Audiovizuális beszédszintézis nyelvultrahang alapon

Csapó Tamás Gábor[1]

[1] *Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék*

Abstract

In this study, we present our initial results in audiovisual text-to-speech synthesis (AV-TTS), which is a subfield of the more general areas of speech synthesis and computer facial animation. The goal of visible speech synthesis is typically to generate face motion or articulatory-related information (e.g., lip, tongue movement, or velum position). We conduct experiments in text-to-articulation prediction, using ultrasound tongue image targets. We extend a traditional deep neural network-based text-to-speech synthesis (DNN-TTS) framework by predicting ultrasound tongue images, of which the continuous tongue motion can be reconstructed in synchrony with synthesized speech. The final output is speech and ultrasound tongue video in 'wedge' orientation. We use the data of eight English speakers (roughly 200 sentences from each) from the UltraSuite-TaL dataset, train several types of deep neural networks, and show that simple DNNs are the most suitable ones for the prediction of sequential articulatory data, as we have limited training material. Objective experiments and visualized predictions show that the proposed solution is feasible and the generated ultrasound videos are mostly close to natural tongue movement but are sometimes oversmoothed. A specific application of audiovisual speech synthesis and text-to-articulation prediction is computer-assisted pronunciation training/computer-aided language learning, which can be beneficial for learners of second languages. With such an AV-TTS, by giving an arbitrary input text, one is able to hear the synthesized speech and, in synchrony with it, see (in 2D or 3D) how to move the tongue to produce target speech sounds. This visual feedback can be helpful for pronunciation training in L2 learning, especially when the target language contains speech sounds that are difficult to articulate (e.g., significantly different from the speaker's mother tongue).

Keywords: AV-TTS, mély neurális hálózatok, DNN, beszédtechnológia

1. Bevezetés

A beszédszintézis (más néven gépi szövegfelolvasás, text-to-speech, TTS) célja, hogy írott szöveget alakítsunk át emberihez hasonló beszéddé. Az audiovizuális beszédszintézis esetén nem csak beszéd a rendszer kimenete, hanem

Email address: csapot@tmit.bme.hu (Csapó Tamás Gábor)

valamilyen más kiegészítő információt, videó formájában megjeleníthető tartalmat is láthatóvá kívánunk tenni az emberi artikulációról. Ez a kiegészített tartalom lehet például beszélő fej (Czap & Mátyás, 2003, 2005), száj- és arcmozgás (Massaro et al., 2012; Schabus et al., 2014), vagy a nyelv mozgására utaló információ (Steiner et al., 2017; Le Maguer et al., 2017; Yu et al., 2019).

A beszédszintézisben napjainkban legtöbbször valamilyen gépi tanulási módszert alkalmaznak. A 2010-es évek elején a rejtett Markov-modell (HMM) alapú megoldások voltak elterjedtek (Tóth & Németh, 2008; Csapó & Németh, 2011), azonban az évtized közepétől ezeket fokozatosan felváltották a mély neurális hálózatot (Deep Neural Network, DNN) alkalmazó rendszerek (Zen et al., 2013; Zainkó et al., 2017). A neurális hálózatok rétegei igen hatékonyan képesek az adatokra jellemző tulajdonságokat megtanulni. Ez azt jelenti, hogy magukból a nyers adatokból tanulja meg a rendszer, hogy milyen absztrakcióval írhatóak le azok; nem pedig ember által megalkotott szabályokat követ. A TTS kutatások iránya a legutóbbi években az 'end-to-end' irányba halad, amely azt jelenti, hogy nyelvészeti tudás nélkül, csak a szöveges tartalom alapján rendeli hozzá a gépi rendszer az adott szöveghez legjobban illeszkedő beszédet, és minden belső komponens neurális hálózat alapú – erre példa a Tacotron2 rendszer (Shen et al., 2018). Ugyanakkor a hagyományos, nem end-to-end típusú, 'klasszikus' DNN-TTS megoldásokat is érdemes használni, amennyiben kevés a rendelkezésre álló adat, hiszen a Tacotron2 rendszer betanításához például több 10 óra nagyságrendű beszédfelvétel szükséges. Főleg akkor beszélhetünk kevés adatról, ha nem csak beszéddel, hanem valamilyen egyéb kiegészítő biológiai jellel is dolgozunk, például nyelvultrahang (Csapó et al., 2017a,b; Hajjej & Csapó, 2020), ajakvideó (Rácz & Csapó, 2020; Arthur & Csapó, 2021), vagy agyi jel (Arthur & Csapó, 2022). A jelen cikk témája, az audiovizuális beszédszintézis esetén szintén tipikusan csak kevés adat (néhány 10 percnyi felvétel) áll rendelkezésre.

1.1. Audiovizuális beszédszintézis

Az audiovizuális beszédszintézis a beszédszintézis és a számítógépes arcanimáció általánosabb területeinek egy része (Massaro et al., 2012). A vizuális

beszédszintézis területe meglehetősen jól megalapozott, és az elmúlt években számos megközelítést fejlesztettek ki (beleértve a szabályalapú (Perrier, 2014) és az adatvezérelt módszereket (Schabus et al., 2014)). A szabályalapú rendszerek magukban foglalják a beszédhangsorozat megtervezését, az izommechanizmusok és a fizikai beszédképző rendszerek modelljét is. A vokális traktus biomechanikai modelljében a nyelv egy végeselemhálóként ábrázolható (Perrier, 2014), és komplex biomechanikai szimulációk szükségesek az emberi artikulációs szervek mozgása során fellépő belső izomfeszültségi állapotok becsléséhez (Stavness et al., 2014). Az adatvezérelt megközelítésekben két fő kategória van: a képalapú rendszerek célja a valós személyről készült videó szintézise, míg a mozgásrögzítésen ('motion capture') alapuló rendszerek az arcpontokból származtatott jellemzők időbeli múlását jelenítik meg (Schabus et al., 2014). Az adatvezérelt, gépi tanulás alapú rendszerekhez olyan beszédkorpuszra van szükség, amely a beszédadatok mellett az arc vagy egyéb artikulációs szervek mozgásának adatait is tartalmazza, szinkron módon rögzítve.

1.2. Artikulációsmozgás-becslés szöveg alapján

Az audiovizuális beszédszintézis egyik típusa az az eset, amikor a szöveg alapján artikulációs mozgást (pl. ajak- vagy nyelvmozgás) szeretnénk megbecsülni a szintetizált beszéddel párhuzamosan. Ehhez speciális biojelek rögzítésére van szükség, amelyek nyomon követik az artikulációs szervek mozgását (pl. elektromágneses artikulográfia, röntgen, mágnesesrezonancia-képalkotás, ajakvideó és nyelvultrahang). Egy ilyen rendszerrel tetszőleges bemeneti szöveghez meghallgatható a szintetizált beszéd, és ezzel szinkronban megtekinthető (2D / 3D-ben), hogyan lehet mozgatni a nyelvet, hogy adott beszédhangokat állítsunk elő. Ez a vizuális visszacsatolás nagy eredményt hozhat az idegennyelv-tanulásban, különösen akkor, ha a célnyelv nehezen artikulálható beszédhangokat tartalmaz.

A legtöbb korábbi tanulmány ezen a területen pontkövető eszközöket, például elektromágneses artikulográfiát (EMA) használt (Ling et al., 2010a,b; Wei et al., 2016; Steiner et al., 2017; Le Maguer et al., 2017; Yu et al., 2019). Ling és munkatársai (2010a) egy HMM-alapú szövegből artikulációs mozgást előre-

jelző rendszert javasoltak, amely képes szintetizálni a beszélő szájmozgását. Itt még nem modellezték az időtartamokat, de egy későbbi tanulmányban az időzítési szempontokat is vizsgálták, és elemezték a kritikus artikulátorokat (Ling et al., 2010b). Wei és munkatársai (2016) DNN-eket használtak a szövegből EMA-ba történő előrejelzéshez. Steiner és munkatársai (2017) hasonlóképpen a szöveg-EMA előrejelzéssel kísérleteztek HMM-ek segítségével (szinkron szövegfelolvasóval), és célként egy geometriai 3D nyelvmodellt is beépítettek. Ezután összehasonlították a HMM-eket és a DNN-eket a szöveg-nyelv modell előrejelzéséhez (Le Maguer et al., 2017). Az eredmények szerint a DNN-ek már 2 óránál kevesebb adattal is felülmúlták a HMM-eket.

Ahogy fent látható, számos tanulmány vizsgálta a szöveg-artikulációs mozgást HMM-ekkel vagy DNN-ekkel, de mindegyik pontkövető berendezést (elektromágneses artikulográfia) használt. Csapó (2021) kezdeti kísérleteket mutatott be nyelvultrahanggal, amit a jelen cikkben az eredmények részletesebb elemzésével egészítünk ki.

1.3. A nyelvultrahang

Az ultrahangot nemzetközi szinten az 1980-as évek kezdete óta használják beszédkutatásra (Stone et al., 1983). Magyarországon az artikulációs vizsgálatok új korszaka nyílt meg az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport 2016-os megalakulásával. Attól függően, hogy a vizsgálófejet (transzdúcer) milyen helyzetben (orientációban) helyezzük az állkapocs alá, többféle irányból is vizsgálható a nyelv. A leggyakrabban a midszagittális orientációt használjuk. A midszagittális felvételek során az ultrahangtranszdúcert az áll alá helyezik; így az ultrahangjelben a legnagyobb változást a nyelv izomzatának felső határa okozza, ami az ultrahangos képeken ideális esetben jól kivehető fehér sávot eredményez.

Az ultrahangos módszer előnye a többi artikulációs rögzítési technikához képest, hogy egyszerűen használható, non-invazív, elérhető árú, valamint nagy felbontású (akár 800 x 600 pixel) és nagy sebességű (akár 100–150 képkocka/másodperc) felvétel készíthető vele. A jó térbeli felbontás azért fontos, hogy a

nyelv alakjáról minél pontosabb képet kapjunk; míg a jó időbeli felbontás ahhoz szükséges, hogy a beszédhangok képzésének gyors változását (pl. zárfelpattanás; koartikuláció) is vizsgálni tudjuk. Az ultrahang hátránya ugyanakkor, hogy a hagyományos beszédkutatási kísérletekhez a rögzített képsorozatból ki kell nyerni a nyelv körvonalát ahhoz, hogy az adatokon további vizsgálatokat lehessen végezni. Ez elvégezhető manuálisan, ami rendkívül időigényes, vagy automatikus módszerekkel, amelyek viszont ma még nem elég megbízhatóak (Csapó & Csopor, 2015; Csapó & Lulich, 2015; Whalen et al., 2019). Az ultrahang használatának bizonyos mértékig hátránya az is, hogy csak a nyelv középső részéről ad információkat, gyakran a nyelvgyök és/vagy a nyelvhegy nem látszik. Emellett előfordulhat, hogy ha a nyelv felülete közel párhuzamos az ultrahangsugárral, akkor a középső részről is hiányos az információ.

1.4. A jelen kutatás célja

A cikk célja az audiovizuális beszédszintézishez való hozzájárulás, azaz a DNN-TTS kiterjesztése artikulációsmozgás-előrejelzéssel, a nyelv ultrahangos képeinek felhasználásával. Megmutatjuk néhány beszélő adatain, hogy a kombinált TTS és szintetizált artikulációs mozgás megvalósítható és elfogadható artikulációsmozgás-videót eredményez. A szöveg alapján történő artikulációsmozgás-előrejelzés hasznos lehet a számítógéppel segített kiejtéstanulás (Computer-Assisted Pronunciation Training, CAPT) alkalmazásokhoz és az artikuláció vizualizációjához.

2. Módszerek

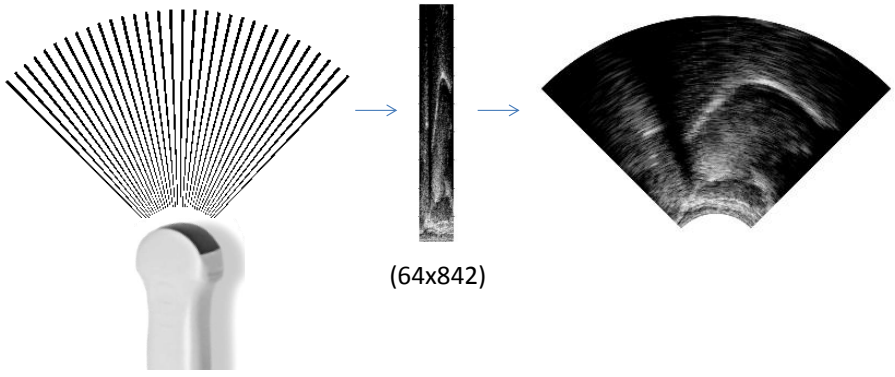
2.1. Adatok

Adatként az UltraSuite-TaL80 adatbázist használtuk fel (Ribeiro et al., 2021) (`https://ultrasuite.github.io/data/tal_corpus/`). Négy angol nyelvű férfi (03mn, 04me, 05ms, 07me) és négy női beszélőt választottunk (01fi, 02fe, 06fe, and 09fe). A beszéddel párhuzamosan a nyelv mozgását midszagittális orientációban rögzítették az Articulate Instruments Ltd. „Micro” ultrahangos

rendszerével, 81,5 képkocka/másodperc sebességgel. Az UltraSuite-TaL80-ban ajakvideót is rögzítettek, de ezt az információt nem használtuk fel a jelenlegi tanulmányban. Az ultrahangadatok és a beszédjelek szinkronizálása az Articulate Instruments Ltd. által biztosított eszközzel (Articulate Assistant Advanced, V219.08) történt. Minden beszélő közel 200 mondatot olvasott fel – a felvételek időtartama beszélőnként kb. 15 perc volt, amit 85-10-5 arányban bontottunk fel tanuló, validációs és teszt adatokra a gépi tanulási kísérletekben.

2.2. Az ultrahangos adatok feldolgozása

Kísérleteinkben a 'nyers' ultrahangos adatokból számított artikulációs jellemzőket használtuk a gépi tanulás további célpontjaként. A 'nyers' adat azt jelenti, hogy az ultrahangeszközből érkező intenzitásinformációt közvetlenül bináris formátumba mentettük (így nem veszett el adat a képpé konvertálás során), és így is dolgoztuk fel. Az 1. ábra mutatja, hogy a letapogatás hogyan történik a „Micro" rendszerrel: az ultrahangfej 64 radiális vonalon (bal oldalon), minden vonalon 842 helyen méri az intenzitást (azaz a szürkeárnyalatos színskálát), 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.



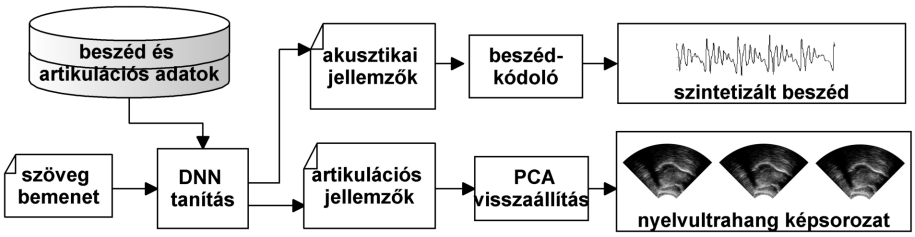
1. ábra. Nyers adatokból ultrahangkép előállítása

A 64 842 pixeles képeket átméreteztük 64 128 pixelre. Az ultrahangos kép viszonylag redundáns, ezért hatékonyan tömöríthető, ami a gépi tanulásnál előnyös lehet, hiszen csak kisebb dimenziójú adattal kell dolgoznunk. Emiatt az 'EigenTongues' (Hueber et al., 2007) módszert követve főkomponens-analízist (Principal Component Analysis, PCA) végeztünk a képeken, melynek során a képpontok varianciájának 70%-át hagytuk meg, így 128 együtthatóra tömörítve az egyes képeket. Ahhoz, hogy az artikulációs adatok szinkronban legyenek az akusztikai adatokkal, újramintavételeztük az előbbit 200 Hz-re, ami 5 ms lépésköznek felel meg.

2.3. DNN-TTS rendszer

A 2. ábra szemlélteti a javasolt megközelítést, azaz a szöveges bemenetből gépi tanulással (mély neuronhálóval) becsült kombinált akusztikus és artikulációs jellemzőket, valamint a kimeneten megjelenő szintetizált beszédet és nyelvultrahang-képsorozatot (videót). A kísérleteket a Merlin DNN-TTS keretrendszert (`https://github.com/CSTR-Edinburgh/merlin`) felhasználva végeztük (Wu et al., 2016), amely angol nyelvre kidolgozott recepteket tartalmaz. A szöveges bemenetet a rendszer először nyelvi jellemzőkké alakítja: ún. környezetfüggő címkék készülnek (Tóth & Németh, 2008; Tóth, 2013). A gépi tanulás bemenete tehát egy-egy beszédhang reprezentációja 425-dimenziós címkeként. A kimenethez a beszédből akusztikai jellemzőket számítunk (60-dimenziós spektrális együtthatók, 5-dimenziós aperiodicitás, és 1-dimenziós alapfrekvencia), míg az artikulációs adatokat a 128-dimenziós PCA komponensekkel reprezentáljuk. A kimeneti paraméterekhez delta és delta–delta jellemzőket is számít a rendszer, így a kimeneti vektor teljes dimenziója 582. A bemeneti és kimeneti jellemzőkből a Merlin rendszer idősort készít, és külön hálózatot tanítunk az időzítésekre, illetve az akusztikai/artikulációs paraméterekre. Egyszerre az idősornak egy-egy eleme megy a hálózatba, és így egy-egy 5 ms-os blokknak megfelelő jellemzőt generál.

A kísérletekben két neurálishálózat-architektúrát hasonlítottunk össze: 1) FC-DNN (fully connected deep neural network), és 2) LSTM (Long-Short Term



2. ábra. A javasolt módszer blokkdiagramja

Memory). Az FC-DNN esetén hat rejtett réteget alkalmaztunk, rétegenként 1024 neuronnal, 256-os batch mérettel, SGD optimalizálóval, batch normalization és dropout nélkül. Az LSTM esetén a háló elején négy teljesen kapcsolt réteg volt (mindegyik 1024 neuronnal), melyet egy LSTM réteg követett (512 neuronnal), batch normalization és dropout nélkül. Optimalizációnak az ADAM algoritmust alkalmaztunk, és a batch méret 256 volt. Minden esetben az MSE hibafüggvényt alkalmaztuk, és a tanítás a hibavisszaterjesztés (backpropagation) algoritmussal történt. Az előbbi, FC-DNN egyszerűbb modell és gyorsabban tanítható; az utóbbi, LSTM bonyolultabb és általában lassabb a betanítása, de 'rekurrens' típusú hálózat, azaz jobban tudja modellezni az adatok időbeliségét, mint az FC-DNN. A két hálózat összparaméterszáma hasonló volt (FC-DNN: 6,28 millió, LSTM: 6,03 millió). Mindegyik hálózatot külön-külön tanítottuk a négy férfi és négy női beszélő adatain, beszélőfüggő modelleket létrehozva.

A szintézis során a PCA-tömörített artikulációs adatokból inverz PCA transzformációval állítjuk elő a 64 128 pixeles 'nyers' elrendezésű nyelvultrahangképeket, melyeket újra átméretezünk az eredeti 64 842 pixeles méretre. Vizualizációs célból ezt a nyers adatot a fonetikai megjelenítésben is használt 'ék'/'legyező' formátumba alakítjuk át, amely megmutatja a nyelv felületének valós arányait (lásd a 4. ábrát). Utóbbi transzformációt az 'ultrasuite-tools' eszközzel végezzük (`https://github.com/UltraSuite/ultrasuite-tools`). Végül a szintetizált beszédet és a nyelvultrahang-képsorozatot folyamatos videóvá fűzzük össze, amely az audiovizuális beszédszintetizátor rendszer kimenete lesz.

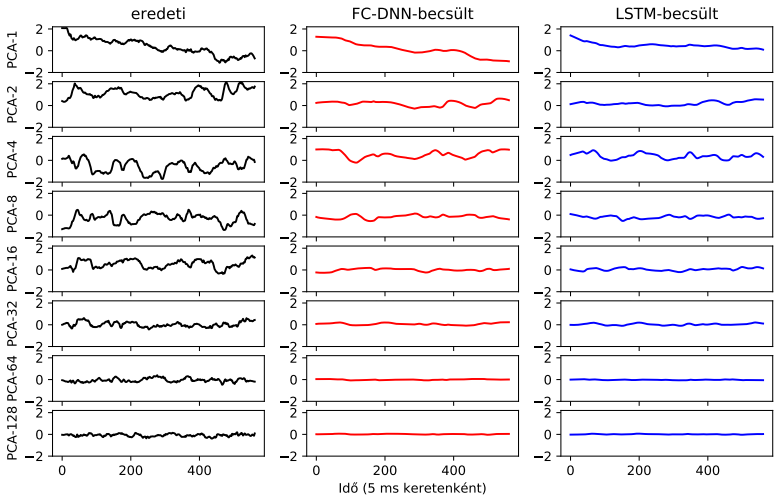
3. Eredmények

A fenti modellek betanítása után beszédet és ultrahangvideót szintetizáltunk az adatbázisok tesztelésre félretett mondatainak szövegei alapján. A validációs és teszthiba mérésére egyrészt spektrális jellemzőkkel kapcsolatos hibát (mel-kepsztrális távolság, Mel-Cepstral Distance, MCD), másrészt artikulációs jellemzőkkel kapcsolatos hibát (átlagos négyzetes hiba, Root Mean Square Error, RMSE) számoltunk. A gépi tanulás során mind az időtartam, mind az akusztikai és artikulációs modelleket betanítottuk, de a hibaszámításokhoz a tesztmondatokat az eredeti időzítésükkel szintetizáltuk. Így a hibamértékek kiszámításához nem volt szükség a jellemzők időbeli vetemítésére.

3.1. Demonstrációs minták

A becsült artikulációs jellemzőkre a 3. ábrán mutatunk egy példát ('01fi' beszélő). Mivel az artikulációs adatokat 128-dimenziós PCA jellemzőkként reprezentáljuk a gépi tanulás során, az ábrának nincs kézzelfogható, artikulációs mozgásokkal összevethető értelmezése. Amit érdemes megfigyelni, hogy az FC-DNN hálózattal becsült és az LSTM hálóval becsült görbék is követik az eredeti artikulációs jellemzők tendenciáit, de a finom részletek a DNN reprezentáció során kisimultak, eltűntek. Ez a túlsimítás jelenség gyakori a statisztikai parametrikus beszédszintézis (HMM-TTS és DNN-TTS) esetén is. Az alacsonyabb számú komponensek (pl. PCA-1, PCA-2, PCA-4) még tartalmaznak az eredeti adathoz hasonló részeket, de a magasabb dimenziójú komponensek (pl. PCA-64, PCA-128) közel konstansak, azaz utóbbiakat a neurális hálózatok nem tudták jól modellezni. Összességében a fenti ábra azt mutatja, hogy a szintetizált nyelvultrahangadatok a nagyobb változásokat (pl. a nyelv vízszintes mozgása) mutatják majd, de a finom részletek várhatóan elvesznek.

Ahhoz, hogy a generált artikulációs adatokat vizuálisan is megjelenítsük, a PCA adatokból nyelvultrahangképeket generáltunk, hiszen ezeken már láthatóvá válik a nyelv mozgása és alakja a szintetizált mondatok esetén is. A 4. ábrán az eredeti videókból és a szintetizáltakból néhány ultrahangképkockát ábrázoltunk az idő függvényében. A '01fi' beszélő esetén a bal oszlopban (eredeti-PCA)

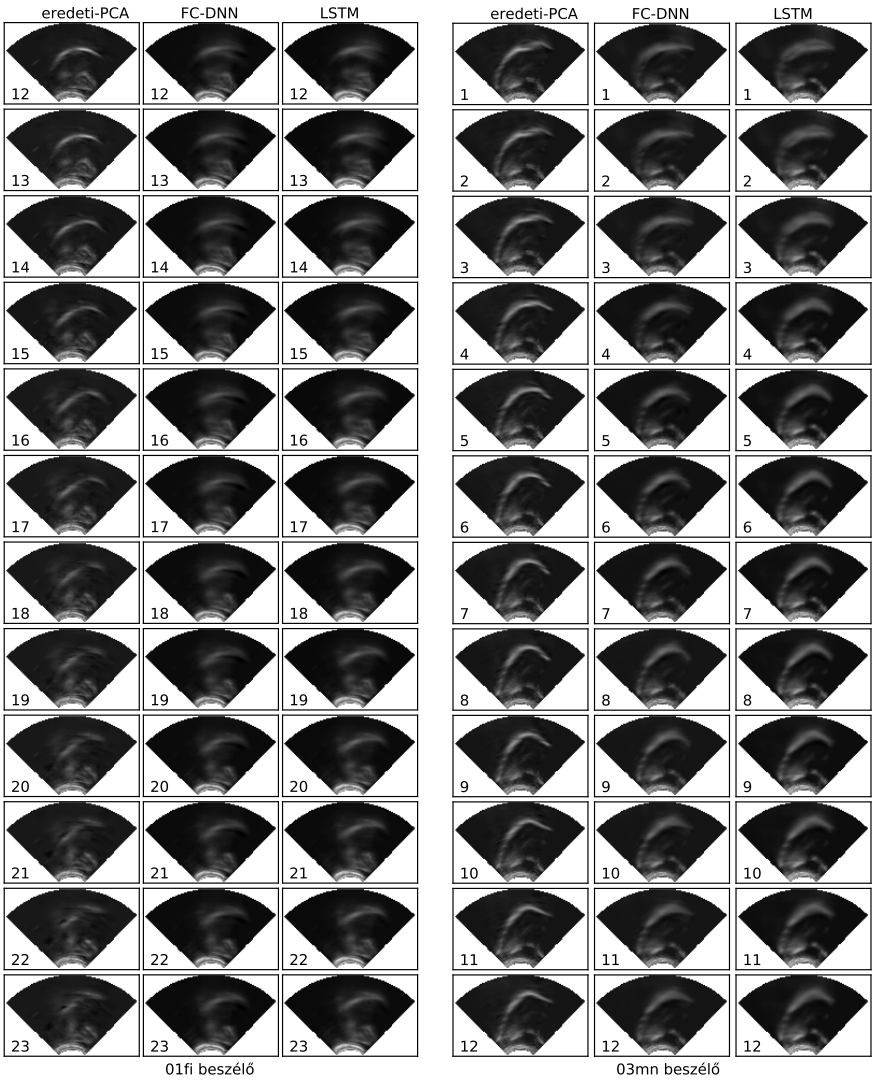


3. ábra. Eredeti és becsült artikulációs jellemzők PCA reprezentációban, '01fi' beszélőtől. Mondat: *"I leave it to nobody," said Shakespeare, sulkily.*

láthatjuk, hogy jelentős nyelvmozgás van, azaz a nyelvcsúcs (jobb oldalon) az idő múlásával feljebb mozdul: míg a kezdő, 12. képkockán még körív jellegű a nyelv látható kontúrja, a 16. képkockától a nyelv első része emelkedni kezd, majd a 20. képkockánál eléri a legmagasabb helyzetet. Mind az FC-DNN, mind az LSTM hálózat előrejelzései követik az eredeti artikulációs mozgást, de a képek helyenként elkentek, és a nyelv kontúrjának éles láthatósága csökken – ismét a statisztikai túlsimítás eredményeként. A '03mn' beszélő esetében hasonló tendenciák figyelhetők meg: a nyelv mozgásának görbülete változik az idő függvényében az eredeti felvételen, de az FC-DNN és az LSTM által előrejelzett képeken a nyelv felülete nem olyan tiszta, mint az eredetiben. Mivel a nyelv szintetizált mozgása jobban látható a valós idejű videókban, néhány mintát elérhetővé tettünk: `http://smartlab.tmit.bme.hu/besztud2022`.

3.2. Objektív mérések

Az 1. táblázat összegzi az MCD eredményeket (kisebb MCD hiba jelenti azt, hogy a szintetizált beszéd közelebb van az eredeti beszédhez). Mivel az MCD értékek logaritmikus skálájúak, ezért az átlagokat az értékek tízes alapú



4. ábra. Eredeti és becsült artikulációs jellemzők nyelvultrahangképként ábrázolva. A bal alsó sarkokban lévő számok a videó képkockáját jelentik

1. táblázat. MCD hibák a validációs/teszt halmazon

	MCD	
Beszélő	FC-DNN	LSTM
01fi	6,995 / 6,971	6,647 / 6,588
02fe	6,095 / 5,803	6,486 / 6,259
03mn	5,781 / 5,785	5,977 / 5,948
04me	5,896 / 6,024	6,318 / 6,312
05ms	6,244 / 6,256	7,235 / 7,083
06fe	5,758 / 5,582	6,444 / 6,330
07me	6,589 / 6,562	6,831 / 6,749
09fe	6,516 / 6,844	7,197 / 7,472
átlag	6,440 / 6,486	6,821 / 6,861

exponenciálisa alapján számítottuk, majd logaritmusra alakítottuk vissza. A tesztmondatok MCD értékei FC-DNN esetén 5,8-7,0 dB (átlag: 6,5 dB), míg LSTM esetén 6,0–7,5 dB (átlag: 6,9 dB) között vannak, ami azt jelzi, hogy a rekurrens neurális hálózat nem javított az akusztikai jellemzők becslésében. A 6 dB körüli MCD értékek is arra utalnak, hogy az így szintetizált beszéd minősége és természetessége gyenge. Ennek az lehet az oka, hogy csak korlátozott méretű artikulációs-akusztikus adatbázisunk van (nagyjából 200 mondat minden beszélőre), ami túl kicsi az LSTM modell betanításához.

Az artikulációs jellemzőre számított RMSE hiba eredményeit a 2. táblázat foglalja össze (kisebb RMSE hiba jelenti azt, hogy a szintetizált artikulációs mozgás közelebb van az eredetihez). A legalacsonyabb hibát a '09fe' beszélő adataival értük el: az FC-DNN-nél 2,9, míg az LSTM-nél 3,1 a tesztadatokon mért hiba. A tendencia hasonló az MCD esetéhez: az LSTM hálózat nem javított az artikulációs jellemzők előrejelzésében, valószínűleg az adatbázisok kis mérete miatt.

Az objektív mérések konklúziója tehát az, hogy a jelen kutatáshoz rendelkezésre álló, beszélőnként kb. 200 mondatot tartalmazó adathalmazok esetén az

2. táblázat. ULTPCA/RMSE hibák a validációs/teszt halmazon

	ULTPCA128/RMSE	
Beszélő	FC-DNN	LSTM
01fi	3,292 / 3,223	3,319 / 3,208
02fe	3,533 / 3,732	3,753 / 3,904
03mn	3,147 / 3,660	3,289 / 3,680
04me	3,849 / 3,985	4,031 / 4,033
05ms	3,133 / 3,233	3,249 / 3,405
06fe	3,439 / 3,250	3,743 / 3,451
07me	3,544 / 3,595	3,498 / 3,461
09fe	3,022 / 2,864	3,234 / 3,133
átlag	3,370 / 3,443	3,515 / 3,534

egyszerű FC-DNN hálózattal jobb eredményt lehet elérni, mint a bonyolultabb LSTM hálózattal.

4. Összefoglalás és következtetések

A fenti kísérletekben bemutattuk, hogy a szöveg alapján történő nyelvultrahangvideó-előrejelzés megvalósítható a hagyományos DNN-alapú beszédszintézis kiterjesztéseként, a viszonylag kis mennyiségű tanítóadat ellenére. Bár a vizuális és a beszédkimenet közötti szinkronizálást a modell nem kényszeríti ki semmilyen módon, a DNN tanítása során az akusztikai és artikulációs jellemzők összekapcsolása biztosítja, hogy a beszéd és a vizuális jellemzők szinkronban legyenek, azaz a generált ultrahangos videókban a nyelv megfelelően mozogjon a szintetizált beszédhez képest. Az akusztikai és artikulációs paraméterek összekapcsolása a gyakorlatban azt jelenti, hogy a tanításuk együttesen történik.

Bár korábban több kísérlet is történt a gépi szövegfelolvasás artikulációs adatokkal való kiterjesztésére, ezen vizsgálatok mindegyike EMA-t használt, amely egy pontkövető berendezés, és kevesebb térbeli információt tartalmaz a nyelvről, mint az ultrahang (Ling et al., 2010a,b; Wei et al., 2016; Steiner et al.,

2017; Le Maguer et al., 2017; Yu et al., 2019). Az ultrahang előnye ebben a kontextusban, hogy az eredményül kapott videó a nyelv nagyobb részét mutatja az EMA-hoz képest.

A szöveges bemenetből származó artikulációsmozgás-előrejelzés hasznos lehet az audiovizuális beszédszintézisben. Egy konkrét alkalmazás a számítógéppel segített nyelvtanulás/számítógéppel segített kiejtésgyakorlás (Katz et al., 2014; Jones, 2017; Agarwal & Chakraborty, 2019), amely hasznos lehet az idegen nyelvet tanulók számára. Egy ilyen kombinált TTS és szöveg-artikuláció becslő rendszerrel tetszőleges bemeneti szöveg megadása után hallhatóvá válik a beszéd, és ezzel szinkronban láthatóvá válik (2D-ben vagy 3D-ben), hogyan kell a nyelvet mozgatni a célhangok képzéséhez. Ez a vizuális visszajelzés hasznos lehet az idegen nyelvek kiejtésének tanulásában, különösen akkor, ha a célnyelv nehezen artikulálható (pl. a beszélő anyanyelvétől lényegesen különböző) beszédhangokat tartalmaz.

A forráskódok, a Merlin receptek, és a betanított modellek elérhetőek a következő oldalon: `https://github.com/BME-SmartLab/txt2ult`.

Köszönetnyilvánítás

A kutatást a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal OTKA programja (FK 142163 projekt), az MTA Bolyai János kutatói ösztöndíja, valamint az Új Nemzeti Kiválóság Program Bolyai+ (ÚNKP-22-5-BME-316) pályázata támogatta.

A kutatást az MTA Bolyai János kutatói ösztöndíja támogatta. A kutatás az Innovációs és Technológiai Minisztérium ÚNKP-21-5 kódszámú (azonosító: ÚNKP-21-5-BME-352) Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült. A mély neuronhálós kísérletekhez használt Titan X GPU az NVIDIA Corporation adománya. Köszönjük a CSTR kutatócsoportnak a Merlin eszköz és az UltraSuite-TaL adatbázis rendelkezésre bocsátását.

Hivatkozások

Agarwal, C., & Chakraborty, P. (2019). A review of tools and techniques for computer aided pronunciation training (CAPT) in English. *Education and Information Technologies*, *24*, 3731–3743. URL: https://doi.org/10.1007/s10639-019-09955-7. doi:10.1007/s10639-019-09955-7.

Arthur, F. V., & Csapó, T. G. (2021). Szájról olvasás automatizálása mély neurális hálózatok és mobilalkalmazás-kezelőfelületet alkalmazásával. *Beszédtudomány - Speech Science*, *2*, 7–23. doi:10.15775/Besztud.2021.7-23.

Arthur, F. V., & Csapó, T. G. (2022). Deep learning alapú agyi jel feldolgozás és beszédszintézis előkészítő munkálatai. In *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY)* (pp. 185–198). online.

Csapó, T. G. (2021). Extending Text-to-Speech Synthesis with Articulatory Movement Prediction using Ultrasound Tongue Imaging. In *Proc. ISCA SSW11* (pp. 7–12). Budapest, Hungary. doi:10.21437/SSW.2021-2. arXiv:2107.05550.

Csapó, T. G., & Csopor, D. (2015). Ultrahangos nyelvkontúr követés automatikusan: a mély neuronhálókon alapuló AutoTrace eljárás vizsgálata. *Beszédkutatás*, *23*, 176–186.

Csapó, T. G., Deme, A., Gráczi, T. E., Markó, A., & Varjasi, G. (2017a). Szinkronizált beszéd- és nyelvultrahang-felvételek a SonoSpeech rendszerrel. In *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY)* (pp. 339–346). Szeged.

Csapó, T. G., Grósz, T., Tóth, L., & Markó, A. (2017b). Beszédszintézis ultrahangos artikulációs felvételekből mély neuronhálók segítségével. In *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY)* (pp. 181–192). Szeged.

Csapó, T. G., & Lulich, S. M. (2015). Error analysis of extracted tongue contours from 2D ultrasound images. In *Proc. Interspeech* (pp. 2157–2161). Dresden, Germany.

Csapó, T. G., & Németh, G. (2011). Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval. In *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY)* (pp. 167–177). Szeged.

Czap, L., & Mátyás, J. (2003). Beszélő fej. In *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY)* (pp. 196–201). Szeged. URL: http://acta.bibl.u-szeged.hu/59393/1/msznykonf_001_196-201.pdf.

Czap, L., & Mátyás, J. (2005). Hungarian Talking Head. In *Proceedings of Forum Acusticum 4th European Congress on Acoustics* (pp. 2655–2658). Budapest, Hungary.

Hajjej, N., & Csapó, T. G. (2020). Realistic Ultrasound Tongue Image Synthesis using Generative Adversarial Networks. *Beszédtudomány - Speech Science*, *1*, 7–21. doi:10.15775/Besztud.2020.7-21.

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., & Stone, M. (2007). Eigentongue feature extraction for an ultrasound-based silent speech interface. In *Proc. ICASSP* (pp. 1245–1248). Honolulu, HI, USA.

Jones, D. (2017). *Development of Kinematic Templates for Automatic Pronunciation Assessment Using Acoustic-to-Articulatory Inversion*. Master's thesis Marquette University. URL: https://epublications.marquette.edu/theses_open/433.

Katz, W., Campbell, T., Wang, J., Farrar, E., Eubanks, J., Balasubramanian, A., Prabhakaran, B., & Rennaker, R. (2014). Opti-speech: A real-time, 3D visual feedback system for speech training. In *Proc. Interspeech* (pp. 1174–1178). Singapore, Singapore.

Le Maguer, S., Steiner, I., & Hewer, A. (2017). An HMM/DNN comparison for synchronized text-to-speech and tongue motion synthesis. In *Proc. Interspeech* (pp. 239–243). Stockholm, Sweden. doi:10.21437/Interspeech.2017-936.

Ling, Z.-H., Richmond, K., & Yamagishi, J. (2010a). An Analysis of HMM-based prediction of articulatory movements. *Speech Communication*, *52*, 834–846. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167639310001147. doi:10.1016/j.specom.2010.06.006.

Ling, Z.-H., Richmond, K., & Yamagishi, J. (2010b). HMM-Based Text-to-Articulatory-Movement Prediction and Analysis of Critical Articulators. In *Proc. Interspeech* (pp. 2194–2197). Makuhari, Japan. URL: https://www.isca-speech.org/archive/interspeech_2010/i10_2194.html.

Massaro, D. W., Cohen, M. M., Tabain, M., Beskow, J., & Clark, R. (2012). Animated speech: research progress and applications. In G. Bailly, P. Perrier, & E. Vatikiotis (Eds.), *Audiovisual Speech Processing* (pp. 309–345). Camridge, UK: Cambridge University Press. URL: http://ebooks.cambridge.org/ref/id/CBO9780511843891A024. doi:10.1017/CBO9780511843891.014.

Perrier, P. (2014). "GEPPETO": A target-based model of speech production including optimal planning and physical modeling. In *Adventures in Speech Science*. Tokyo, Japan. URL: https://hal.archives-ouvertes.fr/hal-01057251.

Rácz, B., & Csapó, T. G. (2020). Ajakvideó alapú beszédszintézis konvolúciós és rekurrens mély neurális hálózatokkal. *Beszédtudomány – Speech Science*, *1*, 57–72. doi:10.15775/Besztud.2020.57-72.

Ribeiro, M. S., Sanger, J., Zhang, J.-X. X., Eshky, A., Wrench, A., Richmond, K., & Renals, S. (2021). TaL: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos. In *2021 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1109–1116). Shenzhen, China. URL: https://arxiv.org/abs/2011.09804. doi:10.1109/SLT48900.2021.9383619. arXiv:2011.09804.

Schabus, D., Pucher, M., & Hofer, G. (2014). Joint audiovisual Hidden Semi-Markov Model-based speech synthesis. *IEEE Journal on Selected Topics in*

Signal Processing, *8*, 336–347. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6589946. doi:10.1109/JSTSP.2013.2281036.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *Proc. ICASSP* (pp. 4779–4783). Calgary, Canada. doi:10.1109/ICASSP.2018.8461368. arXiv:1712.05884.

Stavness, I., Nazari, M. A., Cormac, F., Perrier, P., Payan, Y., Lloyd, J., & Fels, S. (2014). Coupled Biomechanical Modeling of the Face, Jaw, Skull, Tongue, and Hyoid Bone. In N. Magnenat-Thalmann, O. Ratib, & H. F. Choi (Eds.), *3D Multiscale Physiological Human* (pp. 253–274). Springer London. doi:10.1007/978-1-4471-6275-9_11.

Steiner, I., Le Maguer, S., & Hewer, A. (2017). Synthesis of Tongue Motion and Acoustics from Text Using a Multimodal Articulatory Database. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *25*, 2351–2361. doi:10.1109/TASLP.2017.2756818. arXiv:1612.09352.

Stone, M., Sonies, B., Shawker, T., Weiss, G., & Nadel, L. (1983). Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. *Journal of Phonetics*, *11*, 207–218.

Tóth, B. P. (2013). *Rejtett Markov-modell alapú gépi beszédkeltés*. Phd thesis BME TMIT.

Tóth, B. P., & Németh, G. (2008). Rejtett Markov-modell alkalmazása magyar nyelvű gépi szövegfelolvasóhoz. *Beszédkutatás*, *16*, 182–193.

Wei, Z., Wu, Z., & Xie, L. (2016). Predicting articulatory movement from text using deep architecture with stacked bottleneck features. In *Proc. APSIPA* (pp. 1–6). Jeju, South Korea. doi:10.1109/APSIPA.2016.7820703.

Whalen, D. H., Kang, J., Iwasaki, R., Shejaeya, G., Kim, B., Roon, K. D., Mark, K., Tiede, Preston, J., Phillips, E., McAllister, T., & Boyce, S. (2019).

Accuracy assessments of hand and automatic measurements of ultrasound images of the tongue. In *Proc. ICPhS* (pp. 542–546). Canberra, Australia.

Wu, Z., Watts, O., & King, S. (2016). Merlin: An Open Source Neural Network Speech Synthesis System. In *9th ISCA Speech Synthesis Workshop* (pp. 202–207). Sunnyvale, CA, USA. doi:10.21437/ssw.2016-33.

Yu, L., Yu, J., & Ling, Q. (2019). BLTRCNN Based 3D Articulatory Movement Prediction: Learning Articulatory Synchronicity From Both Text and Audio Inputs. *IEEE Transactions on Multimedia*, *21*, 1621–1632. doi:10.1109/TMM.2018.2887027.

Zainkó, C., Tóth, B. P., & Németh, G. (2017). Magyar nyelvű WaveNet kísérletek. In *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY)*. Szeged.

Zen, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP* (pp. 7962–7966). Vancouver, Canada. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6639215. doi:10.1109/ICASSP.2013.6639215.