3D konvolúciós neuronhálón és neurális vokóderen alapuló némabeszéd-interfész

Tóth László[1], Amin Honarmandi Shandiz[1], Gosztolya Gábor[2], Zainkó Csaba[3], Markó Alexandra[4,5], Csapó Tamás Gábor[3,4]

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

Kivonat A némabeszéd-interfészek célja beszédjel előállítása valamilyen, az artikulációs szervek mozgását rögzítő felvételből, például a nyelvmozgást tartalmazó ultrahang-videóból. Jelenleg erre a konverzióra a mély neuronhálókat alkalmazó megoldások tűnnek a legígéretesebbnek. Képek felismerésére már régóta alkalmazzák a konvolúciós neuronhálókat, a legjobb eredményt azonban akkor kaphatjuk, ha a videó egyes képkockáit nem külön-külön, hanem sorozatként dolgozzuk fel. Egy lehetséges megoldás erre, ha a képeket feldolgozó konvolúciós háló kimeneteinek sorozatát egy visszacsatolt neuronhálóval egyesítjük. Jelen cikkben viszont egy másik megoldással próbálkozunk, nevesül 3-dimenziós konvolúciós hálókat használunk, ahol a képek két dimenziója mellett az idő képezi a harmadik tengelyt. A 3D konvolúciós hálóknak is egy speciális változatát alkalmazzuk, amely a térbeli és időbeli konvolúciós lépéseket felbontott formában végzi el – ezt a fajta hálózatot sikeresen használták már más videófelismerési feladatokban is. Kísérleteinkben a 3D neuronháló némileg pontosabb eredményeket adott, mint a kombinált konvolúciós+visszacsatolt modell, ami azt mutatja, hogy ez a megközelítés alternatívája lehet a rekurrens hálókra épülő, általában lassabban és nehézkesebben tanítható modelleknek.
Kulcsszavak: némabeszéd-interfész, CNN, 3D konvolúció, neurális vokóder, nyelv-ultrahang

1. Bevezetés

Az utóbbi években megnőtt az érdeklődés az ún. „articulatory-to-acoustic" átalakítás iránt, amelynek célja az elhangzott beszéd becslése, visszaállítása pusztán az artikulációs szervek mozgása alapján. Ennek a leképezésnek a megoldása tudná a technológiai hátteret nyújtani olyan alkalmazások számára, mint például a némabeszéd-interfész (Silent Speech Interface, SSI (Denby és mtsai, 2010;

Schultz és mtsai, 2017)). Az artikulációs szervek mozgásának valamilyen felvétele alapján ugyanis elvileg olyan esetben is meg lehetne becsülni a beszédjelet, amikor az alany valójában nem is ejt ki hangot, azaz „némán beszél". Az ilyen némabeszéd-interfészek segítségével vissza tudnánk adni olyan betegek beszédkészségét, akik artikulációs szerveiket ugyan képesek mozgatni, de a hangadás képességét elvesztették (például a gégéjüket, hangszalagjaikat érintő műtét vagy sérülés következtében). Emellett olyan alkalmazási területek is felmerülnek, amikor a hangos beszédkommunikáció valamilyen más okból nem lehetséges (pl. nagyon zajos környezetben vagy bizonyos katonai alkalmazásokban). Az artikulációs szervek mozgásának követésére többféle megoldás létezik, a legegyszerűbb (bár nem teljes értékű) ezek közül az ajkak mozgásának videóra rögzítése (Ephrat és Peleg, 2017; Akbari és mtsai, 2018). További lehetőségként kínálkozik az elektromágneses artikulográfia (electromagnetic articulography (EMA), Kim és mtsai (2017a,b)), az ultrahangos nyelvkövetés (ultrasound tongue imaging (UTI), Jaumard-Hakoun és mtsai (2016); Csapó és mtsai (2017); Grósz és mtsai (2018); Kimura és mtsai (2019)) az állandó mágnessel készült artikulográfia (permanent magnetic articulography (PMA), Gonzalez és mtsai (2017)). Az artikulációs izmok elektromiográfiás figyelése (surface electromiography (sEMG), Maier-Hein és mtsai (2005); Janke és mtsai (2012); Janke és Diener (2017)) is lehetséges, illetve több szerző a fenti módszerek párhuzamos, kombinált használatával próbálkozik (Denby és mtsai, 2010). Mi ebben a cikkben a nyelvmozgásról készült ultrahang-videókból fogunk kiindulni.

Az artikulációs mozgásról készült felvételek beszédjellé konvertálásának konvencionálisabb módja a kétlépéses, felismerésből majd szintézisből álló eljárás (Schultz és mtsai, 2017). Mint a neve is mutatja, ez a megközelítés a rögzített jel alapján első lépésben megkísérli felismerni az elhangzott beszédet, majd a felismert szövegből beszédszintézis útján állítja elő a beszédjelet (Denby és mtsai, 2011; Hueber és mtsai, 2010; Wang és mtsai, 2014). Ennek a módszernek a fő hátránya, hogy egyrészt nagy késleltetés keletkezhet az input és az output között, másrészt pedig a beszédfelismerő hibái zavaró módon megjelennek a szintetizált beszédben. További probléma, hogy az esetleges prozódiai információt teljesen elveszítjük az eredeti jelből, pedig a fő prozódiai komponensek – szünetek, hangerő, sőt még az alapfrekvencia is – egész jól rekonstruálhatóak az artikulációs felvételből (Grósz és mtsai, 2018).

A fentiek miatt a jelenlegi SSI megoldások a közvetlen szintézis elvét preferálják, azaz az artikulációs jelet közvetlenül beszédjellé próbálják alakítani, bármiféle közbülső lépés nélkül. A közvetlen leképezés elfogadható minőségű megoldását a mély neuronhálós (Deep Neural Network, DNN) technológia elterjedése tette lehetővé. Az „articulatory-to-acoustic" leképezéssel próbálkozó legújabb cikkek mindegyike mély neuronhálós technológiát alkalmaz, bármilyen jelrögzítési módszerről legyen is szó (Csapó és mtsai, 2017; Grósz és mtsai, 2018; Janke és Diener, 2017; Jaumard-Hakoun és mtsai, 2016; Gonzalez és mtsai, 2017; Moliner és Csapó, 2019; Kimura és mtsai, 2019; Saha és mtsai, 2020).

Jelen cikkben mély neuronhálókat fogunk alkalmazni a nyelvmozgást rögzítő ultrahang-videók beszédjellé alakítására. Habár az ugyanezen problémával fog-

lalkozó legkorábbi cikkek a legegyszerűbb felépítésű, ún. teljes kapcsolású (fully connected) neuronhálókat alkalmaztak (Jaumard-Hakoun és mtsai, 2016; Csapó és mtsai, 2017), mivel a videóink képekből állnak, ésszerűbbnek ígérkezik konvolúciós neuronhálókat (convolutional neural network, CNN) használni. A konvolúciós hálók rendkívül sikeresek a kép alakfelismerésben (Krizhevsky és mtsai, 2012), és számos újabb, SSI-vel foglalkozó tanulmány már konvolúciós hálókra épül (Janke és Diener, 2017; Moliner és Csapó, 2019; Kimura és mtsai, 2019).

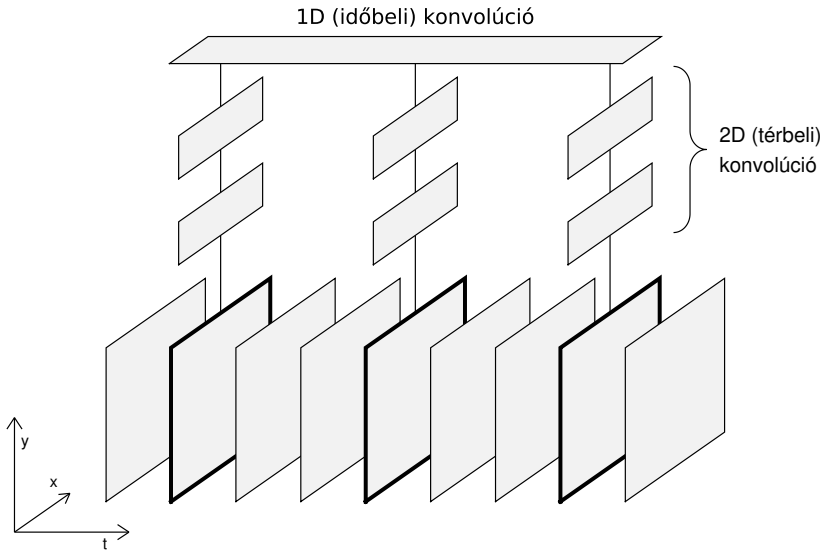
Esetünkben azonban az input egy videófelvétel, azaz nem egyetlen állókép, hanem képek sorozata. Egyetlen statikus képhez viszonyítva a képek sorozata nyilván extra információt tartalmaz a nyelvmozgás irányáról, dinamikájáról, tehát érdemes lehet egyidejűleg több szomszédos képből álló blokkokat feldolgozni. Sorozatok feldolgozására több neuronhálós struktúra is szóba jöhet. Ilyen esetben tipikusan visszacsatolt neuronhálókat szoktunk alkalmazni, például „long short-term memory", röviden LSTM neuronhálót. Mivel az input képekből áll, ezért sztenderdnek mondható megoldás az egyes képeket egy (kétdimenziós) konvolúciós hálóval feldolgozni, majd az egyedi képekből kinyert információt egy LSTM hálóval összevonni (Gonzalez és mtsai, 2017; Moliner és Csapó, 2019; Liu és mtsai, 2018; Kim és mtsai, 2017b). Erre a sémára röviden „CNN+LSTM" modellként fogunk hivatkozni.

Habár az LSTM hálók jól beváltak, folyamatos kritika éri őket a betanításuk lassúsága és nehézkessége miatt. Mivel a visszacsatoló kapcsolatok miatt tanításuk nehezen párhuzamosítható, folyamatos a kutatás a hasonló hatásfokkal működő, de tisztán előrecsatolt hálózatok iránt, például a gépi fordítás (Lakew és mtsai, 2018) vagy a beszédfelismerés (Dong és mtsai, 2018) terén. Esetünkben ilyen alternatívaként merül fel a kétdimenziós konvolúciós háló (2D CNN) kiterjesztése három dimenzióra, ahol a két térbeli kiterjedés mellé az időtengely adja a harmadik dimenziót (Ji és mtsai, 2013; Kimura és mtsai, 2019; Wu és mtsai, 2018). A cikkben ezzel a megoldással fogunk kísérletezni, méghozzá egy speciális 3D hálóstruktúrát, ún. „(2+1)D CNN"-t használva (Tran és mtsai, 2018). Mint látni fogjuk, az eredmények azt mutatják, hogy ez a háló a visszacsatolt hálókéval ekvivalens eredményeket tud elérni ugyanannyi paraméter és rövidebb tanítási idő mellett. Bár további alapos összehasonlító vizsgálatokra lesz szükség, azt mindenképpen kimondhatjuk, hogy az ultrahang-videón alapuló némabeszéd-interfészek esetén a 3D-konvolúcióra épülő előrecsatolt hálózatok a visszacsatolt LSTM-hálózatok életképes alternatívájának tűnnek.

A cikk a következőképp épül fel. A 2. fejezet az alkalmazandó konvolúciós háló technikai részleteit mutatja be. A 3. fejezetben ismertetjük a beszédjeleken és ultrahang-videókon alkalmazott adatkinyerési és -feldolgozási lépéseket, a 4. fejezetben pedig a kísérletek paramétereit. Az 5. fejezet az eredményeket mutatja be és vitatja meg, majd a cikket a levont konklúziókkal zárjuk a 6. fejezetben.

2. Videók feldolgozása konvolúciós neuronhálókkal

Az 'Alexnet' hálózat feltalálása óta az állóképek felismerésében a konvolúciós háló számít a legjobb technológiának (Krizhevsky és mtsai, 2012). Ezek a CNN



1. ábra: A (2+1)D CNN háló szerkezetének illusztrációja. A videó képkockáit (alul) először 2D konvolúciót alkalmazó neurális rétegek dolgozzák fel, majd ezek kimenetét egy 1D konvolúciót végző réteg fogja össze. A modell átugorhat képkockákat; az ugrás lépésköze az időbeli konvolúció „stride" paraméterével szabályozható.

hálózatok a konvolúciót két dimenzióban, a kép két térbeli kiterjedése mentén végzik. Számos alkalmazásban azonban az input egy videó, nem pedig egyetlen állókép. Ilyen esetben a képkockákat sorozatként feldolgozva (külön-külön feldolgozás helyett) gyakran jelentősen javíthatók a felismerési eredmények. A legjobb példa erre az emberi járásmód (gait) felismerése, de általánosságban felhozható bármilyen mozgási esemény detektálása (action recognition, Ji és mtsai (2013); Zhao és mtsai (2018a,b)). Ezekben az esetekben a képkockák sorozata háromdimenziós tömbként fogható fel, ahol a két térbeli dimenzió mellé az idő adja a harmadik dimenziót (lásd 1. ábra).

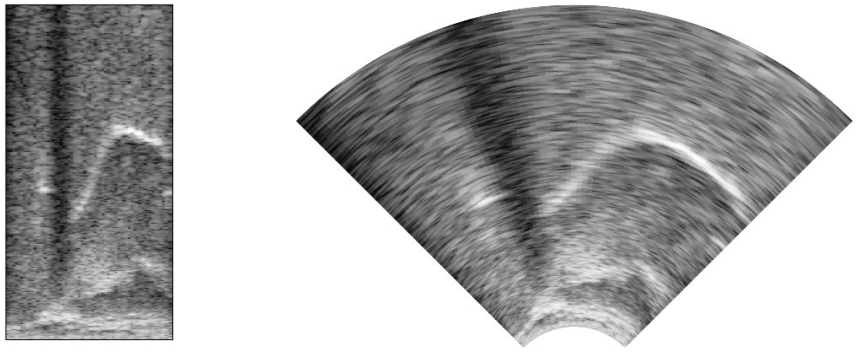
Sorozatok feldolgozásában hagyományosan a visszacsatolt neuronhálók, például az LSTM hálók számítanak a leghatékonyabb, legcélravezetőbb hálózattípusnak (Hochreiter és Schmidhuber, 1997). Ezen hálók betanítása azonban közismerten nehézkes és lassú, ezért több egyszerűsített változatuk is készült, mint például a GRU (gated recurrent unit (Cho és mtsai, 2014)) vagy a „kvázirekurrens" háló (Bradbury és mtsai, 2017). Emellett folyamatos kutatás folyik olyan struktúrák iránt, amelyek hasonló pontosságot tudnak elérni, de visszacsatolt kapcsolatok nélkül. Beszédfelismerésben például az időkésleltetéses háló (time-delay neural network, TDNN) bizonyult kifejezetten sikeresnek (Peddinti és mtsai, 2015; Tóth, 2014), de megemlíthetjük az előrecsatolt szekvenciális memóriájú hálót is (feedforward sequential memory network, FSMN) (Zhang és mtsai, 2018). Videók felismerésére pedig a CNN struktúra többféle, soroza-

tok feldolgozására alkalmas kiterjesztését javasolták (Ji és mtsai, 2013; Zhao és mtsai, 2018a,b). Sajnos a szokványos, 2D konvolúciót sokféleképp lehet kiterjeszteni három dimenzióra, így az optimális modell megtalálása nem triviális. Tran és munkatársai kísérleti úton vetettek össze többféle 3D hálóvariánst, és a legjobb eredményeket a térbeli és időbeli konvolúciós lépések szétválasztásával kapták (Tran és mtsai, 2018). Az általuk „(2+1)D" CNN-nek nevezett modell először 2D konvolúciót végez a két térbeli tengely mentén, ezt követi az 1D konvolúció az időtengely mentén (lásd 1. ábra). Az 1D konvolúció lépésközparaméterének (stride) állításával átugorhatunk képkockákat, lényegében alulmintavételezve őket. Ennek értelme, hogy ily módon – persze rosszabb felbontással, de – szélesebb időtartományból nyerhetünk ki információt anélkül, hogy a hálózat méretét növelnünk kellene. Látni fogjuk, hogy ennek a tényezőnek kulcsszerepe lesz a jó eredmény elérésében, valamint érdekességképp megjegyezzük, hogy lényegében ugyanez az alapötlet a beszédfelismerésben is rendkívül hatékonynak bizonyult (Tóth, 2014). Természetesen ilyen (2+1)D blokkokból többet is egymásra pakolhatunk, ily módon mély hálózatot építve (Tran és mtsai, 2018). Videók felismerésére Luo és munkatársai is hasonló struktúrát találtak optimálisnak (Luo és Yuille, 2019), és az időbeli információ hasonló, hierarchikus összevonása történik a beszédfelismerésben népszerű TDNN hálóban is (Peddinti és mtsai, 2015).

3. Neurális vokóderek

Bár léteznek már olyan neurális struktúrák, amelyek kimenetként beszédjelet állítanak elő (például épp az itt bemutatandó neurális vokóderek), ezek többnyire nagyméretű hálózatok, amelyek betanításához nagyságrendileg nagyobb mennyiségű hanganyag szükséges, mint ami nekünk rendelkezésünkre állt. A kevés adat miatt ezért célravezetőbbnek ígérkezett közvetlenül a beszédjel helyett valamilyen nagyon tömör reprezentációt használni a tanulás kimeneteként. Természetesen ennek a reprezentációnak olyannak kell lennie, hogy abból a beszédjel aztán jó minőségben visszaállítható legyen. E célra kézenfekvően kínálta magát a beszédkódolásban használatos vokóderek által kinyert spektrális reprezentáció: ez egyrészt tömör (hiszen a cél a beszédtömörítés), másrészt visszaállítható belőle a beszéd (hiszen a betömörített beszédet ki is kell tudni tömöríteni).

A nyelvfeldolgozás számos más területe mellett a neuronhálók a beszédkódolásban és a beszédszintézisben is megjelentek. A neuronhálós beszédszintézisben az alábbi kétlépéses eljárás terjedt el: első lépésben a szöveget valamilyen spektrális reprezentációvá, például spektrogrammá vagy mel-spektrogrammá alakítják, a második lépésben pedig a becsült spektrogramból megkapják a beszédjelet (Prenger és mtsai, 2019). Esetünkben az első komponens szerepét a 3D konvolúciós háló fogja átvenni, hiszen az inputunk nem szöveg, hanem egy videó. A második komponenst viszont – tkp. a vokódert – minden változtatás nélkül használni tudjuk. Az utóbbi időben számos, céljainkra használható neurális vokóder született. Mi ezek közül a WaveGlow-t választottuk (Prenger és mtsai, 2019), mivel a korábbiaknál egyszerűbben használhatónak és jobb minőségűnek tűnt,



2. ábra: Az ultrahang-felvételek megjelenítése a) a nyers adattömbnek megfelelő négyzetes elrendezésben b) interpolációval előállítható, anatómiailiag korrekt elrendezésben.

illetve előre betanított modell is rendelkezésre állt hozzá. Habár ez a betanított modell angol nyelvű, korábbi lehallgatásos kísérletek azt mutatták, hogy magyar nyelvű beszéd szintetizálására is remekül használható – olyannyira, hogy a modell magyar nyelvű mintákon való újratanítása sem eredményezett lényegesen jobb minőségű magyar szintetizált beszédet. (Csapó és mtsai, 2020).

3.1. Adatrögzítés és -feldolgozás

Az ultrahang-felvételek egy magyar anyanyelvű női adatközlő közreműködésével készültek. A mondatok felolvasása során nyelvének mozgását az álla alatt rögzített ultrahang-fejjel vettük fel, az Articulate Instruments Ltd. „Micro" fantázianevű eszközrendszerét használva. Ez a berendezés másodpercenként 82 képet készít, mellyel párhuzamosan a beszédjelet is rögzítettük egy Audio-Technica ATR 3350 típusú kondenzátormikrofonnal, melyet a beszélő előtt 20 cm-re helyeztünk el. Az ultrahang-videó és a beszédjel szinkronizálására a berendezéshez tartozó szoftvert használtuk. Összesen 438 mondatot (körülbelül fél órányi hanganyagot) vettünk fel, melyet véletlenszerűen osztottunk tanító, validációs és teszthalmazokra 310-41-87 arányban. Ugyanezt az adatbázist már korábbi tanulmányok is használták (Csapó és mtsai, 2017; Grósz és mtsai, 2018).

A berendezés a 64 pásztázó nyaláb mindegyike mentén 946 mintát rögzít, amelyből a megfelelő szoftverekkel végzett interpoláció útján az anatómiai viszonyoknak megfelelő ábrát kaphatunk (lásd 2. ábra jobb oldala). Azonban ez a fajta ábra a szokatlan alakja miatt nehezebben feldolgozható, miközben nem tartalmaz extra információt az eredeti 946x64 méretű adattömbhöz képest (2. ábra, bal oldal). Emiatt közvetlenül a nyers adatokkal dolgoztunk, sőt, az adattömböt újramintavételezéssel 128x64 méretűre kicsinyítettük. A képek intenzitásértékeit a [-1, 1] intervallumra normalizáltuk.

A beszédjelet 22050 Hz mintavételezéssel rögzítettük, majd egy, a beszédszintézisben is használt függvény-implementációval mel-frekvenciaskálás spektrogrammá konvertáltuk. A feldolgozás lépésközét 270 mintára állítottuk, ez felelt meg legjobban az ultrahang 82 kép/sec felbontásának. Mivel a WaveGlow betanításakor 256-os lépésközt használtak (ez 86,1 kép/sec-et jelent), az eltérésből eredő enyhe elcsúszást a szintézis előtt a mel-spektrogram interpolációjával korrigáltuk. A WaveGlow inputjaként szolgáló mel-spektrogram a 0-8000 Hz-es frekvenciatartományt 80 sávra osztja, így neuronhálóink tanítása során ez a 80 komponensből álló spektrális vektor szolgált tanítási célértékként.

4. Kísérleti konfigurációk

Neuronhálóinkat a Keras keretrendszer segítségével, Tensorflow alapon implementáltuk (Chollet és mtsai, 2015). Öt különböző modellt készítettünk: egy egyetlen képkockát feldolgozó teljes kapcsolású hálót (FCN), egy 2D konvolúciós hálót, amely továbbra is egyetlen képkockán dolgozik (2D CNN), valamint egy 3D konvolúcós hálót, amely már képkockák sorozatát kapja bemenetként (3D CNN); összehasonlítási alapként készítettünk egy hálót, amely a képkockákat feldolgozó 2D-CNN rétegek eredményét egy LSTM segítségével összegzi (CNN+LSTM), valamint ebből a hálóból készítettünk egy kétirányú változatot is (CNN+BiLSTM). Hogy paraméterszám tekintetében összemérhetőek legyenek, mindegyik hálót úgy lőttük be, hogy kb. 4,3 millió tanulható paraméterük legyen. A tanításra Adam optimalizálót használtunk 100-as batchmérettel. Tanítási hibafüggvényként az átlagos négyzetes hibát (mean squared error, MSE) alkalmaztuk.

Teljes kapcsolású háló (FCN): A lehető legegyszerűbb hálóstruktúra, ha teljes kapcsolású (Keras nyelven „Dense") rétegeket használunk. Esetünkben a hálót öt, rétegenként 430 neuront tartalmazó rejtett rétegből építettük fel, míg a kimeneti réteg 80 darab, lineáris aktivációjú neuronból állt, a mel-spektrális célvektornak megfelelően. A háló inputja egyetlen képkocka, azaz 128x64=8192 pixel. A rejtett rétegek a swish aktivációs függvényt használták (Ramachandran és mtsai, 2017), és minden rejtett réteg után egy dropout réteg következett 0,2-es kiejtési valószínűséggel.

Konvolúciós neuronháló (2D CNN): Az előző hálóhoz hasonlóan ez a háló is egyetlen képkockát dolgoz fel, azonban a legfelső, Dense réteg alatti négy réteg mindegyike térbeli konvolúciót végez az adatokon. A rétegek részletes konfigurációja az 1. táblázatból olvasható le. A legjobbnak tűnő meta-paramétereket kísérletezgetéssel kerestük meg, a rejtett rétegek ebben az esetben is a swish aktivációs függvényt alkalmazták.

3D Konvolúciós neuronháló (3D CNN): Ebben a hálózatban a 2D konvolúció helyett 3D konvolúciót alkalmaztunk, mivel ez teszi lehetővé egyetlen képkocka helyett képek rövid sorozatának feldolgozását. A konkrét hálóstruktúra (lásd 1. táblázat) öt képkockát dolgoz fel, amelyek s távolságra találhatók egymástól, ahol s az időtengely mentén végzett konvolúció lépésköze („stride" paramétere). A 2. fejezetben bemutatott (2+1)D konvolúció koncepciójának meg-

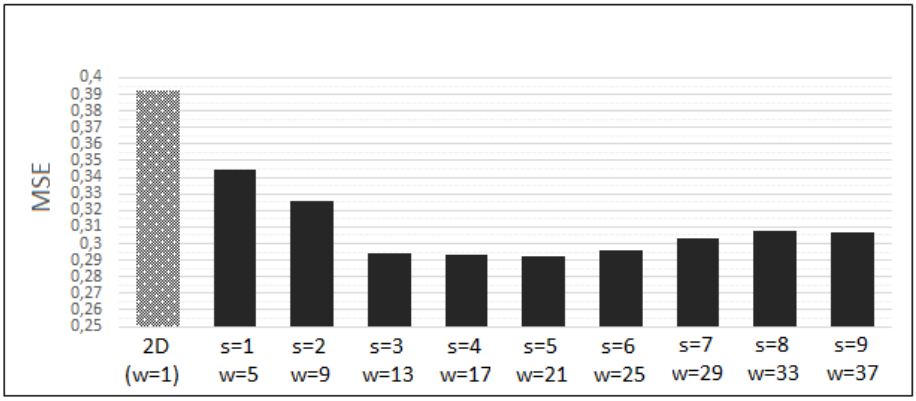
1. táblázat. A 2D és a 3D CNN hálók rétegei a Keras-implementációban, valamint a legfontosabb paramétereik. A különbségeket vastag betűvel kiemeltük.

2D CNN	3D CNN
Conv2D(30, (13,13), strides=(2,2))	Conv**3D**(30, (**5**,13,13), strides=(**s**,2,2))
Dropout(0.2)	Dropout(0.2)
Conv2D(60, (13,13), strides=(2,2))	Conv**3D**(60, (**1**,13,13), strides=(**1**,2,2))
Dropout(0.2)	Dropout(0.2)
MaxPooling2D(pool_size=(2,2))	MaxPooling**3D**(pool_size=(**1**,2,2))
Conv2D(90, (13,13), strides=(2,1))	Conv**3D**(90, (**1**,13,13), strides=(**1**,2,1))
Dropout(0.2)	Dropout(0.2)
Conv2D(150, (13,13), strides=(2,2))	Conv**3D**(**85**, (**1**,13,13), strides=(**1**,2,2))
Dropout(0.2)	Dropout(0.2)
MaxPooling2D(pool_size=(2,2))	MaxPooling**3D**(pool_size=(**1**,2,2))
Flatten()	Flatten()
Dense(1000)	Dense(1000)
Dropout(0.2)	Dropout(0.2)
Dense(13, activation='linear')	Dense(13, activation='linear')

felelően az öt képkockát először külön-külön dolgozza fel a háló, majd ezután végzi el az időtengely menti konvolúciós lépést. Az 1. táblázatban vastag betűvel emeltük ki, hogy ehhez milyen módosításokat kellett eszközölni az eredeti, 2D CNN hálóhoz képest. Megjegyezzük, hogy a legfelső konvolúciós réteg méretének csökkentése azért volt szükséges, hogy a két háló paraméterszáma megközelítőleg ugyanannyi maradjon.

LSTM neuronháló (CNN+LSTM): A teljes kapcsolású, illetve a 2D CNN hálók egyetlen képkockát dolgoznak fel, így sejthető, hogy nem lesznek méltó versenytársai a képek sorozatán működő 3D CNN hálónak. Sorozatok feldolgozására az LSTM hálózat ajánlott, illetve mivel képek sorozatáról van szó, érdemes a visszacsatolt hálót a képeket feldolgozó 2D CNN hálóval kombinálni. Az általunk használt CNN+LSTM modell alsó négy rétege megegyezett a 2D CNN háló alsó négy rétegével; legfelső rejtett rétegként viszont a Dense réteget LSTM rétegre cseréltük. A paraméterszám megőrzésre miatt 500 LSTM neuront tettünk ebbe rétegbe, a hálózat inputját pedig 21 egymást követő képkocka képezte.

Kétirányú LSTM neuronháló (CNN+BiLSTM): Ha nem ragaszkodunk a teljesen valós idejű feldolgozáshoz, akkor az LSTM réteg nem csak időben előre (balról jobbra), hanem időben visszafelé (jobbról balra) haladva is működhet. Szokásos megoldás továbbá egy előre és egy visszafelé haladó réteget is képezni, és ezek kimenetét kombinálni. A CNN+BiLSTM nevű modellünk egy ilyen, úgynevezett kétirányú (bidirectional) hálót takar, amely az előző, CNN+LSTM modelltől csakis az LSTM réteg kétirányúsításában tér el. A paraméterszám megőrzése miatt ebben a modellben az LSTM réteg méretét 320-ra kellett redukálnunk.



3. ábra: 3D CNN háló MSE hibaértéke a validációs halmazon az s paraméter függvényében. Összehasonlításképp a 2D CNN háló hibáját is feltüntettük (bal szélső oszlop).

5. Kísérleti eredmények

A neuronhálóink teljesítményének kiértékelésére többféle lehetőség kínálkozik. A legegyszerűbb megoldás a tanítás során használt célfüggvény (jelen esetünkben az MSE) értékeit összehasonlítani a validációs vagy a tesztadatokon. Az MSE értékek mellett a korreláción alapuló R^2 értékeket is fel fogjuk tüntetni, mivel regressziós feladatok esetén az is egy egyszerűen kiszámolható és népszerű mérőszám. Azonban mivel a kimenetünk egy beszédjel, ezek a matematikai alapon megfogalmazott egyszerű mérőszámok nem feltétlenül tükrözik a hang érzékelt, szubjektív minőségét. Erre vonatkozóan csak lehallgatásos tesztekkel – mint például a MUSHRA teszt (ITU, 2001) – kaphatnánk becslést. Az ilyen, emberi alanyokkal történő szubjektív kiértékelés azonban elég fáradságos, ezért dolgozták ki a különféle objektív metrikákat, amelyek ugyan matematikai úton, de az emberi hallás fő tulajdonságait figyelembe véve próbálják megbecsülni a hang minőségét. Kísérleteink első részében csak a két egyszerű, objektív mérőszámot – a MSE ill. R^2 értékeket – közöljük, a végső összehasonlításnál azonban néhány további, hangminőséget becslő objektív mérőszámot is mutatunk majd.

Mint az elméleti ismertetésben láttuk, a 3D konvolúciós hálónak van egy nagyon fontos meta-paramétere, az s paraméter. A legelső kísérletben ennek hatását vizsgáltuk a hibafüggvény értékére. Ez a paraméter határozza meg, hogy a háló az input mekkora időszakaszáról kap információt: a két szélső képkocka közötti távolság a $w = 4 \cdot s + 1$ képlettel határozható meg. Például $s = 5$ érték esetén a háló által lefedett időablak mérete $w = 21$ képkocka. A videó 82 kép/sec mintavételezési rátáját figyelembe véve, ez körülbelül negyed másodpercnek, nagyságrendileg egy szótag hosszának felel meg.

2. táblázat. A különféle hálóarchitektúrákkal kapott MSE és R^2 értékek a validációs és a teszthalmazon.

Hálózat típusa	Val		Teszt	
	MSE	Mean R^2	MSE	Mean R^2
FCN	0,410	0,600	0,419	0,585
2D CNN	0,392	0,617	0,401	0,603
3D CNN (s=5)	0,292	0,714	0,293	0,710
CNN + LSTM	0,303	0,701	0,296	0,709
CNN + BiLSTM	0,301	0,706	0.296	0,707

A 3. ábra mutatja a 3D CNN hálóval kapott MSE értékeket az s paraméter különböző értéke esetén. Összehasonlításképp a 2D CNN háló (amely csak egyetlen képkockát dolgoz fel) hibáját is feltüntettük. Látható, hogy az aktuális képkocka mellett annak környezetét is figyelembe véve jelentős hibacsökkentést érhetünk el. Már 2-2 közvetlen szomszédot használva (s=1) is jobb eredményt kapunk, de lényegesen nagyobb a javulás 3-6 közti s értékekkel. A tágabb kontextus figyelembe vétele tehát fontos, még képkockák átugrása árán is. A fenti eredmény alapján s értékét 5-re rögzítettük.

A következő kísérletben az ötféle hálóstruktúrát vetettük össze, a validációs és teszthalmazokon kapott MSE és R^2 értékeket a 2. táblázat összegzi (R^2 esetén a nagyobb érték jelent jobb modellt). Látható, hogy az egyetlen képkockát feldolgozó FC és 2D CNN hálók közül a konvolúciós háló ugyan egyértelműen jobb, de sokkal jobb eredményt érhetünk el az egyetlen kép helyett képsorokat feldolgozó hálóvariánsokkal (3D CNN ill. CNN+LSTM hálók).

A 3D CNN és az LSTM-alapú hálók összevetéséhez az LSTM hálók inputját 21 képkockára állítottuk, hiszen a 3D CNN esetén ez bizonyult optimálisnak. Mint a táblázatból látható, a CNN+LSTM modell egyértelműen megverte ugyan az egyetlen képkockás modelleket, de a 3D CNN háló pontosságát nem tudta Meghaladni. Nem változtatott ezen az LSTM réteg kétirányúsítása (BiLSTM) sem: míg ez más feladatokon általában egy pici javulást szokott hozni, itt most gyakorlatilag az egyirányú hálóval ekvivalens eredményt kaptunk. Felvetődött, hogy esetleg a CNN+LSTM modellek számára más lehet az optimális ablakméret, ezért próbáltunk változtatni a 21-es inputméreten, de más értékek esetén sem nem kaptunk lényegesen jobb eredményt. A kapott hibaértékek alapján úgy tűnik, hogy a képkockák alulmintevételezése ugyanolyan hatékonyan segíti az információ fúzióját, mint az összes képkocka feldolgozása az LSTM szofisztikáltabb, visszacsatolást és belső memóriát is alkalmazó technikájával. Az LSTM viszont, épp a rekurrens jellege miatt, nem tud átugrani képkockákat, pedig lehet, hogy ebben az esetben pont erre lenne szükség. Szintén az összes képkocka megőrzéséből kifolyólag az CNN+LSTM háló tanítása jóval hosszabb – kb. 70%-kal több – időt vett igénybe, mint a 3D CNN háló betanítása. A modellek azonos paraméterszáma ellenére érdekes módon az LSTM háló memóriaigénye is nagyobb volt, ennek feltehetően szintén az összes input-képkocka megőrzése az oka.

3. táblázat. Öt modellünk összevetése beszédminőséget mérő objektív mérőszámokkal.

	STOI	PESQ	MCD
FCN	0,661	1,562	4,602
2D CNN	0,658	1,551	4,569
3D CNN (s=5)	0,743	1,831	4,161
CNN + LSTM	0,742	1,792	4,139
CNN + BiLSTM	0,736	1,789	4,152

Végezetül megjegyezzük, hogy korábban Moliner és Csapó is próbálkozott a 2D CNN és LSTM hálók kombinálásával hasonló feladaton (Moliner és Csapó, 2019). Eredményük azonban direkt módon nem összevethető a miénkkel, ugyanis másik vokódert, és ennek megfelelően a tanítás során más célértékeket használtak. Emellett az általuk használt háló jóval nagyobb is volt, több mint négyszeres paraméterszámmal a mi hálóinkhoz képest. Az egyirányú és a kétirányú LSTM-változatok teljesítménye között ők sem tapasztaltak szignifikáns különbséget. Velünk párhuzamosan Saha és munkatársai is próbálkoztak nyelvultrahang-videók feldolgozásával, és tőlünk függetlenül a miénkhez hasonló 3D konvolúciós hálóstruktúrát hoztak ki optimálisnak, valamint ők lényegesen jobb eredményeket kaptak a 3D CNN hálóval, mint a CNN+LSTM kombinációval (Saha és mtsai, 2020).

A hang minőségének kiértékelésére sokféle objektív mérőszámot javasoltak. Ezek valamilyen szinten igyekeznek figyelembe venni az emberi hallás működésének fő tulajdonságait, így valamivel pontosabb becslést adnak a hangminőségre, mint a tanítás során optimalizált MSE hibafüggvény. Az alábbi, 3. táblázatban három ilyen mérőszámot értékeltünk ki az ötféle modellel szintetizált teszthalmazon, ezek sorban a STOI (Short-Term Objective Intelligibility, Taal és mtsai (2011)), a PESQ (Perceptual Evaluation of Speech Quality, ITU-R (2001)), valamint az MCD (Mel-Cepstral Distance, Kubichek (1993)). Előbbi kettő esetén a magasabb érték jelez jobb minőséget, utóbbi esetén a kisebb. Az így kapott a számok is at mutatják, hogy egyértelmű minőségi ugrás van az egyetlen képkockát feldolgozó (FC és 2D CNN), valamint a képsorozatokat konvertáló (3D CNN, LSTM és BiLSTM) hálók között. Két mérőszám a 3D hálót, egy pedig az LSTM hálót hozta ki győztesnek, de az eltérés e két háló teljesítménye között mindhárom metrika szerint minimális, így egyértelmű nyertest nem mernénk hirdetni.

6. Összefoglaló

Cikkünkben egy háromdimenziós konvolúciót végző neuronháló hatékonyságát vizsgáltuk a beszédjel artikulációs felvételekből való visszaállításának feladatkörében. Tran és munkatársainak tanulmánya által motiválva, hálózatunk a 3D konvolúciót két lépésre bontja, így előbbi a térbeli, majd az időbeli konvolúciós lépést végzi el (Tran és mtsai, 2018). Modellünket egy másik, képsorozatok

modellezésére javasolt hálótípussal, a CNN+LSTM háló két változatával vetettük össze. Kísérleteinkben a 3D háló ekvivalens, vagy kicsivel jobb teljesítményt nyújtott, miközben betanítása kevesebb időt igényelt. Mivel mindkét modellnek sok meta-paramétere van, a 2D CNN fölényének kijelentéséhez további alapos mérések kellenének, de annyit bizonyosan állíthatunk, hogy a 3D CNN háló mindenképpen versenyképes alternatívát jelent a CNN+LSTM kombinált hálókkal szemben, ha a célunk nyelvultrahang-videókon alapuló némabeszéd-interfész építése. A jövőben további összehasonlításokat tervezünk egy újabb hálótípussal, az úgynevezett ConvLSTM hálókkal. Mint nevük is mutatja, ezek a hálók egyetlen rétegen belül egyesítik a konvolúciós és az LSTM modellezés előnyeit, tehát elvileg még alkalmasabbak lehetnek videók feldolgozására (Zhao és mtsai, 2019).

Köszönetnyilvánítás

A kutatást a Nemzeti Kutatási Fejlesztési és Innovációs Hivatal FK 124584 kódjelű pályázata, valamint az Innovációs és Technológiai Minisztérium TUDFO/47138-1/2019-ITM programja támogatta. Gosztolya Gábor kutatásait az MTA Bolyai János kutatói ösztöndíja és az Új Nemzeti Kiválóság Program Bolyai+ pályázata (azonosító: ÚNKP-20-5-SZTE-649) támogatta. A kutatáshoz használt grafikus kártya az NVIDIA Corporation ajándéka.

Hivatkozások

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

Bradbury, J., Merity, S., Xiong, C., Socher, R.: Quasi-recurrent neural networks. In: Proc. ICLR (2017)

Cho, K., van Merrienboer, B., Gülcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Proc. EMNLP. pp. 1724–1734 (2014)

Chollet, F., és mtsai: Keras. https://github.com/fchollet/keras (2015)

Csapó, T.G., Grósz, T., Gosztolya, G., Tóth, L., Markó, A.: DNN-based ultrasound-to-speech conversion for a silent speech interface. In: Proc. Interspeech. pp. 3672–3676 (2017)

Csapó, T.G., Zainkó, C., Tóth, L., Gosztolya, G., Markó, A.: Ultrasound-Based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis. In: Proc. Interspeech 2020. pp. 2727–2731 (2020), http://dx.doi.org/10.21437/Interspeech.2020-1031

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Communication 52(4), 270–287 (2010)

Denby, B., Cai, J., Hueber, T., Roussel, P., Dreyfus, G., Crevier-Buchman, L., Pillot-Loiseau, C., Chollet, G., Manitsaris, S., Stone, M.: Towards a practical silent speech interface based on vocal tract imaging. In: Proc. ISSP. pp. 89–94 (2011)

Dong, L., Xu, S., Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: Proc. ICASSP. pp. 5884–5888 (2018)

Ephrat, A., Peleg, S.: Vid2speech: Speech reconstruction from silent video. In: Proc. ICASSP. pp. 5095–5099 (2017)

Gonzalez, J.A., Cheah, L.A., Gomez, A.M., Green, P.D., Gilbert, J.M., Ell, S.R., Moore, R.K., Holdsworth, E.: Direct speech reconstruction from articulatory sensor data by machine learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(12), 2362–2374 (2017)

Grósz, T., Gosztolya, G., Tóth, L., Csapó, T.G., Markó, A.: F0 estimation for DNN-based ultrasound silent speech interfaces. In: Proc. ICASSP. pp. 291–295 (2018)

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)

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), 288–300 (2010)

ITU: ITU-R recommendation BS.1534: Method for the subjective assessment of intermediate audio quality (2001)

ITU-R: ITU-R recommendation P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (2001)

Janke, M., Diener, L.: EMG-to-speech: Direct generation of speech from facial electromyographic signals. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(12), 2375–2385 (2017)

Janke, M., Wand, M., Nakamura, K., Schultz, T.: Further investigations on EMG-to-speech conversion. In: Proc. ICASSP. pp. 365–368 (2012)

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. pp. 1467–1471 (2016)

Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35(1), 221–231 (2013)

Kim, M., Cao, B., Mau, T., Wang, J.: Multiview representation learning via deep CCA for silent speech recognition. In: Proc. Interspeech. pp. 2769–2773 (2017a)

Kim, M., Cao, B., Mau, T., Wang, J.: Speaker-Independent Silent Speech Recognition From Flesh-Point Articulatory Movements Using an LSTM Neural Network. IEEE/ACM Trans. ASLP 25(12), 2323–2336 (2017b)

Kimura, N., Kono, M., Rekimoto, J.: Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In: Proc. of CHI Conf. on Human Factors in Computing Systems (2019)

Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. pp. 1097–1105 (2012)

Kubichek, R.: Mel-cepstral distance measure for objective speech quality assessment. In: Proc. ICASSP. pp. 125–128 (1993)

Lakew, S., Cettolo, M., Federico, M.: A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In: Proc. COLING. pp. 641–652 (2018)

Liu, Z.C., Ling, Z.H., Dai, L.R.: Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation. Speech Communication 99(2017), 161–172 (2018)

Luo, C., Yuille, A.: Grouped spatial-temporal aggregation for efficient action recognition. In: Proc. International Conference on Computer Vision. pp. 5512–5521 (2019)

Maier-Hein, L., Metze, F., Schultz, T., Waibel, A.: Session independent non-audible speech recognition using surface electromyography. In: Proc. ASRU. pp. 331–336 (2005)

Moliner, E., Csapó, T.: Ultrasound-based silent speech interface using convolutional and recurrent neural networks. Acta Acustica united with Acustica 105 (2019)

Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proc. Interspeech. pp. 3214–3218 (2015)

Prenger, R., Valle, R., Catanzaro, B.: Waveglow: A flowbased generative network for speech synthesis. In: Proc. ICASSP. pp. 3617–3621 (2019)

Ramachandran, P., Zoph, B., Le, Q.V.: Swish: a Self-Gated Activation Function. ArXiv e-prints 1710.05941 (2017)

Saha, P., Liu, Y., Gick, B., Fels, S.: Ultra2speech – a deep learning framework for formant frequency estimation and tracking from ultrasound tongue images. In: Proc. MICCAI. pp. 473–482 (2020)

Schultz, T., Wand, M., Hueber, T., Krusienski, D.J., Herff, C., Brumberg, J.S.: Biosignal-based spoken communication: A survey. IEEE/ACM Trans. ASLP 25(12), 2257–2271 (2017)

Taal, C., Hendriks, R., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. ASLP 19(7), 2125–2136 (2011)

Tóth, L.: Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition. In: Proc. ICASSP. pp. 190–194 (2014)

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proc. CVPR (2018)

Wang, J., Samal, A., Green, J.: Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph. In: Proc. SLPAT. pp. 38–45 (2014)

Wu, C., Chen, S., Sheng, G., Roussel, P., Denby, B.: Predicting tongue motion in unlabeled ultrasound video using 3D convolutional neural networks. In: Proc. ICASSP. pp. 5764–5768 (2018)

Zhang, S., Lei, M., Yan, Z., Dai, L.: Deep-FSMN for large vocabulary continuous speech recognition. In: Proc. ICASSP (2018)

Zhao, C., Zhang, J., Wu, C., Wang, H., Xu, K.: Predicting tongue motion in unlabeled ultrasound video using convolutional LSTM neural networks. In: Proc. ICASSP. pp. 5926–5930 (2019)

Zhao, S., Liu, Y., Han, Y., Hong, R., HU, Q., Tian, Q.: Pooling the convolutional layers in deep convnets for video action recognition. IEEE Trans. Circuits and Systems for Video Technology 28(8), 1839–1849 (2018a)

Zhao, Y., Xiong, Y., Lin, D.: Trajectory convolution for action recognition. In: Advances in Neural Information Processing Systems 31. pp. 2204–2215 (2018b)