Final Report of the NKFIH/PD_18-127915 project

‘Development of silent speech interfaces and investigation of speaker dependency using ultrasound tongue imaging and electromagnetic articulography’

Dec 1, 2018 – Nov 30, 2021

1. Introduction

During the last several years, there has been significant interest in the articulatory-to-acoustic conversion research field, which is often referred as “Silent Speech Interfaces” (SSI) [1]. This has the main idea of recording the soundless articulatory movement, and automatically generating speech from the movement information, while the original subject is not producing any sound. Such an SSI system can be highly useful for the speaking impaired (e.g. after laryngectomy), and for scenarios where regular speech is not feasible but information should be transmitted from the speaker (e.g. extremely noisy environments; military applications).

Within the FK-17 and PD-18 projects, we mostly contributed to the SSI field with proposing novel methods which are applying Ultrasound Tongue Imaging (UTI). Originally, we expected to use Electromagnetic Articulography (EMA) as well, but after the grant submission, Prof Jun Wang's research lab (U.Texas, USA) achieved significant research results in EMA-to-speech generation and recognition. Therefore, in the current project, instead of EMA, we focused on ultrasound, lip video and MRI of the vocal tract.

During the PD-18 OTKA project, we published 14 conference papers (the majority at top-ranked international conferences like Interspeech, IJCNN, Speech Synthesis Workshop), 6 international journal papers (mostly with Q1/Q2 ranking, sum impact factor: 9.457; and two more manuscripts are still under revision). Three BSc, 10 MSc and 5 PhD students were involved as part of their project laboratory, thesis writing or individual research project. We had significant discussions with international researchers at renowned conferences (e.g. Interspeech 2018, 2019, 2020, 2021, ISSP 2021, SSW11 and DAGA 2021) about methods that they / we use, and also about potential cooperation possibilities (e.g. with Michael Pucher from ÖAW, Austria, a joint Auistrian-Hungarian cooperation grant is foreseen in the topic of Voicebanks and Speech Interfaces).

With the cooperation of Prof Bruce Denby (Sorbonne Université, France), who was the author of the first Silent Speech Interface related paper in 2010 [1], and Dr Michael Wand (IDSIA, Switzerland), who is expert in advanced deep learning [2], we proposed a special issue of the MDPI Sensors journal, entitled "Future Speech Interfaces with Sensors and Machine Intelligence", https://www.mdpi.com/journal/sensors/special_issues/FSI-SMI (see Fig. 1). We contacted a large number of international colleagues in this field, and envisage to receive many significant manuscripts related to SSI. The call is still open until May 2022; currently there are two published papers, and three other submissions are under review or further processing.

1.1 Key questions, goals of the project

The key goals of the project were to 1) enhance spectral filtering of vocoding using articulatory data 2) test and improve direct synthesis in the field of silent speech interfaces 3) analyze session and speaker dependency of the various articulatory tracking equipment and develop alignment methods.

Fig. 1: MDPI Special Issue certificate.

2 Methods, experiments and results

The project was divided into three work packages (WP), according to the main goals, and the results are summarized separately for each WP.

2.1 WP-A: Articulatory motivated spectral filtering in vocoding

During submission of the FK-17 project, Alexander Sepulveda (Industrial University of Santander, Bucaramangua, Colombia) signed a declaration of international cooperation. As a result of this, we hosted an international student at BME (Dagoberto Porras Plata, from the same university in Colombia) for professional internship between Oct – Dec 2018, who was involved in the acoustic-to-articulatory inversion (AAI) experiments of the PD-18 project [3]. We implemented several different Deep Neural Networks (DNNs) to estimate the articulatory information from the acoustic signal [2]. Ultrasound Tongue Imaging was used as the target articulatory information, and we tested two approaches: 1) the EigenTongue space and 2) the raw ultrasound image, and found that raw target data and a simple neural network with two hidden layers were more suitable for this inversion task. After that, we implemented several advanced DNNs (convolutional and recurrent neural networks, see Fig. 2.), to estimate ultrasound images from the acoustic signal. From these experiments, a journal manuscript has been written, which is still under review [4]. Also, within the AAI scenario, we tested real-time MRI of the vocal tract for the target of the DNNs [5]. As the result, LSTMs can achieve smooth generated MR images of the vocal tract, which are similar to the original MRI recordings. This way, we could compare two articulatory imaging methods (namely ultrasound and MRI) in the AAI field.

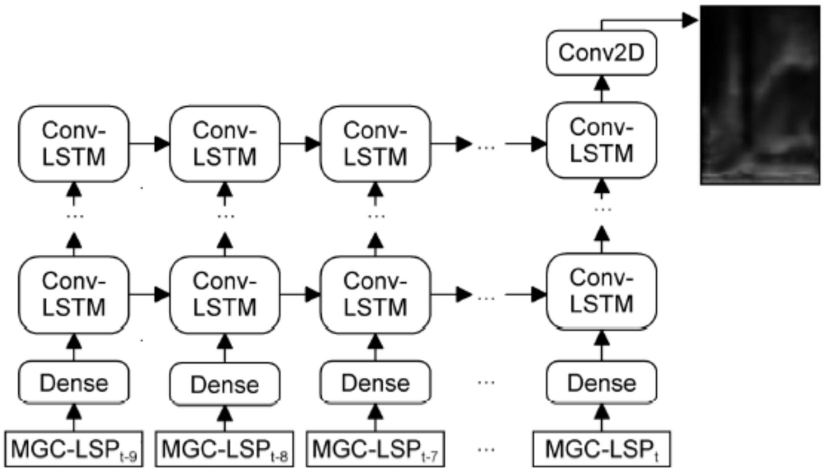

Fig. 2: ConvLSTM network for speech-to-UTI inversion. From [4].

Next, in this WP, we experimented with novel vocoders for text-to-speech (TTS) synthesis and voice conversion (VC) with Mohammed Al-Radhi, being the PhD student of the PI. We proposed a continuous vocoder using continuous F0 (contF0) in combination with Maximum Voiced Frequency (MVF), and applied this for recurrent neural network based voice conversion [6]. The continuous vocoder was applied in DNN-based TTS and tested with both English and Arabic speech [7]. As an extension, Continuous Noise Masking (CNM) was proposed to overcome the issues of simple vocoders (e.g. buzziness) [8]. Within Statistical Voice Conversion (SVC), multiple features from the speech of two speakers (source and target) are converted, using DNNs [9]. We integrated into the SVC framework the continuous vocoder, by converting its contF0, maximum voiced frequency, and spectral features. The continuous vocoder, that we used earlier for UTI-to-F0 prediction [10], was further developed for speech synthesis and voice conversion [11]. We applied Continuous Wavelet Transform (CWT) to characterize and decompose speech features [12]. It can retain the fine spectral envelope and achieve high controllability of the structure closer to human auditory scale. Finally, a demo application 'conTTS' was created [13].

2.2 WP-B: Session and speaker dependency in articulatory-to-acoustic mapping

Within this WP, we cooperated with the participants of the PI's FK-17 project: Gábor Gosztolya, László Tóth, and Amin Shandiz from the University of Szeged, Hungary; and Alexandra Markó from ELTE & Lingual Articulation Research Group, Hungary. The PD-18 and the FK-17 projects have therefore common results.

We compared several types of articulatory acquisition techniques: ultrasound tongue imaging (UTI), lip video (LIP), and vocal tract Magnetic Resonance Imaging (MRI). The advantage of ultrasound is that the full tongue is visible with relatively good spatial and temporal resolution. Lip video, on the other hand, is much simpler to record; but contains significantly less information about the articulation. VT-RMI is significantly more complex and expensive to record, but can provide very detailed information about the full vocal tract.

First, we conducted experiments in ultrasound-to-F0 prediction: we used a simple continuous F0 tracker which does not apply a strict voiced / unvoiced decision [10]. Continuous vocoder parameters (ContF0, Maximum Voiced Frequency and Mel-Generalized Cepstrum) were predicted using separate convolutional neural networks, with ultrasound as input. The methods were tested on four Hungarian speakers (2 males and 2 females), who were recorded during the FK-17 project with the equipment of the Lingual Articulation Research Group. We proposed AutoEncoders for the representation of ultrasound tongue images [14]. Also, we investigated the degree of session-dependency of standard feed-forward DNN-based models for ultrasound-based SSI systems [15]. Besides examining the amount of training data required for speech synthesis parameter estimation, we also showed that DNN adaptation can be useful for handling session dependency [15]. Next, we applied a flow-based neural vocoder (WaveGlow) for ultrasound-to-speech synthesis [16]. Here, the training target was the 80-dimensional mel-spectrogram, which results in a finer detailed spectral representation than earlier methods [16]. Still, using UTI, we tested Generative Adversarial Networks (GAN) for generation of ultrasound images [17]. We compared ultrasound image representations (raw vs. wedge) as the input of SSI systems, of which a journal manuscript is still under review [18]. As a further experiment, we conducted recognition from ultrasound images [19] and compared this with the direct synthesis approach and also with speech synthesis from text [20] (see Fig. 3).

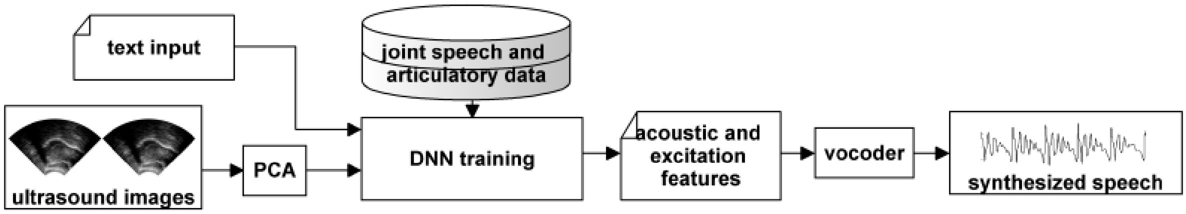

Fig. 3: Comparison of direct ultrasound-to-speech synthesis approach and speech synthesis from text. From [20].

As a second articulatory acquisition technique, we tested lip videos & mouth movement [21]. Inspired by earlier lip-to-speech studies, in our research we designed and implemented models that can generate spectral parameters of speech from lip videos. We used 1000 sentences from a male English speaker of the GRID audiovisual database [22], which contains video from the face of speakers, and synchronous speech. We tested two models that use convolutional and recurrent layers, of which the recurrent neural network was preferred. The results on lip processing might be useful for the PD-18 project, as recording the lip video is simple and cheap compared to other articulatory techniques. We developed a smartphone application that can record silent lip video and synthesize speech from this [23].

The third technique that we investigated within this WP is real-time MRI (rtMRI) of the vocal tract [24]. MRI has not been used before for articulatory-to-acoustic (forward) mapping; although its advantage is that it has a high `relative' spatial resolution. We trained various DNNs for articulatory-to-speech conversion, using rtMRI as input, in a speaker-specific way (Fig. 4 shows a CNN-LSTM network for this purpose). We used American English speakers of the USC-TIMIT articulatory database [25]. We evaluated the results with objective (Normalized MSE and MCD) and subjective measures (perceptual test) and showed that CNN-LSTM networks are preferred (similarly to the earlier cases of UTI and LIP). Another interesting finding of our experiments was that 74% of the recordings of speaker `m1' are out of sync [24]. We contacted the developers of the USC-TMIT database (Asterios Toutios & Shrikanth Narayanan from the University of Southern California, USA) and discussed with them that further investigations are necessary to check this audio-visual synchrony problem. They are open for future collaboration in this field, which can be the basis of a next research grant.

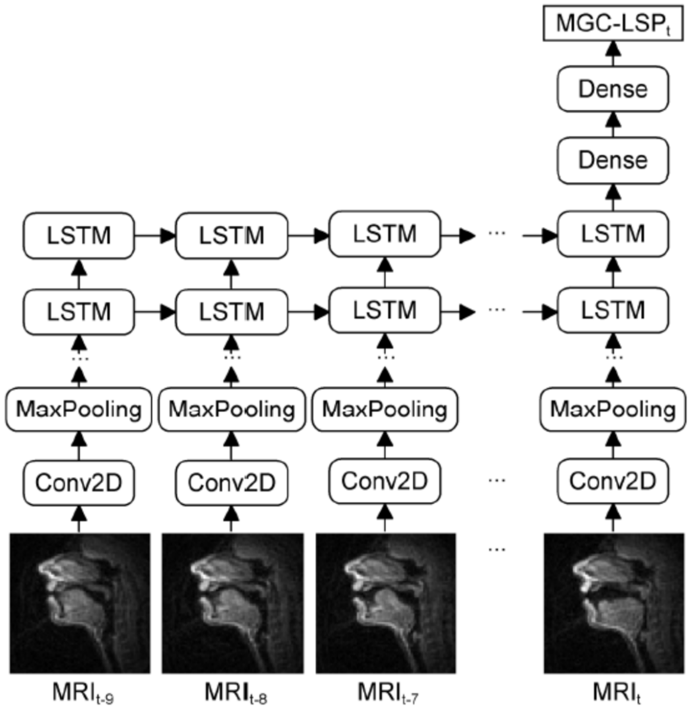

Fig. 4: CNN-LSTM network for MRI-to-speech forward mapping. From [24].

In Sep 2019, as part of the PI's earlier FK-17 and the current PD-18 projects, we invited Maida Percival (PhD student at the University of Toronto, Canada; expert in ultrasound tongue recordings for linguistic purposes) to Budapest; and this co-operation resulted in numerous presentations [26]–[28]. In 2020, during such ultrasound recording sessions, the PI and colleagues observed a serious methodological issue: a limitation of ultrasound tongue imaging is the transducer misalignment during longer data recording sessions. We presented this problem at various conferences last year: Interspeech 2020 [29], UltraFest IX [27], ISSP 2020 [30]; and a large number of researchers confirmed that they have similar issues but do not have the solution yet (e.g. Alan Wrench & Pertti Palo, QMU, UK; Kevin Roon & Wei-Rong Chen & Douglas Whalen, Haskins & YALE & CUNY, USA; Michael Pucher, ÖAW, Austria; Judith Dineley, Augsburg University, Germany; Matthew Faytak, UCLA, USA; Sherman Charles & Steven Lulich, Indiana University, USA; Martijn Wieling, University of Groningen, The Netherlands; Kiyoshi Honda, Tianjin University, China – to mention a few). Therefore, further investigations of speaker dependency and solving the above problem of ultrasound transducer misalignment will be definitely useful for the whole speech community, dealing with articulation or speech production research.

Lastly, we conducted cross-speaker experiments, using x-vectors [31]. Our first attempt to apply them in a multi-speaker silent speech framework brought about a marginal reduction in the error rate of the spectral estimation step. Besides, we created an initial feasibility study for text-to-ultrasound prediction [32]. We extended a traditional (vocoder-based) DNN-TTS framework with predicting PCA-compressed ultrasound images, of which the continuous tongue motion can be reconstructed in synchrony with synthesized speech. Articulatory movement prediction from text input can be useful for audiovisual speech synthesis. A specific application is computer-assisted pronunciation training / computer-aided language learning, which can be beneficial for learners of second languages.

Overall, in this WP, we conducted numerous experiments from various aspects (input representations, machine learning approaches, target representations, and session / speaker dependency), which all can help to create practical Silent Speech Interface systems.

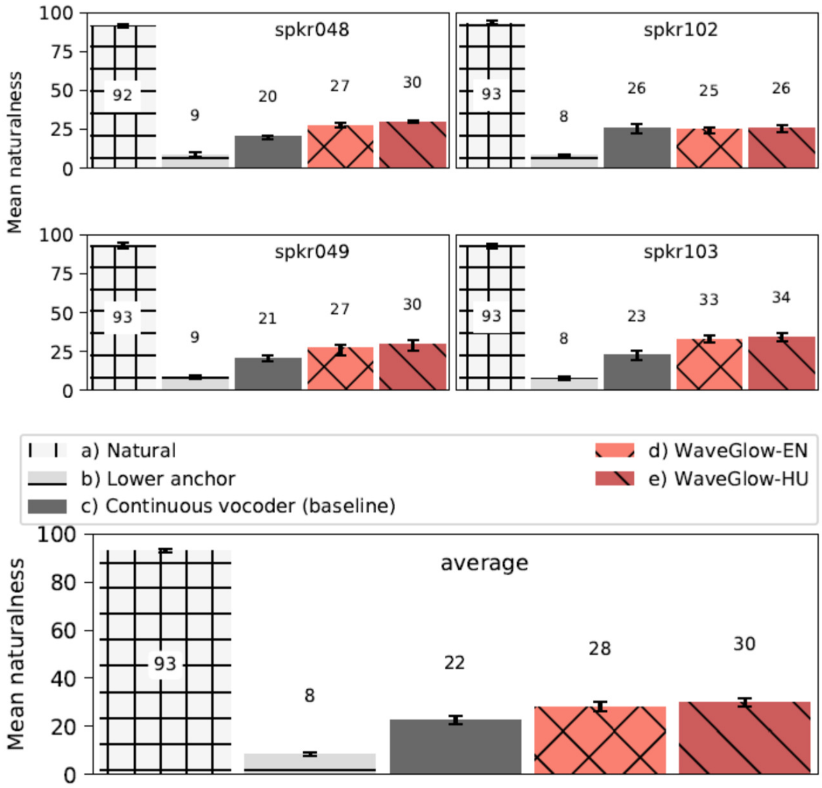

Fig. 5: Results of the subjective evaluation with respect to naturalness, speaker by speaker (top) and average (bottom). The errorbars show the 95% confidence intervals. From [16].

2.3 WP-C: Evaluation and testing with users

Within all of the previous tasks, we performed objective evaluations, and listening tests as a form of subjective evaluations. An example for the results of a MUSHRA-like listening test is presented in Fig. 5, which shows that with the WaveGlow neural vocoder higher quality synthesized speech can be achieved than with previous vocoders [16].

We developed prototype systems for Silent Speech Interfaces. Our developed contributions and prototype systems are available open-source:

- https://github.com/BME-SmartLab/UTI-to-STFT,
- https://github.com/BME-SmartLab/txt-ult2wav,
- https://github.com/BME-SmartLab/txt2ult,
- https://github.com/BME-SmartLab/UTI-to-STFT-Tacotron2,
- https://github.com/BME-SmartLab/UTI-misalignment,
- https://github.com/BME-SmartLab/speech2uti,
- https://github.com/BME-SmartLab/UTI-raw-vs-wedge,
- https://github.com/malradhi/conTTS

Table 1: Students involved in the PD-18 project.

Name	Type / Year	Topic
Dagoberto Porras Plata	MSc internship / 2018-2019	ultrasound & acoustic-to-articulatory inversion
Makrai Márton	PhD individual research topic / 2018-2019	hyperparameter optimization of deep neural networks for ultrasound-to-speech synthesis
Nadia Hajjej	MSc thesis / 2018-2019	application of Generative Adversarial Networks (GANs) for processing of ultrasound images
Amarsaikhan Nasantogtokh	MSc thesis / 2018-2019	applying CycleGANs for ultrasound-speech conversion
Khorkova Mariia	MSc thesis / 2018-2020	lip-to-speech synthesis using deep neural networks
Rácz Bianka	BSc thesis / 2018-2019	lip-to-speech synthesis using convolutional and recurrent neural networks
Varga Kristóf	BSc thesis / 2018-2019	lip-to-speech synthesis using face tracking
Bárány Bálint	MSc thesis / 2018-2020	ultrasound-to-speech synthesis, iterative data loading
Arthur Viktor	MSc thesis / 2018-2020	vid-to-speech mobile application
Maida Percival	PhD individual research topic / 2019-2020	ultrasound session & speaker dependency, linguistic aspects
Amin Shandiz	PhD individual research topic / 2019-2021	data augmentation for UTI; speaker dependency
José Lopez	PhD individual research topic / 2019-2021	testing neural network architectures for ultrasound-to-speech
Mohammed Al-Radhi	PhD individual research topic / 2019-2021	spectral filtering for TTS & VC
Pengyu Dai	MSc thesis / 2019-2020	ultrasound-to-F0
Dóka Zsolt	BSc thesis / 2019-2020	lip-to-speech with MagPhase vocoder
Wu Liang	MSc thesis / 2020-2021	acoustic-to-articulatory inversion using ultrasound
Rysbekova Aliia	MSc thesis / 2020-2021	Comparison of UTI-EMA-MRI-LIP video and its application for deep learning-based articulation-to-speech synthesis
Mohammad Areej Mohammad Mousa	MSc thesis / 2020-2021	Ultrasound Tongue Imaging for Silent Speech Interfaces using deep learning

3 Summary and conclusions

In this project, we first used speaker-dependent neural networks to predict mel-spectrogram parameters from ultrasound tongue image input (in raw scanline representation or wedge orientation). The synthesized speech was either achieved with a continuous vocoder or using WaveGlow inference (trained separately with English and Hungarian data). Later, we investigated solutions for analyzing the misalignment in ultrasound tongue recordings and found that if the transducer moves within the recording session, that can have a negative effect on articulatory-to-acoustic mapping results. Therefore, solutions were proposed for cross-session and cross-speaker ultrasound-to-speech synthesis. During the three years of the project, the main focus was on the ultrasound modality, but we also investigated video of the lip movement and MRI of the vocal tract. Numerous students were involved in the projects, as part of their project laboratory, thesis writing, internship or individual research project. Table 1 summarizes that there were 3 related BSc topics, 10 MSc theses, and 5 PhD research topics.

We presented our results at high-ranked conferences (e.g. Interspeech, ISSP, SSW) and published in top journals (e.g. Multimedia Tools and Applications, Computer Speech and Language). Already during the three years of the project, we have numerous citations in key journals:

- Csapó & Xu, 2020 [29] is cited by Ribeiro et al. 2021 [33] (Speech Communication)
- Al-Radhi et al., 2020 [8] is cited by Aichinger & Pernkopf 2021 [34] (IEEE/ACM Transactions on Audio, Speech and Language Processing)
- Csapó et al., 2020 [16] is cited by Zhang et al., 2021 [35] (IEEE Access)
- Gosztolya et al., 2020 [15] is cited by Gonzalez-Lopez et al., 2020 [36] (IEEE Access)
- Csapó et al., 2019 [10] is cited by Lee et al., 2020 [37] (Sensors)
- Porras et al., 2019 [3] is cited by Eshky et al., 2021 [38] (Speech Communication) and Shahrebabaki et al., 2021 [39] (IEEE/ACM Transactions on Audio, Speech and Language Processing)
- Gosztolya et al., 2019 [14] is cited by Wang et al., 2021 [40] (Journal of the Acoustical Society of America – Express Letters)

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, 2010.

[2] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *Proc. ICASSP*, 2016, pp. 6115–6119.

[3] D. Porras, A. Sepulveda-Sepulveda, and T. G. Csapo, "DNN-based Acoustic-to-Articulatory Inversion using Ultrasound Tongue Imaging," in *Proceedings of the International Joint Conference on Neural Networks*, 2019, vol. 2019-July.

[4] T. G. Csapó and A. Sepúlveda, "Ultrasound Tongue Image Generation for Acoustic-to-Articulatory Inversion using Convolutional and Recurrent Deep Neural Networks," *Submitt. to Multimed. Tools Appl.*, 2021.

[5] T. G. Csapó, "Speaker dependent acoustic-to-articulatory inversion using real-time MRI of the vocal tract," in *Proc. Interspeech*, 2020, pp. 3720–3724.

[6] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "Continuous vocoder applied in deep neural network based voice conversion," *Multimed. Tools Appl.*, vol. 78, no. 23, 2019.

[7] M. S. Al-Radhi, O. Abdo, T. G. Csapó, S. Abdou, G. Németh, and M. Fashal, "A continuous vocoder for statistical parametric speech synthesis and its evaluation using an audio-visual phonetically annotated Arabic corpus," *Comput. Speech Lang.*, vol. 60, p. 101025, Mar. 2020.

[8] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "Continuous Noise Masking Based Vocoder for Statistical Parametric Speech Synthesis," *IEICE Trans. Inf. Syst.*, vol. E103.D, no. 5, pp. 1099–1107, May 2020.

[9] M. S. Al-Radhi, T. G. Csapo, and G. Nemeth, "Parallel voice conversion based on a continuous sinusoidal model," in *2019 10th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019*, 2019.

[10] T. G. Csapó *et al.*, "Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder," in *Proc. Interspeech*, 2019, pp. 894–898.

[11] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "Noise and acoustic modeling with waveform generator in text-to-speech and neutral speech conversion," *Multimed. Tools Appl.*, vol. 80, no. 2, pp. 1969–1994, Jan. 2021.

[12] M. S. Al-Radhi, T. G. Csapó, C. Zainkó, and G. Németh, "Continuous wavelet vocoder-based decomposition of parametric speech waveform synthesis," in *Proc. Interspeech*, 2021, pp. 3266–3270.

[13] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "conTTS: Text-to-Speech Application using a Continuous Vocoder," in *Proc. ISSP*, 2020.

[14] G. Gosztolya, Á. Pintér, L. Tóth, T. Grósz, A. Markó, and T. G. Csapó, "Autoencoder-Based Articulatory-to-Acoustic Mapping for Ultrasound Silent Speech Interfaces," in *International Joint Conference on Neural Networks*, 2019.

[15] G. Gosztolya, T. Grósz, L. Tóth, A. Markó, and T. G. Csapó, "Applying DNN Adaptation to Reduce the Session Dependency of Ultrasound Tongue Imaging-Based Silent Speech Interfaces," *Acta Polytech. Hungarica*, vol. 17, no. 7, pp. 109–124, 2020.

[16] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis," in *Proc. Interspeech*, 2020, pp. 2727–2731.

[17] N. Hajjej and T. G. Csapó, "Realistic Ultrasound Tongue Image Synthesis using Generative Adversarial Networks," *Beszédtudomány - Speech Sci.*, vol. 1, no. 1, pp. 7–21, 2020.

[18] T. G. Csapó, G. Gosztolya, L. Tóth, A. H. Shandiz, and A. Markó, "Optimizing the Ultrasound Tongue Image Representation for Residual Network-based Articulatory-to-Acoustic Mapping," *Submitt. to Multimed. Tools Appl.*, 2021.

[19] C. Zainkó *et al.*, "Adaptation of Tacotron2-based Text-To-Speech for Articulatory-to-Acoustic Mapping using Ultrasound Tongue Imaging," in *Proc. ISCA SSW11*, 2021, pp. 54–59.

[20] T. G. Csapó, L. Tóth, G. Gosztolya, and A. Markó, "Speech Synthesis from Text and Ultrasound Tongue Image-based Articulatory Input," in *Proc. ISCA SSW11*, 2021, pp. 31–36.

[21] B. Rácz and T. G. Csapó, "Ajakvideó alapú beszédszintézis konvolúciós és rekurrens mély

neurális hálózatokkal," *Beszédtudomány -- Speech Sci.*, pp. 57–72, 2020.
[22] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.
[23] F. V. Arthur and T. G. Csapó, "Towards a practical lip-to-speech conversion system using deep neural networks and mobile application frontend," in *2nd International Conference on Artificial Intelligence and Computer Vision (AICV2021)*, 2021.
[24] T. G. Csapó, "Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract," in *Proc. Interspeech*, 2020, pp. 2722–2726.
[25] S. Narayanan *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1307–1311, Sep. 2014.
[26] M. Percival, T. G. Csapó, M. Bartók, A. Deme, T. E. Gráczi, and A. Markó, "Gemination as fortition? Articulatory data from Hungarian," in *LabPhon*, 2020.
[27] M. Percival, T. G. Csapó, M. Bartók, A. Deme, T. E. Gráczi, and A. Markó, "Ultrasound imaging of Hungarian geminates," in *UltraFest IX*, 2020.
[28] M. Percival, T. G. Csapó, M. Bartók, A. Deme, T. E. Gráczi, and A. Markó, "Tongue root and voicing in Hungarian singleton and geminate obstruents," in *12th International Seminar on Speech Production*, 2020.
[29] T. G. Csapó and K. Xu, "Quantification of Transducer Misalignment in Ultrasound Tongue Imaging," in *Proc. Interspeech*, 2020, pp. 3735–3739.
[30] T. G. Csapó, K. Xu, A. Deme, T. E. Gráczi, and A. Markó, "Transducer Misalignment in Ultrasound Tongue Imaging," in *12th International Seminar on Speech Production*, 2020.
[31] A. H. Shandiz, Ĺ. Toth, Ǵabor Gosztolya, A. Marḱo, and T. Ǵabor Csaṕo, "Neural speaker embeddings for ultrasound-based silent speech interfaces," in *Proc. Interspeech*, 2021, vol. 1, pp. 151–155.
[32] T. G. Csapó, "Extending Text-to-Speech Synthesis with Articulatory Movement Prediction using Ultrasound Tongue Imaging," in *Proc. ISCA SSW11*, 2021, pp. 7–12.
[33] M. S. Ribeiro, J. Cleland, A. Eshky, K. Richmond, and S. Renals, "Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors," *Speech Commun.*, vol. 128, pp. 24–34, Apr. 2021.
[34] P. Aichinger and F. Pernkopf, "Synthesis and Analysis-By-Synthesis of Modulated Diplophonic Glottal Area Waveforms," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 914–926, 2021.
[35] J. Zhang, P. Roussel, and B. Denby, "Creating Song from Lip and Tongue Videos with a Convolutional Vocoder," *IEEE Access*, vol. 9, pp. 13076–13082, 2021.
[36] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martin Donas, J. L. Perez-Cordoba, and A. M. Gomez, "Silent Speech Interfaces for Speech Restoration: A Review," *IEEE Access*, vol. 8, pp. 177995–178021, Sep. 2020.
[37] W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov, and S. Lee, "Biosignal sensors and deep learning-based speech recognition: A review," *Sensors*, vol. 21, no. 4, pp. 1–22, 2021.
[38] A. Eshky, J. Cleland, M. S. Ribeiro, E. Sugden, K. Richmond, and S. Renals, "Automatic audiovisual synchronisation for ultrasound tongue imaging," *Speech Commun.*, vol. 132, pp. 83–95, Sep. 2021.
[39] A. Sabzishahrebabaki, G. Salvi, T. K. Svendsen, and S. M. Siniscalchi, "Acoustic-to-Articulatory Mapping with Joint Optimization of Deep Speech Enhancement and Articulatory Inversion Models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 1–1, 2021.
[40] H. Wang, P. Roussel, and B. Denby, "Improving ultrasound-based multimodal speech recognition with predictive features from representation learning," *JASA Express Lett.*, vol. 1, no. 1, p. 015205, Jan. 2021.