Szájról olvasás automatizálása mély neurális hálózatok és mobilalkalmazás-kezelőfelület alkalmazásával

Arthur Frigyes Viktor[1], Csapó Tamás Gábor[1,2]

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

Abstract

Automatic lipreading is a technique to predict the spoken content using lip video input. The advantage of lip video compared to other articulatory techniques (e.g. ultrasound tongue imaging and MRI) is that it is easily available and affordable: most modern smartphones have a front camera. There are already a few solutions for lip-to-speech synthesis, but they mostly concentrate on offline training and inference. In this research, we propose a system built from three components: a backend for deep neural network training and inference, a webservice responsible for the communication between the server and the client, and a frontend as a form of a mobile application. We trained two approaches, both using convolutional and recurrent neural networks. In the first case, we record the mimic movements of the whole face and from this information, we deduce the phonetic information. In the latter case, only the mouth area is available as input data for the neural network. Our initial evaluation shows that the scenario is feasible: a top-5 classification accuracy of 74% is combined with feedback from the mobile application user, making sure that the speaking impaired might be able to communicate with this solution. The results of the articulatory-to-text conversion can contribute to the development of 'Silent Speech Interface' (SSI) systems. The essence of SSI is recording the articulation organs while the user of the device actually does not make a sound but yet the machine system is capable to synthesize speech based on the movement of the organs.

Keywords: vid2speech, lip-reading, lip video, DNN, speech technology

1. Bevezetés

A beszéd megértése képi információk alapján rendkívül nehéz feladat. Nehezen általánosítható, tekintve, hogy minden beszélő artikulációja kisebb-nagyobb mértékben eltér egymástól. Az artikuláció-akusztikum konverziós módszerek célja, hogy artikulációs mozgás alapján szintetizáljanak beszédet. Az artikulációs információ lehet például a nyelv mozgása ultrahanggal rögzítve (Csapó

Email addresses: hello@victorarthur.com (Arthur Frigyes Viktor), csapot@tmit.bme.hu (Csapó Tamás Gábor)

et al., 2017a,b), elektromágneses artikulográf (Cao et al., 2018), felszíni eletromiográfia (Diener & Schultz, 2018), vagy mágnesesrezonancia-képalkotás (Csapó, 2020).

1.1. Artikuláció-akusztikum átalakítás ajakvideó alapján

A konverzió egy másik lehetséges megoldása a beszédszintézis kizárólag egy arcról vagy ajakról készült videó képkockáiból (Rácz & Csapó, 2020). 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), Akbari et al. (2018) és Rácz & Csapó (2020). Le Cornu & Milner (2015) a felvett ajakmozgás alapján mély neurális hálózatot tanítanak be a beszéd spektrális paramétereinek megbecsülésére, majd ebből egy vokóderrel beszédet szintetizálnak. Ephrat & Peleg (2017) konvolúciós hálózatokat (CNN) alkalmaznak, és a GRID audiovizuális adatbázison megmutatják, hogy a mély neuronháló betanítása során nem látott szavakat is képes a rendszer érthetően szintetizálni, 52%-os pontossággal. Akbari et al. (2018) egy újfajta spektrális reprezentációt (hallási spektrogram) alkalmaznak, és a gépi tanulást autóenkóder, konvolúciós, és rekurrens hálózatok (RNN) kombinációjával valósítják meg. Rácz & Csapó (2020) szintén CNN és RNN hálózatokat használnak a közvetlen 'lip-to-speech' beszédszintézisre.

Wand et al. (2016) és Sun et al. (2018) munkájukban a közvetett módszert mutatják be, azaz automatikus szájról olvasást végeznek. Wand et al. (2016) mutatták be az egyik első teljesen mély neuronháló alapú megoldást, melyben rekurrens hálózatot (LSTM, Long-Short Term Memory) alkalmaznak. Korábbi munkákhoz hasonlóan a GRID adatbázison tesztelték módszerüket, melynek eredménye 79.6%-os szófelismerési arány lett. Sun et al. (2018) a Lip-Interact rendszert mutatták be, melynek segítségével néma ajakmozgással lehet irányítani a mobiltelefon bizonyos funkcióit. Parancsszavas felismerést valósítanak meg (44 parancsot megkülönböztetve) a mobiltelefon első kameráját felhasználva. Az eredmények szerint a Lip-Interact elsősorban akkor hasznos, ha a

felhasználó egyik vagy mindkét keze foglalt (pl. vezetés közben), és ilyenkor a mobiltelefonos interakciók az automatikus szájról olvasás felhasználásával hatékonyabbá válnak.

A némabeszéd-interfész (Silent Speech Interface, SSI) az artikuláció-akusztikum konverziós módszerek egy olyan távlati alkalmazása, amelynek használatával némán beszélve, „tátogva" adhatunk ki hangot (Denby et al., 2010; Csapó et al., 2017b; Kimura et al., 2019; Gonzalez-Lopez et al., 2020). 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.

1.2. A jelen kutatás célja

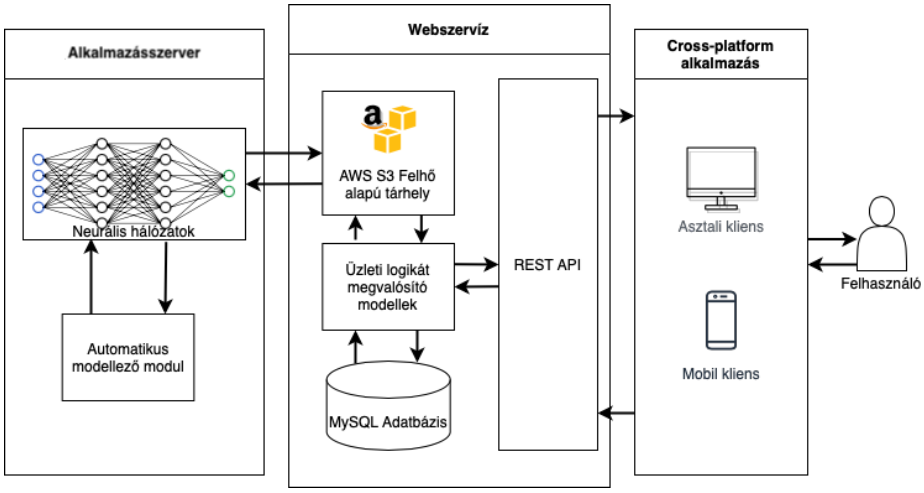
A kutatás célja egy megvalósíthatósági tanulmány: azt teszteltük, hogy egy mobil kliens alapú automatikus szájról olvasó rendszerhez milyen komponensek szükségesek. Választásunk a fenti lehetőségek közül egy 'közvetlen' módszer megvalósítására esett, mert így a felhasználónak lehetősége van az artikuláció-beszéd felismerés után a legvalószínűbb eredmények közül választani, és így mindenképp a helyes szöveg kerül a szövegfelolvasó bemenetére. Távlati célunk az alkalmazással, hogy hangtalan, illetve kis hangintenzitással beszélő pácienseknek számára egy alternatív kommunikációs eszközként használható megoldást nyújtson.

2. Módszerek

A kutatás során megvalósítottuk a teljes architektúrát, a háttérszerverrel és mobil kliens alkalmazással együtt.

2.1. A rendszer főbb komponensei

A prototípus rendszer három fő komponensből és azok alárendelt rendszereiből tevődik össze, melyek az 1. ábrán láthatók.



1. ábra. Rendszer architektúra

Alkalmazásszerver Az első komponens az alkalmazásszerver, mely közvetlenül az AWS S3 tárterületre felkerült új, a felhasználóra specifikus videó anyagot tanító adatként használva elvégzi a mély neuronháló tanítását, azaz frissíti a neurális hálózat súlymátrixát. Az így elkészült predikciós modell alapján később a mobilalkalmazás beszédpredikcióra képes. A legenerált modelleket az alkalmazásszerver felmásolja a tárterületre, ahonnan azt a későbbi felhasználás során elérheti.

Webszerviz Az alkalmazásszerver és a mobilalkalmazás között helyezkedik el egy webszerviz, melynek főbb feladatai a felhasználói autentikáció, adatok rögzítése, hívások kezelése lokális adatbázisban, valamint a tanulási adatok mentése és feltöltése felhő alapú tárterületre.[1]

[1] A kiértékeléshez felhasznált adatok AWS S3 szerveren kerültek tárolásra, míg a webszerviz DigitalOcean szerverein került kiszolgálásra, az SSL tanúsítványt a LetsEncrypt bocsájtotta ki.

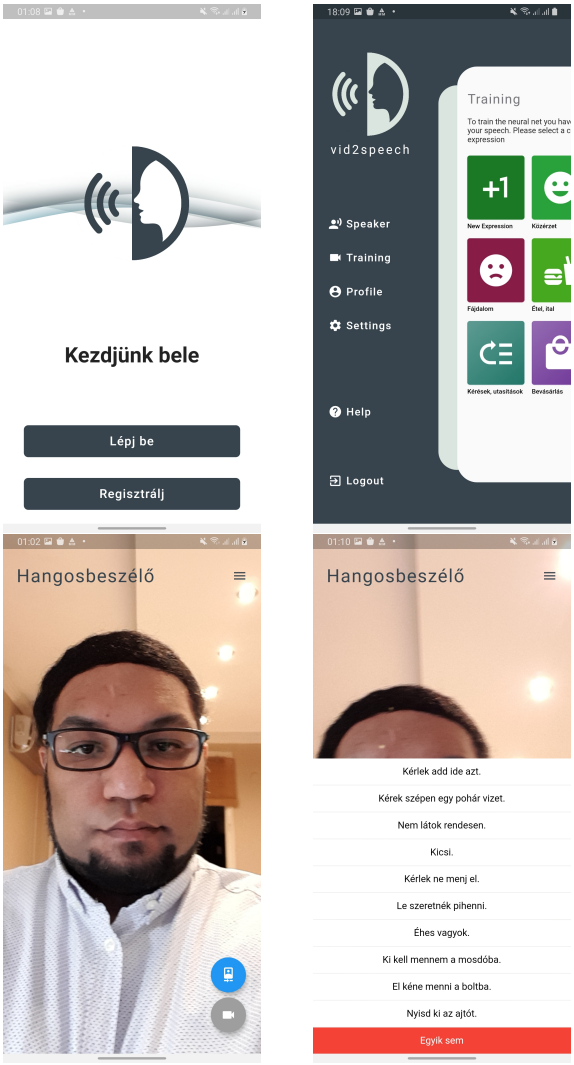
Cross-platform mobilalkalmazás A felhasználó közvetlenül a mobilalkalmazással van kapcsolatban, mely a Flutter keretrendszer által biztosított platformfüggetlenség előnyeit élvezi. Ennek okán iOS és Android mobil operációs rendszereken egyaránt elérhető a szolgáltatás. A betanítási fázisban a felhasználónak előre megadott mondatokat kell némán felolvasnia, melyeket utána a webszervizen keresztül az alkalmazásszerverre töltünk fel; melyből a gépi tanulási módszerrel egy predikciós modell készül.

A mobilalkalmazáson belül a predikciós mód esetén az ajakvideó alapú felismerés után a felhasználó számára megjelenítjük a legvalószínűbb találatokat, melyekből választhat. A kiválasztott szöveget átadjuk a rendszerben lévő szövegfelolvasónak (Androidon lehetséges alternatívák például: gTTS, ProfiVox és ProfiVox-HMM). A jelen kutatás során a szövegfelolvasó modulok előnyeit/hátrányait nem vizsgáltuk. A fejlesztés során Samsung Galaxy Note 9, valamint Huawei P10 mobil eszközön futó Android 9.0 és 10.0 operációs rendszereken történt a kiértékelés és az adatrögzítés. A 2. ábra a mobilalkalmazás angol és magyar változataiból jelenít meg képernyőképeket.

A leírt három részből álló struktúra lehetővé teszi a nagy számítási kapacitást igénylő gépi tanulási folyamat elkülönítését: az ilyen műveletek aszinkron módon történnek meg az alkalmazásszerveren. Az itt lezajló kiértékelés eredménye minimális késleltetéssel a mobil eszközön megjelenik, ezzel a kliens oldalt mentesítve a nagyobb erőforrás igényű feladatok elvégzésétől. Továbbá megvalósítja a "Single-point-of-change" elvét, tehát elegendő egy helyen változtatunk a rendszert és annak hatását vizsgálhatjuk a kimeneten, a többi komponens megváltoztatása nélkül. Ilyen módon a rendszer alkalmas új neurális hálózatok hatásának tesztelésére.

2.2. Tanító adatok rögzítése a mobil kliens segítségével

Mind a tanító adatokat, mind pedig a predikció során használt videóanyagot a mobilalkalmazás rögzíti.



2. ábra. Képernyőképek a mobilalkalmazásról (angol és magyar nyelvű verziók). Bal felső: Bejelentkezési képernyő. Jobb felső: Navigációs és tanítási képernyő. Bal alsó: Néma ajakmozgás rögzítése. Jobb alsó: A szájról szövegre konvertálás eredményének megjelenítése (több eredmény egy listában).

A tanítási adatok szöveges tartalmához 88 kifejezést választottunk ki. A 88 példamondat a StrokeAid[2] elnevezésű segédprogramból került átemelésre, aminek célja, hogy segítséget nyújtson a stroke-on átesett páciensek számára. A StrokeAid programban található kifejezések segítségével a páciens gyorsan tud reagálni egy beszélgetésben.

Tanítási módban a rögzítés során a felhasználókat arra kérjük, hogy ötször ismételjék el ezeket a mondatokat, így minden példamondathoz öt felvétel keletkezik (azaz összesen 440 videó). Az arcról készült videót 720x1280 képpontos felbontással és 25 képkocka/másodperces sebességgel rögzíti a mobilalkalmazás az okostelefon elülső kamerájával, és minden mondat rögzítése után az adatokat a webszervizen keresztül elküldjük az alkalmazásszerverre. A megvalósíthatósági tanulmányunk során egy férfi beszélővel (a cikk első szerzője) teszteltük a tanítást, aki felmondta a 440 kifejezést.

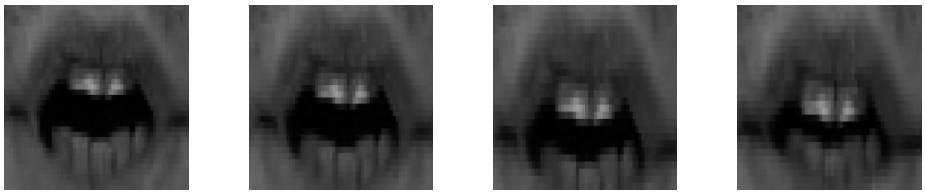
2.3. Videóadatok feldolgozása

Miután megtörtént a tanító adatok gyűjtése a mobilalkalmazással, a feldolgozás hátralévő részéért az alkalmazásszerver felelős. Az arc egyes jellemző pontjainak meghatározásához három eljárást teszteltünk: 2d106det MobileNet (Deng et al., 2019), Google Firebase ML Kit FireVision[3], és 'shape_predictor_68_face_landmarks' modell, mely a DLib programcsomag része[4]. Miután a sebesség és kompatibilitási teszteket elvégeztük, a választásunk a DLib módszerre esett, mely 68 jellemző pontot jelöl ki az arcon. Ennek megfelelően a neurális hálózat bemeneteként a videókból csak a száj környéki régiót mutató, 299x299 pixelre átméretezett részeket használtuk fel. Néhány minta képkocka látható a 3. ábrán.

[2] `https://play.google.com/store/apps/details?id=com.onlab.monddki`

[3] `https://firebase.google.com/docs/ml-kit/detect-faces`

[4] `https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/`



3. ábra. Néhány a videókból kivágott ajak-régió minta, amelyek a DNN bemenetére kerültek.

2.4. DNN tanítása az alkalmazásszerveren

A rögzített 440 videó 60%-át tanítási és 40%-át validációs halmazra osztottuk. A jellemzők kinyerésére konvolúciós hálózatot, a végső osztályzáshoz pedig rekurrens hálózatot használtunk. Automatikus hiperparaméter optimalizálást végeztünk, melynek során a mély neuronháló különböző paramétereit több tartományban állítottuk, majd a betanított rendszereket kiértékeltük. A tanítás során a korai leállítást 10 epoch türelemmel alkalmaztuk. A hálózatot osztályozási módban tanítottuk, kategorikus kereszt-entrópia költségfüggvény és ADAM optimalizáló használatával (tanulási ráta: 10^{-5}).

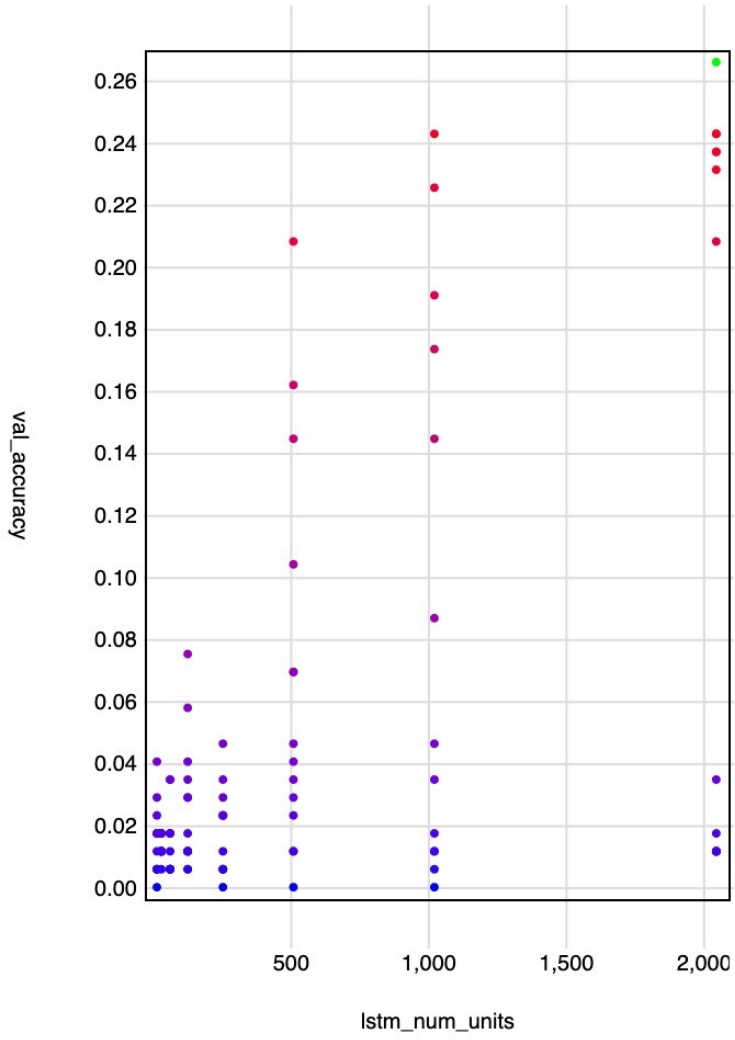
3. Eredmények és diszkusszió

3.1. Mély neuronhálós tanítás

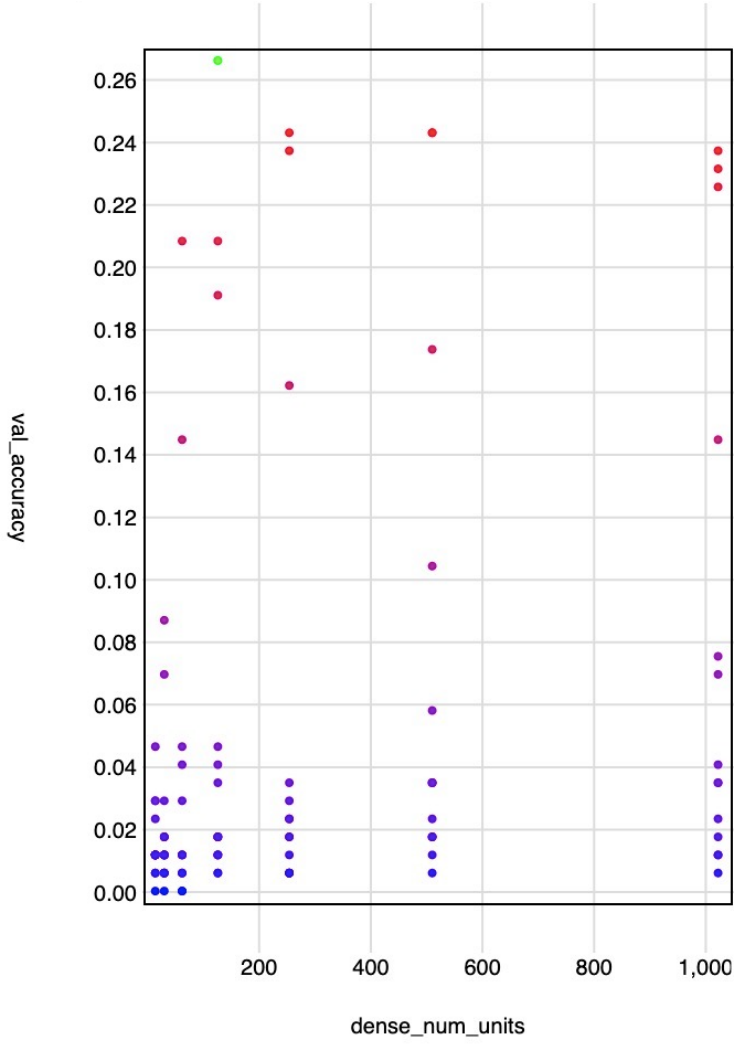
A mély neuronhálók hiperparaméter optimalizálása során kapott eredményeket a 4–6. ábrák mutatják. A 4. ábrán az LSTM rétegben (a hálózat rekurrens része, mely a szekvenciális adatok feldolgozásáért felel) lévő neuronok száma, az 5. ábrán az előrecsatolt rétegben lévő neuronok száma, míg a 6. ábrán a hálózatban használt DropOut hatása látható. Az optimális hálózati struktúra a következő: InceptionV3 a jellemző kinyeréséhez, amelyet egyetlen LSTM réteg követ, 2048 neuronnal, majd 10% DropOut és egy teljesen kapcsolt réteg, amelynek végén 128 neuron található.

3.2. Tesztelés új felvételekkel

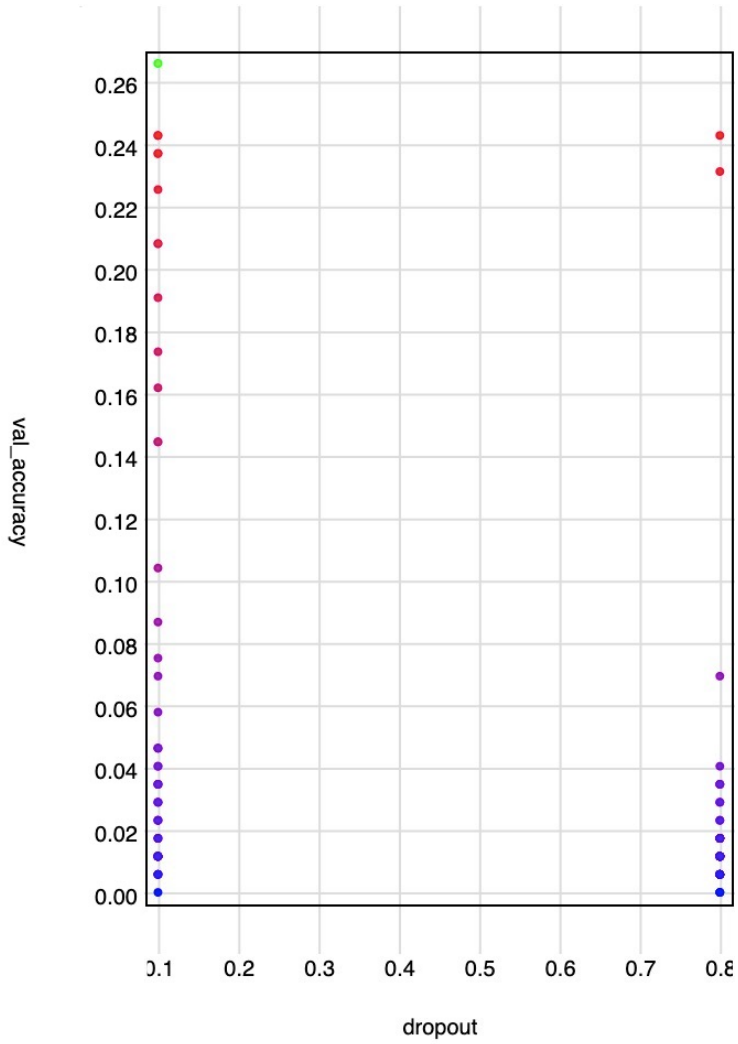
Tesztelés céljából új videofelvételeket rögzítettünk (a 88 magyar mondat mindegyikéhez egy-egy bemondás), biztosítva, hogy tanító és a kiértékelés során használt adatok ne kerüljenek ismételt felhasználásra, azaz függetlenek legyenek.



4. ábra. Top-1 validációs pontosság az LSTM rétegben lévő neuronok számának függvényében.



5. ábra. Top-1 validációs pontosság az előrecsatolt rétegben lévő neuronok számának függvényében.



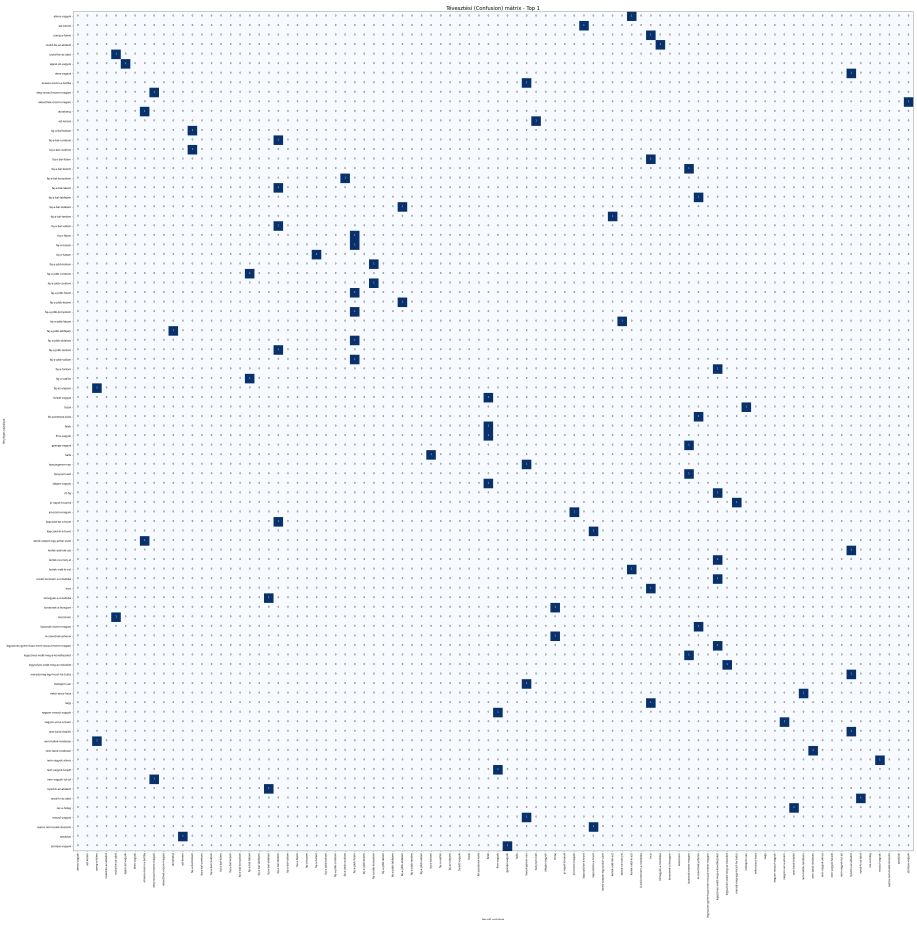
6. ábra. Top-1 validációs pontosság hálózatban használt DropOut függvényében.

Az osztályozás eredményeként a Top-1 tévesztési mátrix a 7. ábrán látható, míg a 8. ábra a Top-5 pontosságot mutatja. Összességében a végső modell 53% top-1 és 74% top-5 pontosságot ért el. Összehasonlíthatjuk ezt az emberi szájról olvasás teljesítményével, amely körülbelül 30% (Altieri et al., 2011). A tévesztési mátrixokban az optimális eset az lenne, ha a mátrix átlójában szerepelne a legtöbb eredmény. Top-1 pontosság esetén (7. ábra) ezt nem sikerült elérni: sok esetben a néma videókat rosszul osztályozta a rendszer. Másrészt a top-5 tévesztési mátrix (8. ábra) több elemet tartalmaz az átló körül, ami azt mutatja, hogy a hálózat elfogadható teljesítménnyel találta meg az alany által kimondott szöveget. A szakirodalmi áttekintés során a GRID adatbázison a többi rendszer nagyságrendileg hasonló eredményt ért el (Ephrat & Peleg, 2017: 52%, Wand et al., 2016: 79.6%), bár ott a nyelv és a szótárméret is jelentősen különbözött a mi kísérleteinktől.

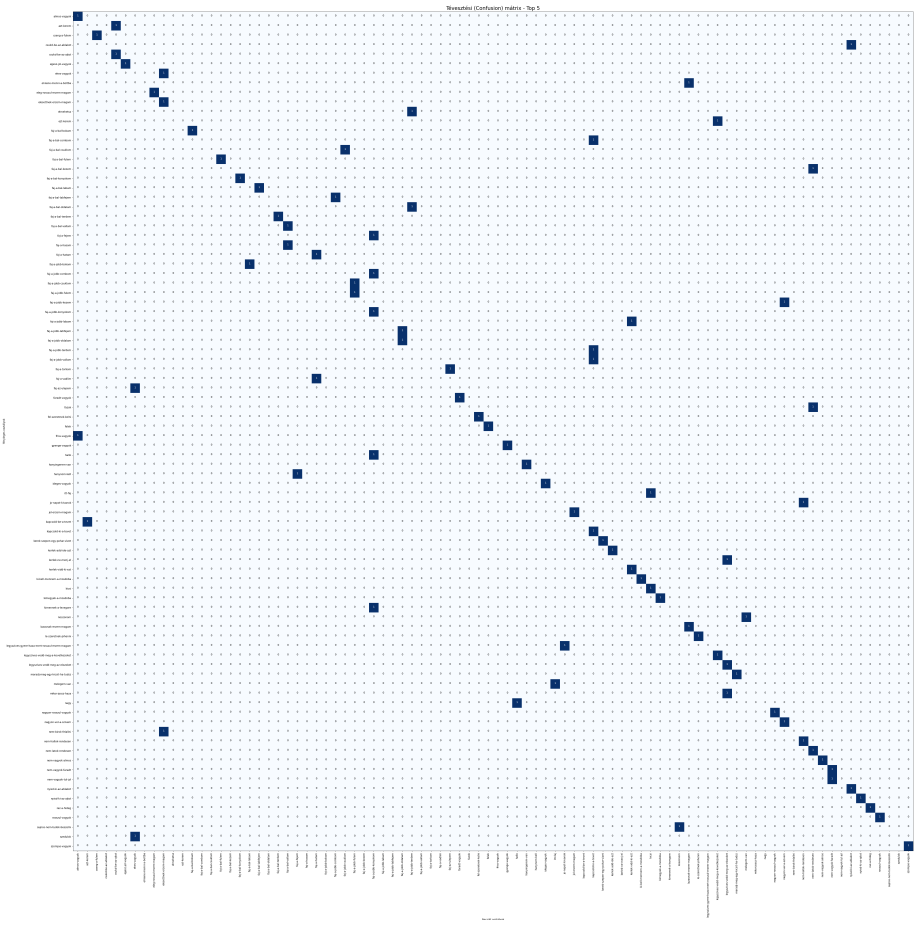
A gyakorlati megvalósításban az alkalmazásszerveren futtatott predikció után a legvalószínűbb felismerési eredményeket visszaküldi a mobil kliensnek. Itt a felhasználó kiválaszthatja, hogy melyik volt a ténylegesen kimondott mondat (lásd 2. ábra, jobb alsó kép), mielőtt elküldené a rendszer szövegfelolvasó moduljába. Ez a lépés biztosítja a megfelelő mondat hangos felolvasását valós kommunikáció esetén.

4. Összegzés

A kutatás során automatikus szájról olvasásra betanított mély neurális hálózatokat terveztünk, majd egy end-to-end alkalmazás architektúrát fejlesztettünk. A rendszer mobilalkalmazást használ a felhasználóval történő interakcióra, míg a gépi tanulási lépések egy háttérszerveren történnek.



7. ábra. Top-1 tévesztési mátrix. (teljes méretben: `https://github.com/victorarthur/vid2speech_images`)



8. ábra. Top-5 tévesztési mátrix. (teljes méretben: `https://github.com/victorarthur/vid2speech_images`)

A kezdeti kiértékelésünk azt mutatja, hogy az eljárás megvalósítható, és az alkalmazást a potenciális végfelhasználók használni tudják majd. A némabeszéd-interfészek elsődleges célfelhasználói a beszédszervi sérüléssel élő emberek (Denby et al., 2010; Gonzalez-Lopez et al., 2020). Ezenkívül az automatikus szájról olvasás hasznos lehet, ha figyelembe vesszük az adatvédelmi aggályokat: egyesek nem érzik jól magukat, ha hangosan kell beszélniük okostelefonjukkal, amikor mások a közelben vannak.

A jövőbeli munkánk során a rendszert a célfelhasználói csoport több tagjával is tesztelni tervezzük. Más, összetettebb hálózatokat is szándékozunk használni, figyelembe véve a valós idejű kommunikációhoz szükséges gyors válaszsebességet.

Köszönetnyilvánítás

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

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.

Altieri, N. A., Pisoni, D. B., & Townsend, J. T. (2011). Some normative data on lip-reading skills (L). *The Journal of the Acoustical Society of America*, *130*, 1–4. doi:10.1121/1.3593376.

Cao, B., Kim, M., Wang, J. R., Van Santen, J., Mau, T., & Wang, J. (2018). Articulation-to-Speech Synthesis Using Articulatory Flesh Point Sensors' Orientation Information. In *Proc. Interspeech* (pp. 3152–3156). Hyderabad, India. doi:10.21437/Interspeech.2018-2484.

Csapó, T. G. (2020). Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract. In *Proc. Interspeech* (pp. 2722–2726). Shanghai, China. doi:10.21437/Interspeech.2020-0015.

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

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

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

Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., & Zafeiriou, S. (2019). The Menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking. *International Journal of Computer Vision*, *127*, 599—624.

Diener, L., & Schultz, T. (2018). Investigating Objective Intelligibility in Real-Time EMG-to-Speech Conversion. In *Proc. Interspeech* (pp. 3162–3166). Hyderabad, India. doi:10.21437/Interspeech.2018-2080.

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

Gonzalez-Lopez, J. A., Gomez-Alanis, A., Martin Donas, J. M., Perez-Cordoba, J. L., & Gomez, A. M. (2020). Silent Speech Interfaces for Speech Restoration: A Review. *IEEE Access*, *8*, 177995–178021. doi:10.1109/access.2020.3026579. arXiv:2009.02110.

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.

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

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

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. doi:10.1145/3242587.3242599.

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