Ajakvideó alapú beszédszintézis konvolúciós és rekurrens mély neurális hálózatokkal

Rácz Bianka[1], Csapó Tamás Gábor[1,2]

[1] *Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék*
[2] *MTA-ELTE „Lendület" Lingvális Artikuláció Kutatócsoport*

Abstract

Articulatory-to-acoustic mapping methods have the aim to convert articulatory movement to acoustic speech signal. For articulatory acquisition, complex techniques (e.g. ultrasound, MRI) are suitable – but also, the lip movement contains relevant information about the speech sounds. There have been several studies applying deep neural networks for the lip-to-speech problem, and also for automatic lipreading. Inspired by the earlier studies, in this paper we designed and implemented models that can generate spectral parameters of speech from lip videos. Later, from the predicted spectral parameters, we synthesized the speech using a vocoder. For the experiments, we used 1000 sentences from a male English speaker of the GRID audiovisual database, which contains video from the face of speakers, and synchronous speech. Based on the literature, we extended the baseline deep neural network model and identified two models that use convolutional and recurrent layers. The convolutional network has single images as input, whereas the recurrent network can take into account the sequential nature of the input data: it has eight consecutive face images as input. We compared these two new models to the original baseline model in a multi-step experiment. In an objective test, we generated speech by the vocoder and by the DNN models. We calculated the Mel Cepstral Distortion between synthesized and reference sentences, and found that the recurrent model has significantly lower error than the baseline FC-DNN, while the output of the convolutional model was not better. After this we collected several subjects' opinions during an online subjective test. They had to evaluate how natural the speech utterances they heard sounded. Similarly to the objective experiment, in the subjective test the recurrent neural network (which takes eight consecutive images as input) was preferred. The results might be useful for application in Silent Speech Interfaces or for lipreading systems.

1. Bevezetés

Az artikuláció (a beszédképző szervek koordinált mozgása) és az akusztikum (a keletkező beszédjel) kapcsolata az 1700-as évek óta foglalkoztatja a beszédkutatókat (Kempelen, 1791). Ahhoz, hogy a beszédképző szervek (pl. hang-

Email addresses: rczbianka1@gmail.com (Rácz Bianka), csapot@tmit.bme.hu (Csapó Tamás Gábor)

Beszédtudomány – Speech Science 2020. 57–72. DOI: 10.15775/Besztud.2020.57-72

szalagok, nyelv, lágy szájpad) mozgását vizsgálni tudjuk, speciális eszközökre van szükségünk, mivel a legtöbb ilyen szerv nem látható folyamatosan beszéd közben. Az artikulációs szervek közül a szükséges eszközök tekintetében a legegyszerűbb az ajakmozgás vizsgálata, hiszen ehhez egy egyszerű videokamera is elegendő, amely a beszéd közbeni arcmozgást rögzíti.

Magyarországon többek között Bolla kísérletezett az ajkak (fotolabiogram) vizsgálatával, viszont a vizsgálatok csak statikus állóképeken alapultak (Bolla, 1995). Az MTA-ELTE „Lendület" Lingvális Artikuláció Kutatócsoport eszközeivel 2016 óta több magyar beszélőtől is rögzítettünk dinamikus nyelvultrahang és ajakvideó felvételeket (Csapó et al., 2017a). Emellett nemzetközi szinten egyre több adatbázis áll rendelkezésre, melyek alapján az ajakmozgás vizsgálható – pl. a GRID korpuszban 34 beszélőtől rögzítettek 1000–1000 rövid mondatot angol nyelven (Cooke et al., 2006).

1.1. Artikuláció-akusztikum átalakítás ajakvideó alapján

Az artikuláció-akusztikum konverziós módszerek célja, hogy artikulációs mozgás alapján szintetizáljanak beszédet. Az artikulációs információ lehet például a nyelv mozgása ultrahanggal rögzítve (Csapó et al., 2017c,b). A konverzió egy másik lehetséges megoldása a beszédszintézis kizárólag egy arcról vagy ajakról készült videó képkockáiból. A megoldás kétféle megközelítéssel lehetséges: 1) közvetlen 'lip-to-speech', 2) közvetett 'lip-to-text', majd 'text-to-speech' lépésekben. A közvetlen módszerek gyorsabbak, hiszen nincs szükség külön szövegfelolvasó modulra. Ezekre mutat példát Le Cornu & Milner (2015), Ephrat & Peleg (2017), és Akbari et al. (2018). A közvetett módszerek tulajdonképpen automatikus szájról olvasást végeznek, például Wand et al. (2016) és Sun et al. (2018).

A némabeszéd-interfész (Silent Speech Interface) az artikuláció-akusztikum konverziós módszerek egy olyan távlati alkalmazása, amelynek használatával némán beszélve, „tátogva" adhatunk ki hangot (Denby et al., 2010; Csapó et al., 2017c; Kimura et al., 2019). A némabeszéd-interfésszel segíthetünk olyan embereknek kommunikálni, akik egy betegség vagy baleset következtében elvesztették

a hangalkotási képességüket, viszont még tudnak artikulálni. De nem csak az egészségügyben használhatjuk ezt az eszközt. A mindennapi életben is hasznos lehet, ha egy megbeszélésen ülve hang nélkül tudunk válaszolni egy telefonhívásra anélkül, hogy megzavarnánk a társainkat.

A jelen kutatás célja egy olyan rendszer létrehozása, amely az ajakról készült videofelvételekből beszédet tud szintetizálni. Ehhez mély neurális hálón alapuló gépi tanulást alkalmaztunk, melynek bemenete az arcról készült videó volt, kimenete pedig a beszéd spektrális paraméterei. A gépi tanulás által becsült spektrális paraméterekkel egy vokóder használatával mondatokat szintetizáltunk. Az így szintetizált beszéd ugyan nem érthető teljesen, de sok esetben szótöredékek vagy szavak is érthetőek lettek, így a kezdeti eredményeket biztatónak tartjuk.

2. Módszerek

A jelen cikkben bemutatjuk Rácz (2019) szakdolgozata során készült kutatás módszereit és eredményeit.

2.1. Felvételek és adatok

A videókat a GRID adatbázisból töltöttük le (Cooke et al., 2006), amely nyilvánosan elérhető: `http://spandh.dcs.shef.ac.uk/gridcorpus/`. Minden videón egy embert látunk szemből, ahogy egy hat szavas mondatot mond el angol nyelven. A videókat 25 képkocka/sebességgel rögzítették. Minden videó 3 másodperc hosszú, és 75 képkockából áll. A GRID adatbázis 34 beszélőjéből csak egyet választottunk ki a jelen cikk demonstrációs kísérleteihez: az S3-as beszélővel készített felvételeket használtuk fel az egy-beszélős modellek tanításához. Az S3 beszélő választásának oka, hogy Ephrat & Peleg (2017) is ezen beszélő felvételeivel végzett hasonló kísérletet. Az 1. ábra bal oldalán látható egy képkocka az S3 beszélő eredeti videójából. Az 1000 videóból az utolsó 10-et választottuk a teszteléshez és a maradék 990 videó 80%-át használtuk a tanításhoz, 20%-át pedig a validáláshoz.

1. ábra. Bal oldal: egy képkocka a GRID adatbázis S3 beszélő eredeti videójából; jobb oldal: egy képkocka az előfeldolgozás után.

2.1.1. Az ajakvideó előfeldolgozása

A bemeneti képek előfeldolgozása során automatikus módszerekkel (az OpenCV modullal, 'Haar Cascades' alapon) kivágtuk az arcot a képből, és az így keletkezett képkockákat szürkeárnyalatossá, és 128×128-as méretűvé konvertáltuk. A felvételek egységességét (pl. fényerő változások) nem vizsgáltuk. Az eredményre az 1. ábra jobb oldalán látható példa. A szürkeárnyalatos képek pixelei képezték a neurális hálózatok bemenetét.

2.1.2. A beszédjel előfeldolgozása

A beszédjel paraméterekre bontására és a későbbi visszaállításra egy egyszerű vokódert választottunk, a korábbi ultrahangos kísérleteinkhez hasonlóan (Csapó et al., 2017c,b). Először spektrális elemzést végeztünk mel-általánosított kepsztrum (Mel-Generalized Cepstrum, Line Spectral Pair, MGC-LSP, Tokuda et al., 1994) módszerrel, melyet statisztikai parametrikus beszédszintézisben széles körben használnak. Az elemzéshez 12-ed rendű MGC-t számítottunk $\alpha = 0{,}42$ és $\gamma = -1/3$ értékekkel – ezen paraméterek széles körben használtak statisztikai parametrikus beszédszintézishez (Csapó & Németh, 2014; Drugman et al., 2009). Ahhoz, hogy a beszédjel analízise során kapott paraméterek szink-

ronban legyenek az arcképekkel, a kereteltolást 1 / FPS értékre választottuk (40 ms). A viszonylag nagyméretű kereteltolás (szemben a statisztikai parametrikus beszédszintézisben tipikusan használt 5 ms-os nagyságrenddel, Csapó & Németh, 2014) önmagában is ronthatja az újraszintetizált beszéd minőségét, azonban az audiovizuális adatok szinkronitásához ez volt a célszerű választás. A gépi tanulás kimenete tehát a fenti vokóder 13-ad rendű spektrális paraméterei voltak.

A beszéd visszaállításához fehérzaj gerjesztést generáltunk, majd a gerjesztést és az MGC-LSP paramétereket felhasználva MGLSADF szűrővel (Imai et al., 1983) visszaállítottuk a szintetizált beszédet (suttogás jellegű beszédet generálva). A fenti vokóder az SSI témakörében tehát úgy használható, hogy a beszéd visszaállításához a zaj gerjesztés mellett nem az eredeti spektrális paramétereket használjuk fel, hanem az arcképek alapján gépi tanulással becsülteket.

2.2. Gépi tanulás

A modellek feladata, hogy a bemenetükön kapott videókból vagy képkockákból előállítsák a beszédszintetizáláshoz szükséges együtthatókat. Az MGC-paraméterek a beszéd spektrális burkolóját írják le, a neuronháló feladata ezeknek a paramétereknek a minél pontosabb becslése volt az arckép / ajak alapján. Mivel ezek a paraméterek folytonos értékűek, ezért regressziós módban használtuk a mély hálókat. 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 modellek az átlagos négyzetes hiba (Mean Squared Error, MSE) hibafüggvényt használták. Annak érdekében, hogy elkerüljük a túltanulást, fontos, hogy a megfelelő időben állítsuk meg a tanítási folyamatot. Ezt 'early stopping' módszerrel oldottuk meg: ha a validációs hiba nem javul 5 epochon keresztül, akkor a tanítás leáll és

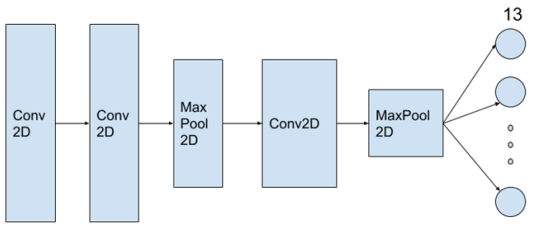
az utolsó legjobb eredményt menti el a hálózat. Automatikus hiperparaméter optimalizálást nem végeztünk.

2.2.1. Előrecsatolt mély neurális hálózat (FC-DNN alaprendszer)

Alaprendszernek egy 5 rejtett réteges, rétegenként 1000 neuront tartalmazó neuronháló struktúrát használtunk lineáris kimeneti réteggel. Az alaprendszer bemeneteként nem az arcvideó pixeleit használtuk, hanem az ezekből főkomponens-analízissel 100 dimenziósra tömörített adatokat, melyhez az EigenFaces módszert alkalmaztuk (Hueber et al., 2007). Enélkül a neuronháló nem tudta megtanulni a bemenet és kimenet közti összefüggést.

2.2.2. Konvolúciós neurális hálózat (CNN)

A konvolúciós neurális hálózatot (Convolutional Neural Network, CNN) gyakran alkalmazzák a képek osztályozására, feldolgozására. A tanulás folyamán képes megtanulni a képek különböző tulajdonságait, mint például a különböző élek, görbék kinézetét. Több egymás utáni konvolúciós rétegből és a hozzájuk tartozó aktivációs függvényekből áll. A hálózat egy képet kap a bemenetén, amelyen egy $n \times n$ méretű szűrővel csúszóablakszerűen végig haladva tömöríti a pixelekből kinyert információt. Így tanulja meg a hálózat a kis $n \times n$-es képekből az eredeti kép különböző tulajdonságait.



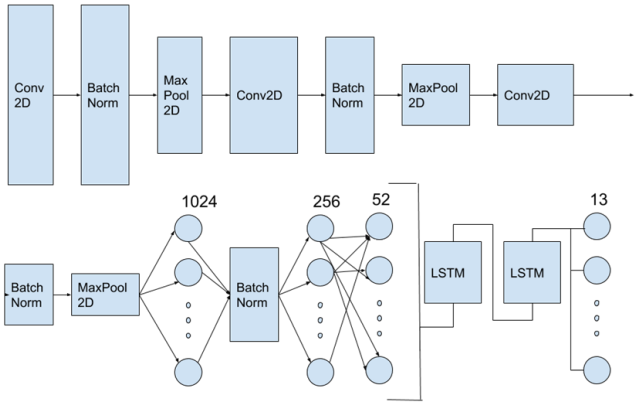
2. ábra. A CNN modell architektúrája.

Az ajakvideó-beszéd átalakításhoz második modellként egy CNN hálózatot használtunk. A hálózat a videókat képkockánként kapta meg. Így a bemeneti adatok nem tartalmazták azt az információt, hogy ezek a képek időben összefüggenek. A 2. ábrán látható a modell felépítése. Ez egy egyszerű hálózat, két

konvolúciós réteg után egy maxpool réteg, majd egy konvolúciós és egy maxpool réteg következik, végezetül pedig tizenhárom neuron adja a kimenetet, mivel a beszédszintetizáláshoz szükséges tizenhárom együtthatót kell meghatároznia a modellnek. Optimizációs algoritmusnak Adam-ot használtunk. Aktivációs függvénynek LeakyReLU-t (Leaky Rectified Linear Unit) alkalmaztunk, amelyet gyakran használnak a konvolúciós hálóknál.

2.2.3. Konvolúciós és rekurrens neurális hálózat (CNN-LSTM)

Bizonyos adatok szekvenciális formában állnak rendelkezésre, azaz az adatsorban az egymás utániság is hasznos információval bír. Például ha egy részvény jövőbeli értékét szeretnénk megjósolni, ahhoz nem elég, ha csak az utolsó nap értékét látjuk. Egy értékből nem tudjuk megállapítani, hogy a részvény ára javuló vagy romló tendenciát mutat vagy éppen stagnál. Az ilyen és ehhez hasonló esetekben hasznos lehet egy olyan mély neurális hálózat, amely rendelkezik valamilyen memóriával, amiben képes tárolni az előzőleg kapott adatokat és ezek alapján meg tudja tanulni az összefüggéseket. Az előrecsatolt hálózatokkal ellentétben a rekurrens neurális hálózatok (Recurrent Neural Networks, RNN) rendelkeznek ilyen memóriával. A Long Short-Term Memory (LSTM) a rekurrens hálózatok egy változata, amely hatékonyabban tanítható, mint a hagyományos RNN.



3. ábra. A CNN-LSTM modell architektúrája.

A harmadik modellünk a konvolúciós hálózat végére csatolt LSTM-mel egészült ki. Ettől a modelltől vártuk a legjobb eredményt, mivel a konvolúciós hálózat képes megtanulni a videó különböző jellegzetességeit, tulajdonságait, az utána kötött LSTM viszont az időbeli összefüggéseket. A hálózat egyszerre több 128×128-as képkockát is megkapott a bemenetén, ez lehetővé tette az időbeliség megtanulását. A bemeneti képek optimális darabszámát egyrészt kísérletezéssel, másrészt a korábbi hasonló kutatások tapasztalatai alapján állítottuk be (Ephrat & Peleg, 2017; Akbari et al., 2018). A vektorokból nyolc képet összefűztünk, majd ezekhez a képkockacsoportokhoz a csoport első képéhez tartozó kimenetet rendeltük. A második modell konvolúciós hálózatát kiegészítettük néhány fully-connected réteggel és két Long Short-Term Memory réteggel is. A 3. ábrán látható a modell felépítése. A fully-connected rétegekre azért volt szükség, mert a hálózat enélkül nem volt képes tanulni a rétegek közti nagy paraméterszám különbség miatt. A hálózat működését kipróbáltuk az Adam és az SGD optimizációs algoritmusokkal is és végül az SGD-t használtuk, mivel így a tanítások jobb eredményeket mutattak.

2.3. Kiértékelési módszerek

A tesztelés folyamán mindhárom modellnek el kellett végeznie egy predikciót egy-egy olyan videón, amit sem a tanító és sem a validációs halmazban nem látott még. A tesztek előkészítéseként a kódoló függvénnyel elvégeztük ugyanazt a kódolást a bemeneten, mint a tanító adatok esetén, majd a megfelelő bemeneti struktúrába rendeztük őket. Itt a hálózat már nem kapja meg az elvárt kimenetet, mivel neki kell egy predikciót készítenie.

A következő lépésekben úgy ellenőriztük le a modellek helyességét, hogy a tanítás során keletkezett tanító és validációs hibákat kísértük figyelemmel, majd a kimenetként kapott együtthatókból hangot generáltunk és meghallgattuk azokat. Hasonlóképpen vetettük össze a modelleket egymással is: megkerestük, hogy melyik hálózatnál volt a legkisebb a validációs hiba értéke, és hogy melyik hálózat hány epochig volt képes tanulni, mielőtt az 'early stopping' leállítot-

ta volna. A generált audió fájlokat meghallgatva összevetettük, hogy melyik modell érte el a legjobb eredményt.

A tesztelés előtt össze kellett állítani egy teszthalmazt, amely segítségével össze tudjuk hasonlítani a modelleket. Kiválasztottunk tíz olyan videót, amelyet nem használtunk fel előtte sem a tanításhoz, sem a validációhoz, így a hálózatok számára teljesen ismeretlenek voltak. Ezeket a videókat megkapták a hálózatok és a kimenetként kapott együtthatókból hangot generáltunk. Három modellt használtunk fel a tesztelés során: az előrecsatolt hálózatot (EigenFaces bemenettel), a konvolúciósat (amely képkockákból tanult), és a Long Short-Term Memory rétegeket használót (amelynek nyolc egymás utáni képkocka volt a bemenete). A tesztelést kétfelé bontottuk: objektív és szubjektív tesztelésre.

2.3.1. Objektív kiértékelés

Az objektív kiértékelés során olyan mérőszámot kerestünk, amellyel pontosabban meg lehet határozni a modellek egymáshoz viszonyított eredményességét, mint a validációs hiba értékével. Az objektív teszt során két-két hangfájl közötti spektrális távolságot (Mel Cepstral Distortion, MCD) számoltunk, Kubichek (1993) alapján. Az MCD-t a beszédszintetizáló rendszerek minőségének felmérésére használják. Minél kisebb az MCD értéke a szintetizált és a természetes beszéd között, annál jobban sikerült a szintetizált beszédnek reprodukálnia a természetes beszédet.

2.3.2. Szubjektív meghallgatásos teszt

A szubjektív tesztelés során az erre a célra készített internetes teszt (`http://leszped.tmit.bme.hu/rb2019/`) kitöltői is meghallgatták a generált hangokat. Az ajakvideó-beszédszintézis kutatásának célja, hogy a jövőben az emberek könnyedén használhassák a vele alkotott szolgáltatásokat, termékeket. Így a tesztelés folyamán a felhasználói élmény felmérése kiemelten fontos, hisz lehet bármilyen hasznos egy alkalmazás, ha a felhasználóknak kényelmetlen, nehézséget okoz a használata, akkor nem fogják használni. Éppen ezért egy szubjektív hallgatásos tesztet végeztettünk el, hogy felmérjük melyik modell hogyan tel-

jesít. A teszt során a teszthalmazban szereplő videókból szintetizált hangokat kellett meghallgatniuk a kitöltőknek. Egy-egy oldalon mindegyik hangminta ugyanahhoz a videóhoz tartozott; MUSHRA-jellegű tesztként (ITU-R Recommendation BS.1534). Referenciaként szerepelt az eredeti hangfájl is. A kitöltők feladata az volt, hogy minden egyes mintát osztályozzanak egy skálán aszerint, hogy mennyire hangzik természetesnek az adott beszéd (0: teljesen természetellenes, 100: teljesen természetes). A tesztben a fent említett három hálózat által szintetizált hangok szerepeltek (az eredeti mondatok F0-ját megtartva a szintézis során), valamint az eredeti beszéd és a vokóderrel szintetizált is; mindegyik rendszerből 10–10 mondat. A kitöltők nem tudták, melyik minta melyik modellhez tartozott és a sorrendjük tesztesetenként meg is volt keverve. A szubjektív teszt egy másik felületén figyelemmel követhettük az eredményeket. A tesztet összesen hét kísérleti alany végezte el (hat férfi és egy nő; 23–45 évesek, átlagos életkor: 31 év; egyikük sem volt beszédtechnológiai szakértő). Itt láthattuk, hogy melyik modell átlagosan milyen értékelést kapott az egyes tesztesetekben vagy a teszt egészén.

3. Eredmények és diszkusszió

3.1. Objektív kiértékelés

Összehasonlítási alapnak nem a természetes beszédet választottuk, mivel a tanítóminták előkészítése során az audió fájlból egy kódoló segítségével nyerjük ki az együtthatókat, amikből majd egy dekódolóval újra hangot generálunk. Ez a vokóder torzítja az eredeti hangot, így a tanítás során is jelen van ez a torzítás. Ezért a referenciának a teszthalmazban szereplő eredeti hangfájlokból vokóderrel szintetizált beszédet választottuk. Ezt hasonlítottuk össze a már említett három modellel. Modellenként mind a tíz tesztmintára lefuttattuk az MCD számítást. Az 1. táblázatban láthatóak a teszt eredményei. Az MCD-t használva annál jobb eredményről beszélhetünk, minél kisebb a kapott érték. Az egész teszt legjobb értékét a CNN-LSTM érte el, átlagosan 4,05-öt. Ezután következett az alaprendszer (MCD: 5,20), majd végül a konvolúciós hálózat

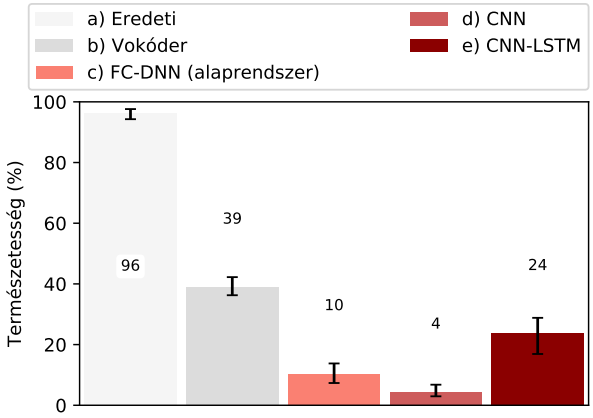
1. táblázat. A különböző modellek MCD értékei az egész tesztet nézve

mondat	Előrecsatolt neurális hálózat (FC-DNN, alaprendszer)	Konvolúciós neurális hálózat (CNN)	Konvolúciós és rekurrens neurális hálózat (CNN-LSTM)
1	5,01	6,15	3,74
2	5,45	6,31	3,88
3	5,31	6,29	4,28
4	5,51	7,05	3,81
5	4,72	5,15	3,61
6	4,27	6,08	3,73
7	5,57	6,47	4,12
8	5,94	6,54	3,70
9	5,23	6,80	4,84
10	5,26	5,63	4,81
átlag	5,20	6,25	4,05

(MCD: 6,25). A konvolúciós hálózat valószínűleg azért teljesített rosszul, mert a bemeneti képek túl nagy mennyiségű információt tartalmaznak, amiből nem tudta hatékonyan megtalálni a spektrális paraméterekkel való összefüggést. Az alaprendszer EigenFaces dimenziócsökkentő eljárást használt, így ott ez nem fordult elő. A CNN-LSTM pedig azért eredményezhetett kisebb hibát, mert ott a nyolc egymás utáni kép összefűzéséből származó információ kompenzálni tudta a konvolúciós rétegeket.

3.2. Szubjektív meghallgatásos teszt

A 4. ábra megmutatja, hogy a szubjektív meghallgatásos tesztben átlagosan milyen értékelést kaptak a modellek az egész tesztre vetítve (a 95%-os konfidenciaintervallumokat is feltüntetve). Az 'eredeti' címkéjű hangot használtuk referenciának. A teszt során a maximális 100 ponthoz nagyon közeli értékeket ért el, tehát a kitöltők is azt a mintát tartották a legtermészetesebbnek, ami a valóságban is az. A tesztből kiderült, hogy a vokóder jelentősen rontja a beszéd minőségét, a kitöltőktől átlagosan 39 pontot kapott a 'vokóder' típus.



4. ábra. A különböző modellek átlagos szubjektív értékei az egész tesztet nézve (a 95%-os konfidenciaintervallumokat is feltüntetve).

Mivel a tanítás során ennek a segítségével szintetizáljuk a beszédet, így várható volt, hogy egyik modell sem fog magasabb pontszámot elérni. Az általunk alkotott hálózatok közül, ebben a tesztben is a CNN-LSTM modell érte el a legjobb eredményeket. A teljes tesztre kiszámolt átlagból látszik, hogy a hálózatnak van még hova fejlődnie, a vokódertől nagyjából 15 ponttal van lemaradva. A legrosszabbul a csak konvolúciót használó hálózat teljesített. Ez a modell nagyon kicsi, 5-höz közeli pontszámot kapott.

Összegezve megállapíthatjuk, hogy a CNN-LSTM hálózat sokkal tisztább, kevésbé zajos hangokat képes szintetizálni, mint a képkockasorokat használó konvolúciós hálózat.

A szakirodalmi áttekintés során találtunk néhány hasonló kutatást, melyek a 'lip-to-speech' témakörrel foglalkoztak. Le Cornu & Milner (2015) az arcról készült képek előfeldolgozásával próbált jobb eredményeket elérni, míg mi ezt a feldolgozást a neurális hálózatokra bíztuk. Ephrat & Peleg (2017) csak konvolúciós hálózatot használt, míg mi rekurrens módszereket is teszteltünk. Akbari et al. (2018) egy komplex beszédkódolót használt a spektrális paraméterekből beszéd szintéziséhez; a saját fenti kísérletekben pedig egy egyszerű vokódert alkalmaztunk.

4. Következtetések

A kutatásban az ajakvideó alapú beszédszintézisre mutattunk be egy kísérletet. Konvolúciós és rekurrens neurális hálózat architektúrákat teszteltünk. Megtapasztaltuk, hogy milyen fontos az adatok megfelelő ismerete és előfeldolgozása, hogy milyen problémák adódhatnak. A hálózatok teljesítményét több lépcsős teszteléssel össze is hasonlítottuk. Először egy objektív teszten összehasonlítottuk a Mel Cepstral Distortion érték szerint a különböző modelleket. Majd egy szubjektív hallgatásos tesztet töltettünk ki néhány emberrel, ahol a hallott mintákat kellett értékelniük aszerint, hogy mennyire érzik természetesnek őket. A tesztekből egyértelműen kiderült, hogy a CNN-LSTM hálózat érte el a legjobb eredményt. Bár a modelljeink folyamatosan fejlődtek, még a CNN-LSTM hálózat által szintetizált beszéd sem érthető teljesen, de szótöredékek felismerhetőek. Az eredmények alkalmazhatóak lehetnek némabeszéd-interfészekben vagy automatikus szájról olvasó rendszer kidolgozásához (Sun et al., 2018).

Egy továbbfejlesztési lehetőség egy teljesen más architektúra, mint például a Generative Adversarial Network (GAN) (Goodfellow et al., 2014) típusú hálózat használata lehetne. CNN hálózat esetében lehetséges 8 keretet felhasználni a bemenethez, akár 2D konvolució esetén 8 csatornával, vagy 3D konvolució felhasználásával (Tóth & Shandiz, 2020). További lehetőség a rekurrens hálózat seq2seq módon történő tanítása (encoder-decoder architektúrával), amely az eredményeket javíthatja, mivel a hosszú távú információ is a hálózat rendelkezésére áll (Sutskever et al., 2014).

Köszönetnyilvánítás

A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta (FK 124584 és PD 127915 projektek). Köszönjük a meghallgatásos teszt résztvevőinek a teszt kitöltését.

Hivatkozások

Akbari, H., Arora, H., Cao, L., & Mesgarani, N. (2018). LIP2AUDSPEC : Speech reconstruction from silent lip movements video. In *Proc. ICASSP* (pp. 2516–2520). Calgary, Canada.

Bolla, K. (1995). *Magyar fonetikai atlasz. A szegmentális hangszerkezet elemei*. Budapest: Nemzeti Tankönyvkiadó.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, *120*, 2421–2424. URL: `http://asa.scitation.org/doi/10.1121/1.2229005`. doi:`10.1121/1.2229005`.

Csapó, T. G., Deme, A., Gráczi, T. E., Markó, A., & Varjasi, G. (2017a). Synchronized speech, tongue ultrasound and lip movement video recordings with the "Micro" system. In *Challenges in analysis and processing of spontaneous speech*.

Csapó, T. G., Grósz, T., Gosztolya, G., Tóth, L., & Markó, A. (2017b). DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. In *Proc. Interspeech* (pp. 3672–3676). Stockholm, Sweden. URL: `http://dx.doi.org/10.21437/Interspeech.2017-939`. doi:`10.21437/Interspeech.2017-939`.

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

Csapó, T. G., & Németh, G. (2014). Statistical parametric speech synthesis with a novel codebook-based excitation model. *Intelligent Decision Technologies*, *8*, 289–299.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, *52*, 270–287. URL: `http://dx.doi.org/10.1016/j.specom.2009.08.002`. doi:`10.1016/j.specom.2009.08.002`.

Drugman, T., Wilfart, G., Moinet, A., & Dutoit, T. (2009). Using a Pitch-Synchronous Residual Codebook for Hybrid HMM/frame Selection Speech Synthesis. In *Proc. ICASSP* (pp. 3793 – 3796). Taipei, Taiwan.

Ephrat, A., & Peleg, S. (2017). Vid2speech: Speech Reconstruction from Silent Video. In *Proc. ICASSP* (pp. 5095–5099). New Orleans, LA, USA. URL: http://arxiv.org/abs/1701.00495. arXiv:1701.00495.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)* (pp. 2672–2680). URL: https://papers.nips.cc/paper/5423-generative-adversarial-nets.

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.

Imai, S., Sumita, K., & Furuichi, C. (1983). Mel Log Spectrum Approximation (MLSA) filter for speech synthesis. *Electronics and Communications in Japan (Part I: Communications)*, *66*, 10–18. URL: http://doi.wiley.com/10.1002/ecja.4400660203. doi:10.1002/ecja.4400660203.

Kempelen, F. (1791). *Az emberi beszéd mechanizmusa, valamint a szerző beszélőgépének leírása [Eredeti cím: Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine]*. Bécs, Ausztria: Degen.

Kimura, N., Kono, M. C., & Rekimoto, J. (2019). Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). Glasgow, UK. doi:10.1145/3290605.3300376.

Kubichek, R. F. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proc. ICASSP* (pp. 125–128). Victoria, Canada. doi:10.1109/pacrim.1993.407206.

Le Cornu, T., & Milner, B. (2015). Reconstructing intelligible audio speech from visual speech features. In *Proc. Interspeech* (pp. 3355–3359). Dresden, Germany.

Rácz, B. (2019). *VID2SPEECH: beszédszintézis ajak videóból konvolúciós és rekurrens mély neurális hálózatokkal*. Technical Report BME TMIT. URL: https://diplomaterv.vik.bme.hu/hu/Theses/VID2SPEECH-beszedszintezis-ajak-videobol.

Sun, K., Yu, C., Shi, W., Liu, L., & Shi, Y. (2018). Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *UIST 2018 - Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (pp. 581–593). Berlin, Germany. URL: http://dl.acm.org/citation.cfm?doid=3242587.3242599. doi:10.1145/3242587.3242599.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* (pp. 3104–3112). arXiv:1409.3215.

Tokuda, K., Kobayashi, T., Masuko, T., & Imai, S. (1994). Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. In *Proc. ICSLP* (pp. 1043–1046). Yokohama, Japan.

Tóth, L., & Shandiz, A. H. (2020). 3D Convolutional Neural Networks for Ultrasound-Based Silent Speech Interfaces. In *Proc. ICAISC*. Zakopane, Poland.

Wand, M., Koutník, J., & Schmidhuber, J. (2016). Lipreading with long short-term memory. In *Proc. ICASSP* (pp. 6115–6119). Shanghai, China.