neurális hálózatot [22]. A kutatásban ultrahang- és ajakvideó-alapú artikulációs adatok alapján alkalmaztak autoencoder neuronhálózatot, illetve előrecsatolt hálózatot (MLP) egy egyszerű vokóder spektrális (egész pontosan ún. LSF) paramétereinek becslésére, végül ez alapján éneket hoztak létre egy artikulációs szintetizátorral. Az eredmények és a hangminták szerint a becslési feladat megoldása előremutató, de még további kutatást igényel.

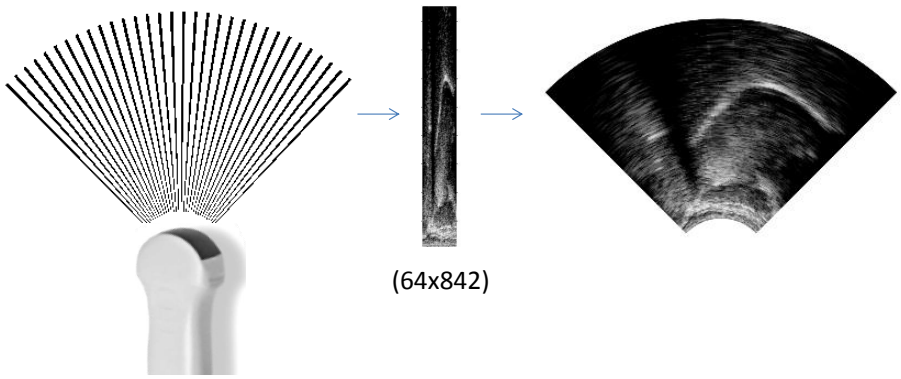
1.1. A jelen kutatás célja

A szakirodalmi áttekintés szerint az artikuláció-akusztikum konverzió még kezdeti stádiumban van, és a valós időben működő SSI rendszerek kifejlesztése a feladat minél pontosabb megoldását igényli. A jelen tanulmányban bemutatjuk az első erre irányuló kísérletünket, amelyben egy magyar beszélő ultrahangos felvételei alapján beszédet szintetizálunk.

2. Módszerek

2.1. Felvételek és adatok

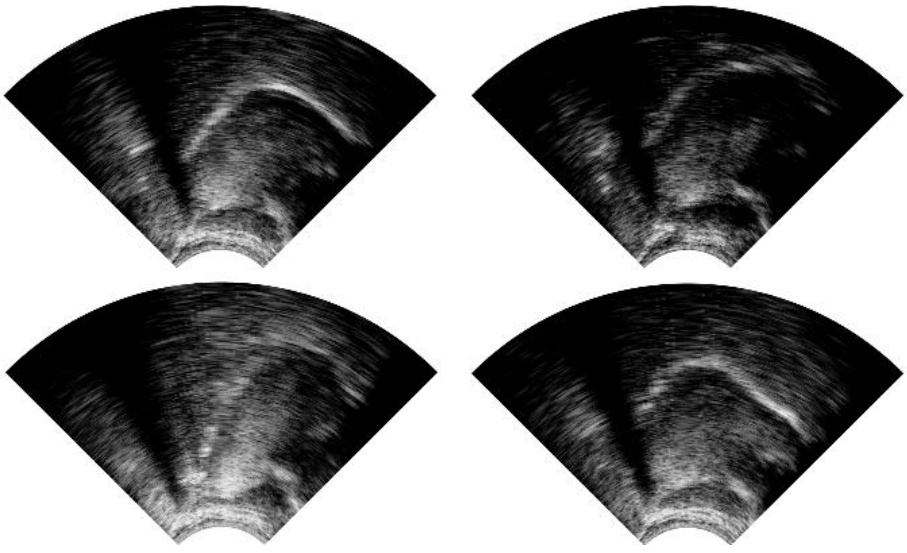
A kutatáshoz egy női beszélőtől (MA) rögzítettünk párhuzamos ultrahang- és beszédfelvételeket. A felvételek az ELTE Fonetikai Tanszék egyik csendes szobájában készültek, a szakirodalomban javasolt helyzetben és beállításokkal [5]. A beszélő a PPBA adatbázis [27] első 176 mondatát olvasta fel. A nyelv középvonalának (szagittális) mozgását a SonoSpeech rendszerrel rögzítettük (Articulate Instruments Ltd.) egy 2–4 MHz frekvenciájú, 64 elemű, 20 mm sugarú konvex ultrahang-vizsgálófejjel, 82 fps sebességgel. A felvételek során ultrahang-rögzítő sisakot is használtunk (Articulate Instruments Ltd., fénykép: [28]). A beszédet egy Audio-Technica – ATR 3350 omnidirekcionális kondenzátor-mikrofonnal rögzítettük, amely a sisakra volt csípteve, a szájtól kb. 20 cm-re. A hangot 22050 Hz mintavételi frekvenciával digitalizáltuk egy M-Audio – MT-RACK PLUS hangkártyával. Az ultrahang és a beszéd szinkronizációja a SonoSpech rendszer 'Frame sync' kimenetét használva történt: minden elkészült ultrahangkép után ezen a kimeneten megjelenik egy néhány ns nagyságrendű impulzus, amelyet egy 'Pulse stretch' egység szélesebb négyszög ugrássá alakít, hogy digitalizálható legyen [28]. Ez utóbbi jelet szintén a hangkártya rögzítette. A felolvasandó mondatok képernyőn megjelenítését és az adatok felvételét a kísérlet vezetője végezte az Articulate Assistant Advanced (Articulate Instruments Ltd.)



1. ábra. Nyers adatokból ultrahangkép előállítása.

szoftver használatával. A ultrahangból származó nyers adatokat ezután közvetlenül bináris formátumba mentettük (így nem veszett el adat a képpé konvertálás során). Az 1. ábra mutatja, hogy a letapogatás hogyan történik a SonoSpeech rendszerrel: az ultrahangfej 64 radiális vonalon (bal oldalon), minden vonalon 842 helyen méri az intenzitást, és a nyers adatban minden intenzitásértéket 8 biten tárol (ennek eredménye látható középen). Ha ezt a szokásos ultrahangképpé akarjuk alakítani, akkor az adatokat poláris koordinátarendszerben lehet ábrázolni szürkeárnyalatos képként, mely a jobb oldalon látható.

A 2. ábra néhány példát mutat a nyelvről készített ultrahangfelvételre a fenti női beszélőtől. A felvételeken bal oldalon látható a nyelvgyök, jobb oldalon a nyelvhegy; a kettő között a nyelv felső felülete. A bal oldali sötétebb rész a nyelvcsont helyére, míg a jobb oldali sötétebb rész az állkapocscsont helyére utal (mivel az ultrahang-hullám a csontokon nem tud áthatolni). A felvételek során az ultrahang-vizsgálófejet az áll alá helyeztük; így az ultrahangjelben a legnagyobb változást a nyelv izomzatának felső határa okozza, ami az ultrahangos képeken ideális esetben jól kivehető fehér sávot eredményez. Mivel a hullámok nagy része nem jut tovább a nyelv felső határán, így a távolabbi szövetpontokról, a szájpadlásról kevesebb az információnk. A 2. ábrán az is látható, hogy a képek minősége széles skálán mozog, mivel az ultrahangos technológia nem mindig nyújt teljesen tökéletes nyelvkontúrt. A bal felső és jobb alsó képen jól kivehető a nyelv kontúrja; ezzel szemben a bal alsó képen a kontúr nem folytonos, hanem szakadás vagy ugrás látható. A jobb felső képen a nyelvkontúr kevésbé erőteljesen látszik.



2. ábra. Különböző minőségű ultrahangképek ugyanazon beszélőtől.

2.2. A beszédjel előfeldolgozása

A beszédfelvételek és szöveges átiratuk alapján egy magyar nyelvű kényszerített felismerővel [29] meghatároztuk a hanghatárokat, majd a hanghatárok alapján a felvételek elején és végén található csendet nem vettük figyelembe a gépi tanulási adatok generálása során.

A beszédjel paraméterekre bontására és a későbbi visszaállításra egy egyszerű impulzus-zaj gerjesztésű vokódert választottunk (PySPTK implementáció: https://github.com/r9y9/pysptk). Az alapfrekvenciát (F0) a SWIPE algoritmussal mértük. A következő lépésben spektrális elemzést végeztünk mel-általánosított kepsztrum (Mel-Generalized Cepstrum, MGC, [30]) módszerrel, melyet statisztikai parametrikus beszédszintézisben széles körben használnak. Az elemzéshez 25-öd rendű MGC-t számítottunk $\alpha = 0{,}42$ és $\gamma = -1/3$ értékekkel. Ahhoz, hogy a beszédjel analízise során kapott paraméterek szinkronban legyenek az ultrahangképekkel, a kereteltolást 1 / FPS értékre választottuk (ahol FPS az adott ultrahangfelvétel képkocka/másodperc sebessége).

A beszéd visszaállításához az F0 paraméterből először impulzus-zaj gerjesztést generáltunk, majd a gerjesztést és az MGC paramétereket felhasználva MGLSADF szűrővel [31] visszaállítottuk a szintetizált beszédet. A fenti vokóder az SSI témakörében tehát úgy használható, hogy a beszéd visszaállításához az eredeti F0 paraméterek mellett nem az eredeti spektrális paramétereket használjuk fel, hanem az ultrahangképek alapján gépi tanulással becsülteket.

2.3. Az ultrahangadatok előfeldolgozása

Az ultrahangadatokon a csendes szakaszok kivágásán kívül egyéb előfeldolgozást nem végeztünk, azaz közvetlenül az ultrahangos rögzítés során előálló nyers adatok (az 1. ábra középső része) képezték a gépi tanulás inputját, ami gyakorlatilag megfelel annak, mint ha magukon az ultrahangképeken tanítanánk. Így 64 x 842 méretű jellemzővektorokkal kellett dolgoznunk, ami meglehetősen magas jellemzőszámot jelent. A 2.4. fejezetben bemutatunk egy nagyon egyszerű jellemzőkiválasztási módszert, amellyel megpróbáltuk kiszűrni az ultrahangképek azon régióit, ahol nem történik olyan változás, amely a tanulás során fontos lenne a modell számára, így az ide tartozó pixelértékek eldobhatók.

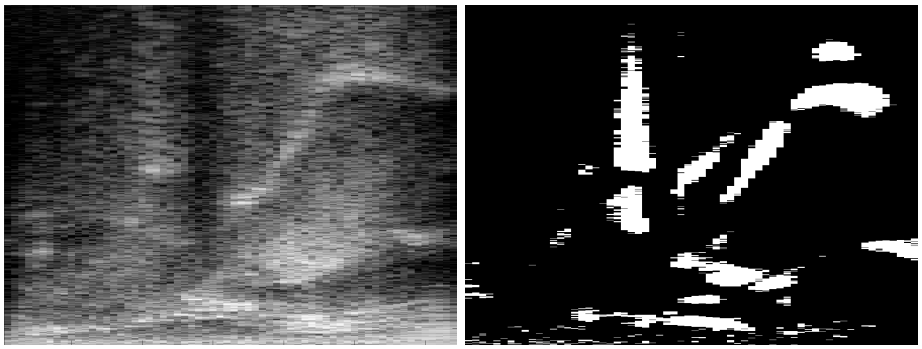
2.4. Gépi tanulás

Az ultrahangfelvételeken teljesen kapcsolt (fully connected) mély ,,egyenirányított" (rectifier) neurális hálókat [32] tanítottunk. A rectifier hálók esetén a rejtett neuronok a rectifier aktivációs függvényt (max(0,x)) alkalmazzák, ennek köszönhetően körülményes előtanítási módszerek nélkül, hagyományos backpropagation algoritmussal is hatékonyan taníthatóak [32]. A megtanulandó célértékeket a vokóder MGC paraméterei képezték. Mivel feltevéseink szerint az utrahangadatokból a hangmagasság értéke (F0) egyáltalán nem, a hangosság értéke (az MGC első dimenziója) pedig csak kis eséllyel állítható vissza, ezért ezt a két paramétert kihagytuk a gépi tanulásból, és a szintézis során

az eredeti értékeket használtuk. A fennmaradó 25 MGC-paraméter a beszéd spektrális burkolóját írja le, a neuronháló feladata ezeknek a paramétereknek a minél pontosabb becslése volt az ultrahang alapján. Mivel ezek a paraméterek folytonos értékűek, ezért osztályozás helyett regressziós módban használtuk a mély hálót. Egyelőre – jobb híján – az átlagos négyzetes hibafüggvény (MSE) segítségével tanítottunk. A későbbiekben érdemes lehet majd ezt leváltani egy olyan mértékre, amely figyelembe veszi az emberi percepciót is. Jaumard-Hakoun és munkatársai például a kiértékelésnél a spektrális torzítást mérték (bár a tanulás során feltehetően ők is az MSE-hibát használták, ez nem derül ki egyértelműen a tanulmányukból) [22]. A multidimenziós regressziós tanítást ők úgy oldották meg, hogy minden regressziós jellemzőre külön neuronhálót tanítottak. Munkánkban mi kipróbáltuk, hogy minden MGC jellemzőre külön hálót tanítva jobb eredményt kapunk-e, mint egy hálót tanítva egyszerre a teljes MGC vektorra.

Kísérleteink során egy 5 rejtett réteges, rétegenként 1000 neuront tartalmazó neuronháló struktúrát használtunk lineáris kimeneti réteggel. Tekintve, hogy az MGC paraméterek különböző skálán mozogtak, tanítás előtt standardizáltuk őket, hogy várható értékük 0, szórásuk pedig 1 legyen. A standardizálás egy fontos lépés, hiszen amennyiben ezt nem tesszük meg, úgy a regressziós tanulás során a nagyobb értékekkel rendelkező MGC jellemzőt tanulja meg a háló nagy pontossággal, míg a kisebb értéktartományon mozgót kevésbé az MSE hibafüggvény miatt.

A neuronhálók bemeneteként kezdetben az egész ultrahangképet használtuk, ami rendkívül zajos, és sok felesleges részt is tartalmaz (lásd 2. ábra), ezért egy egyszerű jellemzőkiválasztási eljárást is kipróbáltunk. A módszer lényege, hogy minden pixelre kiszámítottuk annak korrelációját a 25 MGC jellemzővel, majd vettük ezen korrelációk maximumát, és küszöböltünk, azaz csak azokat a pixeleket tartottuk meg, ahol a korreláció egy küszöbérték fölé esett. A 3. ábra egy példát mutat az eredeti felvételre, illetve a kapott szűrési maszkra (a fehér pontok jelentik a megtartott pixeleket). Az így kapott maszk alapján tudtuk szűrni, hogy a kép mely részeit érdemes figyelni. A bemeneti jellemzőkészlet redukálása



3. ábra. Ultrahangkép és a jellemzőkészlet szűréséhez használt maszk.

révén jelentősen, körülbelül a tized részére – 53 888-ról 5 572-re – redukáltuk a jellemzők számát. Ez a lépés lehetővé tette, hogy ne csak az aktuális ultrahangképet, hanem annak időbeli szomszédait is felhasználjuk a tanítás során. A beszédfelismerésben teljesen szokványos lépés az aktuális adatvektor mellett az időben szomszédos vektorokat is bemenetként megadni a hálónak, innen jött az ötlet erre a megoldásra. A kísérletekben az aktuális képen kívül 4-4 szomszédot használtunk fel inputként, ami összesen 9 szomszédos jellemzővektort jelent; így végső soron a szomszédokat is figyelembe vevő háló nagyságrendileg ugyanakkora inputvektoron dolgozott, mind amekkora az eredeti, redukálatlan inputvektor volt.

3. Kísérleti eredmények

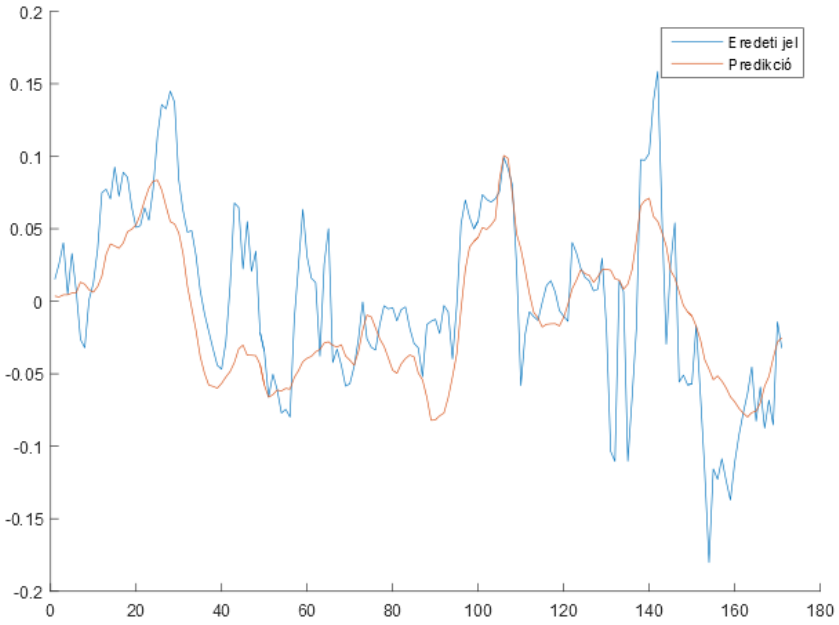
A 176 rendelkezésre álló felvételből 158-at használtunk a neuronhálók tanítására, a maradék 28-at pedig tesztelésre. A neuronháló különböző változataival a teszthalmazon elért átlagos négyzetes hiba (MSE) értékeit az 1. táblázat foglalja össze. A bemeneti jellemzők esetén két variációt próbáltunk meg. „Teljes" jellemzőkészletnek fogjuk hívni azt az esetet, amikor a teljes képet, azaz az összes, 53 888 rögzített adatot használtuk inputként. A korábban ismertetett jellemzőkiválasztási módszerrel előállított 5 572 elemű jellemzőkészletre „redukált" készletként hivatkozunk. A bemeneti képek száma 1 vagy 9 lehet, a 9 jelenti azt, hogy 9 egymást követő kép alkotta az inputot, ami természetesen csakis a redukált jellemzőkészlet esetén jön szóba. A betanított hálók oszlopában az 1-es értékek azt jelentik, hogy egyetlen hálót tanítottunk 25 kimenettel, míg a másik esetben 25 hálót tanítottunk külön-külön a 25 MGC-paraméter becslésére.

A táblázat első és harmadik sorát összevetve láthatjuk, hogy a jellemzők számának radikális csökkentése csak minimális mértékben növelte a hibát, azaz a jellemzőkiválasztási módszerünk jól teljesített. A harmadik és a negyedik sor összevetéséből pedig az olvasható ki, hogy a szomszédos 4-4 kép felhasználása körülbelül 10%-kal csökkentette a hibát. Végezetül, a többi sort is vizsgálva azt látjuk, hogy az egyes paraméterek közelítésére külön-külön tanított hálók nem javítottak számottevően, viszont betanításuk lényegesen több időt vett igénybe.

1. táblázat. A különböző módon tanított neuronhálókkal elért átlagos négyzetes hibák.

Bemeneti jellemzőkészlet	Bemeneti képek száma	Betanított hálók száma	MSE
Teljes	1	1	0,00194
	1	25	0,00190
Redukált	1	1	0,00203
	9	1	**0,00180**
	1	25	0,00199
	9	25	0,00184

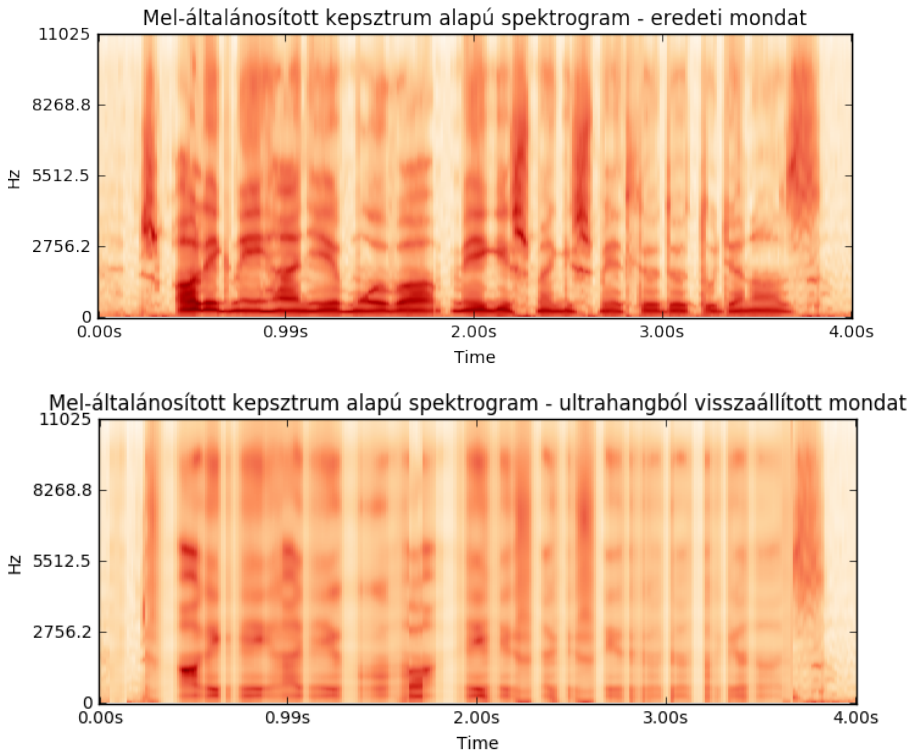
Az MSE hiba értéke sajnos nem túl informatív arra nézve, hogy milyen minőségű lett a visszaállított beszéd. A hiba érzékeltetésére a 4. ábrán kirajzoltuk egy konkrét MGC-paraméter időbeli görbéjét, valamint annak neuronhálóval kapott közelítését. Megfigyelhetjük, hogy a neuronháló alapvetően követi ugyan a görbe trendjét, de a finom részleteket sok esetben képtelen visszaadni. Az ebből eredő hiba csökkentésére tervezzük megvizsgálni, hogy az MGC-paraméterek mekkora időbeli simítást bírnak el minőségromlás nélkül, majd ezekkel a simított paraméterekkel fogjuk tanítani a hálót.



4. ábra. Egy MGC-paraméter időbeli görbéje és annak becslése a legjobb eredményt elérő neuronhálóval.

A hiba további érzékeltetésére az 5. ábrán példát mutatunk egy mondat eredeti, illetve a rekonstrukció után kapott spektrogramjára. Ugyan a neuronháló nem tudta pontosan megtanulni az eredeti beszédre jellemző összes spektrális komponenst (pl. formánsok), de a tendenciák alapján látható, hogy a gépi tanulás eredményeként kapott spektrogram is emlékeztet beszédre (pl. 0,5 s körül a formánsok egészen jól kivehetőek).

Az ultrahangból visszaállított felvételeken precíz, többalanyos lehallgatásos kiértékelést nem végeztünk, de a szubjektív benyomásunk az volt, hogy bár a felvételek nagyon torzak, sok esetben szavak, sőt némely esetben teljes mondatok is érthetőek. Ezt biztató kezdeti eredménynek tartjuk, tekintve, hogy a feldolgozás összes lépésében a lehető legegyszerűbb megoldást alkalmaztuk.



5. ábra. Felül: eredeti MGC-alapú spektrogram. Alul: gépi tanulással artikulációs adatokból becsült MGC-alapú spektrogram.

4. Összefoglalás, következtetések

A tanulmányban bemutattunk egy kísérletet, amelynek a célja az volt, hogy nyelvultrahang-képekből kiindulva beszédet szintetizáljunk. A kutatás során egy női beszélőtől rögzítettünk közel 200 bemondáshoz tartozó szinkronizált beszéd- és nyelvultrahang-felvételt. A beszédből az alapfrekvencia- és a spektrális paramétereket nyertük ki. Ezután mély neurális háló alapú gépi tanulást alkalmaztunk, melynek bemenete a nyelvultrahang volt, kimenete pedig a beszéd spektrális paraméterei. A tesztelés során egy impulzus-zaj gerjesztésű vokódert alkalmaztunk. Az eredeti beszédből származó F0 paraméterrel és a gépi tanulás által becsült spektrális paraméterekkel mondatokat szintetizáltunk. Az így szintetizált beszédben sok esetben szavak, vagy akár teljes mondatok is érthetőek lettek.

A jelen cikkben elért kezdeti eredményeket biztatónak tartjuk. A továbbiakban a rendszernek gyakorlatilag minden pontján finomításokat tervezünk. Meg fogjuk vizsgálni, hogy a szintézis mely paramétereinek becslése a legmegfelelőbb, tervezzük variálni az optimalizálandó célfüggvényt, a ne-

uronháló struktúráját (pl. teljesen kapcsolt helyett konvolúciós), és a jellemzőkinyerési-jellemzőredukciós lépés is rengeteg kísérleti lehetőséget kínál. Emellett a szájpadlás helyzetéről kinyert információ [33] hozzáadása is segítheti a feladat megoldását.

A mai 'Silent Speech Interface' rendszerek ugyan még kísérleti fázisban vannak, de a jövőben várhatóan valós időben is megvalósítható lesz az artikuláció-akusztikum becslés problémája. Az SSI rendszerek hasznosak lehetnek a beszédsérültek kommunikációjában, illetve zajos környezetben történő beszéd során [13]. A beszélőfüggetlen SSI rendszerek elkészítése egyelőre kihívást jelent, de a legújabb kutatások szerint konvolúciós hálózatokkal ebben a témakörben is nagy előrelépést lehet elérni [34].

Az artikuláció és az akusztikum (elsősorban beszéd) kapcsolatának vizsgálata a beszédkutatás alapkérdéseinek megválaszolása mellett hasznos lehet nyelvoktatásban, beszédrehabilitációban, illetve beszédtechnológiában, audiovizuális beszédszintézisben is.

Köszönetnyilvánítás

A kutatás során Csapó Tamás Gábort és Markó Alexandrát az MTA „Lendület" programja; Grósz Tamást az Emberi Erőforrások Minisztériuma ÚNKP-16-3 kódszámú Új Nemzeti Kiválóság Programja támogatta.

Hivatkozások

1. Öhman, S., Stevens, K.: Cineradiographic studies of speech: procedures and objectives. The Journal of the Acoustical Society of America **35** (1963) 1889
2. Bolla, K.: A magyar magánhangzók és rövid mássalhangzók képzési sajátságainak dinamikus kinoröntgenográfiai elemzése. Magyar Fonetikai Füzetek **8**(8) (1981) 5–62
3. Bolla, K., Földi, É., Kincses, G.: A toldalékcső artikulációs folyamatainak számítógépes vizsgálata. Magyar Fonetikai Füzetek **15**(4) (1985) 155–165
4. Stone, M., Sonies, B., Shawker, T., Weiss, G., Nadel, L.: Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. Journal of Phonetics **11** (1983) 207–218
5. Stone, M.: A guide to analysing tongue motion from ultrasound images. Clinical Linguistics & Phonetics **19**(6-7) (2005) 455–501
6. Schönle, P.W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., Conrad, B.: Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. Brain and Language **31**(1) (1987) 26–35
7. Mády, K.: Magyar magánhangzók vizsgálata elektromágneses artikulográffal normál és gyors beszédben. Beszédkutatás 2008 (2008) 52–66
8. Baer, T., Gore, J., Gracco, L., Nye, P.: Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. The Journal of the Acoustical Society of America **90**(2) (1991) 799–828

9. Woo, J., Murano, E.Z., Stone, M., Prince, J.L.: Reconstruction of high-resolution tongue volumes from MRI. IEEE Transactions on Bio-medical Engineering **59**(12) (2012) 3511–3524
10. Cheah, L.A., Bai, J., Gonzalez, J.A., Ell, S.R., Gilbert, J.M., Moore, R.K., Green, P.D.: A user-centric design of permanent magnetic articulography based assistive speech technology. In: Proc. BioSignals. (2015) 109–116
11. Csapó, T.G., Csopor, D.: Ultrahangos nyelvkontúr követés automatikusan: a mély neuronhálókon alapuló AutoTrace eljárás vizsgálata. Beszédkutatás 2015 (2015) 177–187
12. Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M.: Eigentongue feature extraction for an ultrasound-based silent speech interface. In: Proc. ICASSP, Honolulu, HI, USA (2007) 1245–1248
13. Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Communication **52**(4) (2010) 270–287
14. Bocquelet, F., Hueber, T., Girin, L., Badin, P., Yvert, B.: Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications. In: Proc. Interspeech. (2014) 2288–2292
15. Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., Yvert, B.: Real - time Control of a DNN - based Articulatory Synthesizer for Silent Speech Conversion : a pilot study. In: Proc. Interspeech. (2015) 2405–2409
16. Wang, J., Samal, A., Green, J.: Preliminary Test of a Real-Time, Interactive Silent Speech Interface Based on Electromagnetic Articulograph. In: Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies. (2014) 38–45
17. Denby, B., Stone, M.: Speech synthesis from real time ultrasound images of the tongue. In: Proc. ICASSP, Montreal, Quebec, Canada, IEEE (2004) 685–688
18. Hueber, T., Benaroya, E.L., Chollet, G., Dreyfus, G., Stone, M.: Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Communication **52**(4) (2010) 288–300
19. Hueber, T., Benaroya, E.l., Denby, B., Chollet, G.: Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface. In: Proc. Interspeech, Florence, Italy (2011) 593–596
20. Hueber, T., Bailly, G., Denby, B.: Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface. In: Proc. Interspeech, Portland, OR, USA (2012) 723–726
21. Hueber, T., Bailly, G.: Statistical conversion of silent articulation into audible speech using full-covariance HMM. Computer Speech and Language **36** (2016) 274–293
22. Jaumard-Hakoun, A., Xu, K., Leboullenger, C., Roussel-Ragot, P., Denby, B.: An Articulatory-Based Singing Voice Synthesis Using Tongue and Lips Imaging. In: Proc. Interspeech. (2016) 1467–1471
23. Gonzalez, J.A., Moore, R.K., Gilbert, J.M., Cheah, L.A., Ell, S., Bai, J.: A silent speech system based on permanent magnet articulography and direct synthesis. Computer Speech and Language **39** (2016) 67–87
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. (2015) http://arxiv.org/abs/1506.01497.
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105

26. Xie, S., Tu, Z.: Holistically-Nested Edge Detection. In: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE (2015) 1395–1403
27. Olaszy, G.: Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai. Beszédkutatás 2013 (2013) 261–270
28. Csapó, T.G., Deme, A., Gráczi, T.E., Markó, A., Varjasi, G.: Szinkronizált beszéd- és nyelvultrahang-felvételek a SonoSpeech rendszerrel. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017), Szeged, Magyarország (2017)
29. Mihajlik, P., Tüske, Z., Tarján, B., Németh, B., Fegyó, T.: Improved Recognition of Spontaneous Hungarian Speech—Morphological and Acoustic Modeling Techniques for a Less Resourced Task. IEEE Transactions on Audio, Speech, and Language Processing **18**(6) (2010) 1588–1600
30. Tokuda, K., Kobayashi, T., Masuko, T., Imai, S.: Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. In: Proc. ICSLP, Yokohama, Japan (1994) 1043–1046
31. 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
32. Glorot, X., Bordes, A., Bengio, Y.: Deep Sparse Rectifier Neural Networks. In Gordon, G.J., Dunson, D.B., eds.: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). Volume 15., Ft. Lauderdale, FL, USA, Journal of Machine Learning Research - Workshop and Conference Proceedings (2011) 315–323
33. Epstein, M.A., Stone, M.: The tongue stops here: ultrasound imaging of the palate (L). The Journal of the Acoustical Society of America **118**(4) (2005) 2128–31
34. Xu, K., Roussel, P., Csapó, T.G., Denby, B.: Convolutional neural network-based automatic classification of midsagittal tongue gestures using B-mode ultrasound images. submitted manuscript (2016)