INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Ultrasound-based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis

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

[1]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
[2]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
[3]Institute of Informatics, University of Szeged, Hungary
[4]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[5]Department of Applied Linguistics and Phonetics, Eötvös Loránd University, Budapest, Hungary

{csapot, zainko}@tmit.bme.hu, {tothl, ggabor}@inf.u-szeged.hu, marko.alexandra@btk.elte.hu

Abstract

For articulatory-to-acoustic mapping using deep neural networks, typically spectral and excitation parameters of vocoders have been used as the training targets. However, vocoding often results in buzzy and muffled final speech quality. Therefore, in this paper on ultrasound-based articulatory-to-acoustic conversion, we use a flow-based neural vocoder (WaveGlow) pre-trained on a large amount of English and Hungarian speech data. The inputs of the convolutional neural network are ultrasound tongue images. The training target is the 80-dimensional mel-spectrogram, which results in a finer detailed spectral representation than the previously used 25-dimensional Mel-Generalized Cepstrum. From the output of the ultrasound-to-mel-spectrogram prediction, WaveGlow inference results in synthesized speech. We compare the proposed WaveGlow-based system with a continuous vocoder which does not use strict voiced/unvoiced decision when predicting F0. The results demonstrate that during the articulatory-to-acoustic mapping experiments, the WaveGlow neural vocoder produces significantly more natural synthesized speech than the baseline system. Besides, the advantage of WaveGlow is that F0 is included in the mel-spectrogram representation, and it is not necessary to predict the excitation separately.

Index Terms: articulatory-to-acoustic mapping, articulation-to-F0, end-to-end

1. Introduction

Articulatory-to-acoustic mapping methods aim to synthesize speech signal directly from articulatory input, applying the theory that articulatory movements are directly linked with the acoustic speech signal in the speech production process. A recent potential application of this mapping is a “Silent Speech Interface” (SSI [1, 2]), which has the main idea of recording the soundless articulatory movement, and automatically generating speech from the movement information, while the subject is not producing any sound. Such an SSI system can be highly useful for the speaking impaired (e.g. after laryngectomy or elderly people), and for scenarios where regular speech is not feasible, but information should be transmitted from the speaker (e.g. extremely noisy environments or military applications).

For the articulatory-to-acoustic mapping, the typical input can be electromagnetic articulography (EMA) [3, 4], ultrasound tongue imaging (UTI) [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], permanent magnetic articulography (PMA) [16], surface electromyography (sEMG) [17, 18], Non-Audible Murmur (NAM) [19], electro-optical stomatography [20] or video of the lip movements [6, 21, 22]. From another aspect, there are two distinct ways of SSI solutions, namely ‘direct synthesis’ and ‘recognition-and-synthesis’ [2]. In the first case, the speech signal is generated without an intermediate step, directly from the articulatory data [3, 4, 5, 7, 8, 10, 11, 13, 14, 15, 16, 17, 18, 19, 21]. In the second case, silent speech recognition (SSR) is applied on the biosignal which extracts the content spoken by the person (i.e. the result of this step is text); this step is then followed by text-to-speech (TTS) synthesis [6, 9, 12, 20, 22]. In the SSR+TTS approach, any information related to speech prosody is lost, whereas it may be kept with direct synthesis. Also, the smaller delay by the direct synthesis approach might enable conversational use.

For the direct conversion, typically, vocoders are used, which synthesize speech from the spectral parameters predicted by the DNNs from the articulatory input. One of the spectral representations that was found to be useful earlier for statistical parametric speech synthesis is Mel-Generalized Cepstrum in Line Spectral Pair form (MGC-LSP) [23, 24].

The early studies on articulatory-to-acoustic mapping typically applied a low-order spectral representation, for example, only 12 coefficients were used in [8, 10]. Later, our team also experimented with using 22 kHz speech and 24-order MGC-LSP target [15] (with the gain, having 25 dimensions altogether). Still, the 24-order MGC-LSP target is a relatively low-dimensional spectral representation, and this simple vocoder that we used in previous studies [10, 11, 12, 13, 14, 15] can be a bottleneck in the ultrasound-to-speech mapping framework.

Besides the spectrum, the other aspect of direct conversion is to predict the source / excitation information, e.g. the fundamental frequency of speech. There have been only a few studies that attempted to predict the voicing feature and the F0 curve using ultrasound as input. Hueber et al. experimented with predicting the V/UV parameter along with the spectral features of a vocoder, using ultrasound and lip video as input articulatory data [7]. They applied a feed-forward deep neural network (DNN) for the V/UV prediction and achieved 82% accuracy. In [11], we experimented with deep neural networks to perform articulatory-to-acoustic conversion from ultrasound frames (raw scanlines), with an emphasis on estimating the voicing feature and the F0 curve from the ultrasound input. We attained a correlation rate of 0.74 between the original and the predicted F0 curves, and an accuracy of 87% in

http://dx.doi.org/10.21437/Interspeech.2020-1031

V/UV prediction (using the voice of only one speaker). However, in several cases, the inaccurate estimation of the voicing feature caused audible artefacts. Most recently, we used a continuous vocoder (which does not have a strict voiced/unvoiced decision [23, 24, 25, 26]), for the CNN-based UTI-to-F0 and UTI-to-MGC-LSP prediction [15]. In the experiments with two male and two female speakers, we found that the continuous F0 was predicted with lower error, and the continuous vocoder produced similarly natural synthesized speech as the baseline vocoder using standard discontinuous F0. Also, the advantage of the improved vocoder is that, as all parameters are continuous, it is not necessary to train a separate network in classification mode for the voiced/unvoiced prediction.

1.1. Neural vocoders

Since the introduction of WaveNet in 2016 [27], neural vocoders are an exciting way of generating the raw samples of speech during text-to-speech synthesis (TTS). However, a problem with early WaveNet-like models was that they were computationally extremely expensive. Currently, state-of-the-art TTS models are based on parametric neural networks using improved versions of WaveNet-like neural vocoders. TTS synthesis is typically done in two steps: 1) the first step transforms the text into time-aligned features, such as a mel-spectrogram, 2) the second step transforms these spectral features to the speech signal. If we replace the first step (text-to-spectrogram) with articulation-to-spectrogram prediction, we can use the recent advances of the latter step directly for the purpose of articulation-to-speech synthesis.

One of the most recent types of neural vocoders, WaveGlow [28] is a flow-based network capable of generating high-quality speech from mel-spectrograms. The advantage of the WaveGlow model is that it is relatively simple, yet the synthesis can be done faster than real-time. It can generate audio by sampling from a distribution (zero mean spherical Gaussian), conditioned on a mel-spectrogram.

1.2. Contributions of this paper

In this paper, we extend our earlier work on ultrasound-based articulatory-to-acoustic mapping. From the ultrasound tongue raw scanline input, we predict 80-dimensional STFT spectral representation, from which we synthesize speech with a WaveGlow model. We show that the use of a neural vocoder is advantageous compared to earlier vocoders, which applied source-filter separation.

2. Methods

2.1. Articulatory data acquisition

Two Hungarian male and two female subjects were recorded while reading sentences aloud (altogether 209 sentences each). The tongue movement was recorded in midsagittal orientation using the “Micro” ultrasound system of Articulate Instruments Ltd. at 81.67 fps. The speech signal was recorded with a Beyerdynamic TG H56c tan omnidirectional condenser microphone. The ultrasound data and the audio signals were synchronized using the tools provided by Articulate Instruments Ltd. In our experiments, the raw scanline data of the ultrasound was used as input of the networks, after being resized to 64×128 pixels using bicubic interpolation. More details about the recording set-up and articulatory data can be found in [10]. The duration of the recordings was about 15 minutes, which was partitioned into training, validation and test sets in a 85-10-5 ratio.

2.2. Speech data for neural vocoder

WaveGlow [28] provides a pretrained model trained on the LJSpeech database, from 24 hours of English audiobooks with a single female speaker. Our informal listening tests showed that the single speaker WaveGlow model can generate both male and female voice samples, but it performs weakly with low F0 values (which is typical for male speakers). However, since we have both male and female articulatory and acoustic recordings, we hypothesized that a multispeaker WaveGlow model will be more suitable for synthesizing speech. Therefore, we chose 5 male and 6 female Hungarian speakers (altogether 23k sentences, roughly 22 hours) from the PPSD database [29].

2.3. Continuous vocoder (baseline)

In the baseline vocoder, first, the speech recordings (digitized at 22 kHz) were analyzed using MGLSA [30] at a frame shift of 22 050 Hz / 81.67 fps = 270 samples, which resulted in 24-order spectral (MGC-LSP) features [31]. Next, continuous F0 (ContF0) is calculated on the input waveforms using the simple continuous pitch tracker [32]. After this step, Maximum Voiced Frequency (MVF) is calculated from the speech signal [33, 24]. The continuous vocoder parameters (MGC-LSP, log-ContF0 and log-MVF) served as the training targets of the neural network in our speech synthesis experiments.

During the synthesis phase, voiced excitation is composed of residual excitation frames overlap-added pitch synchronously [23, 24, 26]. This voiced excitation is lowpass filtered frame by frame at the frequency given by the MVF parameter. In the frequency range higher than the actual value of MVF, white noise is used. The voiced and unvoiced excitation is added together. Finally, an MGLSA filter is used to synthesize speech from the excitation and the MGC parameter stream [30].

2.4. WaveGlow neural vocoder

During analysis, the mel-spectrogram was estimated from the Hungarian speech recordings (digitized at 22 kHz). Similarly to the original WaveGlow paper [28], 80 bins were used for mel-spectrogram using librosa mel-filter defaults (i.e. each bin is normalized by the filter length and the scale is the same as in HTK). FFT size and window size were both 1024 samples. For hop size, we chose 270 samples, in order to be in synchrony with the articulatory data. This 80-dimensional mel-spectrogram served as the training target of the neural network.

NVIDIA provided a pretrained WaveGlow model using the LJSpeech database (WaveGlow-EN). Besides, another WaveGlow model was trained with the Hungarian data (WaveGlow-HU). This latter training was done on a server with eight V100 GPUs, altogether for 635k iterations.

In the synthesis phase, an interpolation in time was necessary, as the original WaveGlow models were trained with 22 kHz speech and 256 samples frame shift; for this we applied bicubic interpolation. Next, to smooth the predicted data, we used a Savitzky-Golay filter with a window size of five, and cubic interpolation. Finally, the synthesized speech is the result of the inference with the trained WaveGlow model (EN/HU) conditioned on the mel-spectrogram input [28].

2.5. DNN training with the baseline vocoder

Similarly to our previous study [15], here we used convolutional neural networks (CNN), but we further optimized manu-

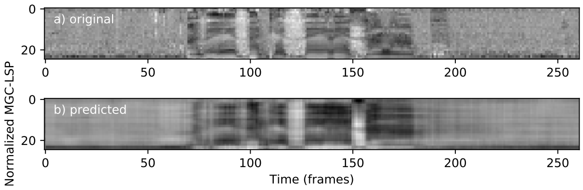

Figure 1: *Demonstration samples from a female speaker: normalized MGC-LSP using the baseline system.*

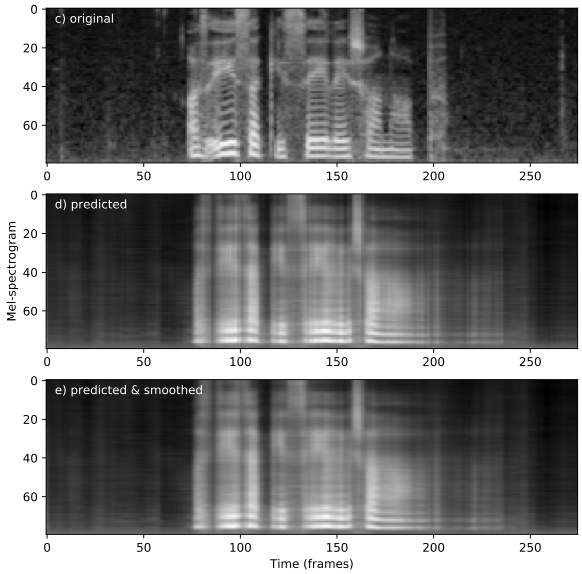

Figure 2: *Demonstration samples from a female speaker: normalized Mel-spectrogram using the proposed system.*

ally the network structure and parameters. We trained speaker-specific CNN models using the training data (roughly 180 sentences). For each speaker, two neural networks were trained: one CNN for predicting the excitation features (log-ContF0 and log-MVF), and one for predicting the 25-dimensional MGC-LSP. All CNNs had one 64×128 pixel ultrasound image as input, and had the same structure: two convolutional layers (kernel size: 13×13, number of filters: 30 and 60), followed by max-pooling; and again two convolutional layers (filters: 90 and 120), followed by max-pooling. Finally, a fully connected layer was used with 1000 neurons. In all hidden layers, the Swish activation was used [34], and we applied dropout with 0.2 probability. The cost function applied for the regression task was the mean-squared error (MSE). We used the SGD optimizer with manually chosen learning rate. We applied early stopping to avoid over-fitting: the network was trained for 100 epochs and was stopped when the validation loss did not decrease within 3 epochs.

2.6. DNN training with the WaveGlow neural vocoder

In the proposed system, one CNN is used for each speaker, with the same structure as for the baseline system: two convolutional layers, max-pooling, two convolutional layers, max-pooling, and a fully connected layer with 1000 neurons. The network had 64×128 pixel images as input and was predicting the 80-dimensional mel-spectrogram features. The training procedures were the same as in the baseline setup.

Table 1: *MCD scores on the test set.*

	Mel-Cepstral Distortion (dB)		
Speaker	Continuous Vocoder	WaveGlow-EN	WaveGlow-HU
Speaker #048	5.54	5.27	5.34
Speaker #049	5.67	5.66	5.65
Speaker #102	5.26	5.20	5.18
Speaker #103	5.41	5.34	5.37
Mean	5.47	5.37	5.38

3. Experimental results

3.1. Demonstration sample

A sample sentence (not being present in the training data) was chosen for demonstrating how the baseline and the proposed systems deal with the prediction of spectral parameters. Fig. 1 shows the spectral features of the baseline system (original and predicted MGC-LSP, normalized to zero mean and unit variance). In general, the generated speech starts at similar time as in the original recording, but it lasts longer – probably as a result of inaudible post-speech tongue movement (around frames 190–220). Fig. 2 shows an example for the mel-spectogram prediction with the proposed system. The reason for the misaligned time scale is the different hop size (Fig. 1: 270 samples, Fig. 2: 256 samples). In the proposed system, the spectral details are similarly over-smoothed as in the baseline system, but the 80-dimensional mel-spectrogram contains more detailed information. Another difference between the baseline and the proposed systems is the way how they handle the voicing feature of speech excitation (i.e., in the proposed approach, the F0 information is included in the mel-spectrogram).

3.2. Objective evaluation

After training the CNNs for each speaker and feature individually, we synthesized sentences, and measured objective differences. For this, Mean Square Error is not a suitable measure, as the (normalized) MGC-LSP features and the mel-spectrogram values have different scales; therefore, the MSE values by the baseline and proposed systems are not directly comparable. Instead of MSE, the objective metric chosen in this test is the Mel-Cepstral Distortion (MCD, [35]). Lower MCD values indicate higher similarity. This metric not only evaluates the distance in the cepstral domain but also uses a perceptual scale in an effort to improve the accuracy of objective assessments, and is a standard way to evaluate text-to-speech synthesis systems. In general, the advantage of MCD is that it is better correlated with perceptual scores than other objective measures [35].

Table 1 shows the MCD results in dB for all the synthesized sentence types and speakers (lower values indicate higher spectral similarity). The best (lowest) MCD result, with an average value of 5.37 dB, was achieved when predicting mel-spectogram features and synthesizing with the English WaveGlow model. The Hungarian WaveGlow model scores second with an average MCD of 5.38 dB. Finally, the highest average error is attained by the baseline vocoder (MCD: 5.47 dB). By checking the MCD values speaker by speaker, we can see that both WaveGlow-EN and WaveGlow-HU were better than the baseline for all speakers, while there is some speaker dependency between the English and Hungarian versions (but these differences are mostly small in scale). Overall, according to the Mel-Ceptral Distortion measure, the proposed system is clearly better than the baseline vocoder.

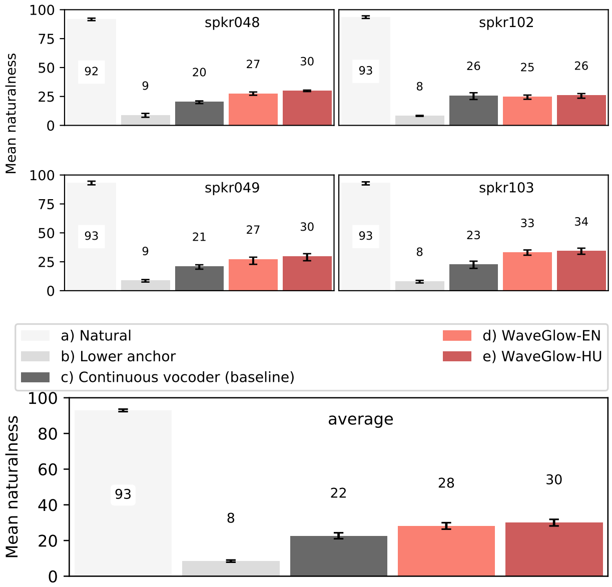

Figure 3: *Results of the subjective evaluation with respect to naturalness, speaker by speaker (top) and average (bottom). The errorbars show the 95% confidence intervals.*

3.3. Subjective listening test

In order to determine which proposed version is closer to natural speech, we conducted an online MUSHRA-like test [36].

Our aim was to compare the natural sentences with the synthesized sentences of the baseline, the proposed approaches and a lower anchor system (the latter having constant F0 and predicted MGC-LSP). In the test, the listeners had to rate the naturalness of each stimulus in a randomized order relative to the reference (which was the natural sentence), from 0 (very unnatural) to 100 (very natural). We chose four sentences from the test set of each speaker (altogether 16 sentences). The variants appeared in randomized order (different for each listener). The samples can be found at `http://smartlab.tmit.bme.hu/interspeech2020_UTI-to-STFT`.

Each sentence was rated by 22 native Hungarian speakers (24 females, 2 males; 19–43 years old). On average, the test took 13 minutes to complete. Fig. 3 shows the average naturalness scores for the tested approaches. The lower anchor version achieved the lowest scores, while the natural sentences were rated the highest, as expected. The proposed neural vocoder based versions (WaveGlow-EN and WaveGlow-HU) were preferred over the baseline continuous vocoder, except for speaker #102, for whom they were rated as equal. In all cases, WaveGlow-HU was slightly preferred over WaveGlow-EN.

To check the statistical significances we conducted Mann-Whitney-Wilcoxon ranksum tests with a 95% confidence level. Based on this, both WaveGlow-EN and WaveGlow-HU are significantly different from the baseline vocoder, but the difference between the English and Hungarian WaveGlow versions is not statistically significant. When checking the significances speaker by speaker, the same tendencies can be seen (WaveGlow-EN = WaveGlow-HU > baseline), except for speaker #102, for whom the differences are not statistically significant between the baseline and proposed systems.

As a summary of the listening test, a significant preference towards the proposed WaveGlow-EN/WaveGlow-HU models could be observed.

4. Discussion and conclusions

In this paper, we used speaker-dependent convolutional neural networks to predict mel-spectrogram parameters from ultrasound tongue image input (in raw scanline representation). The synthesized speech was achieved using WaveGlow inference (trained separately with English and Hungarian data). We compared the proposed model with a baseline continuous vocoder, in which continuous F0, Maximum Voiced Frequency and MGC-LSP spectral features were predicted separately.

The results of the objective evaluation demonstrated that during the articulatory-to-acoustic mapping experiments, the spectral features are predicted with lower Mel-Cepstral Distortion using the proposed WaveGlow/mel-spectrogram model than with the baseline (5.37 dB vs. 5.47 dB). According to the subjective listening test, the WaveGlow flow-based neural vocoder produces more natural synthesized speech compared to the continuous vocoder baseline, for three out of the four speakers. By informally listening to the synthesized sentences of the male speakers, we found that the 80-dimensional mel-spectrogram representation of WaveGlow was not enough to capture changes in F0. Therefore, the final synthesized sentences for male speakers (especially for #102) are less natural. This could be improved by either a higher dimensional spectral representation (which, in practice is not easy, as both WaveGlow models were trained with 80-D mel-spectrogram), or by non-linearly reshaping the mel-spectrogram during the articulatory-to-acoustic mapping, i.e. adding more emphasis to the lower part of the spectrum, which contains F0 information.

The advantage of WaveGlow is that F0 is included in the mel-spectrogram representation, and it is not necessary to predict the excitation separately. In the baseline continuous vocoder [15], separate CNN models were used for the excitation and spectral prediction. Although here we did not measure the accuracy of F0 prediction separately, from the subjective listening test it is clear that the mel-spectrogram can represent the excitation information well for high F0, but not for low F0 speakers. The disadvantage of WaveGlow is that for training the neural vocoders, a huge amount of speech data is necessary (24 hours were used for the English model [28] and 22 hours for the Hungarian model). From this point of view, the continuous vocoder is much simpler, as it has 2-dimensional excitation, 25-dimensional spectral features, and no data is required to train the vocoder itself. Besides, it gives controllability, which usually is not fully supported by neural vocoders.

In the future, we plan to apply the above articulatory-to-acoustic prediction framework with the flow-based neural vocoder for other articulatory modalities (e.g. lip or rtMRI).

The keras implementations are accessible at `https://github.com/BME-SmartLab/UTI-to-STFT/`.

5. Acknowledgements

The authors were partly funded by the NRDIO of Hungary (FK 124584, PD 127915 grants). T.G.Cs. and Cs.Z. were also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825619 (AI4EU). G.G. and L.T. were funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences, and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-20-5. L.T. was supported by the TUDFO/47138-1/2019-ITM project. The Titan Xp used for this research was donated by the NVIDIA Corporation. We thank the listeners for participating in the subjective test.

6. References

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

[2] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-Based Spoken Communication: A Survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, dec 2017.

[3] B. Cao, M. Kim, J. R. Wang, J. Van Santen, T. Mau, and J. Wang, "Articulation-to-Speech Synthesis Using Articulatory Flesh Point Sensors' Orientation Information," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 3152–3156.

[4] F. Taguchi and T. Kaburagi, "Articulatory-to-speech conversion using bi-directional long short-term memory," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 2499–2503.

[5] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in *Proc. ICASSP*, Montreal, Quebec, Canada, 2004, pp. 685–688.

[6] T. Hueber, E.-L. Benaroya, G. Chollet, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[7] T. Hueber, E.-l. Benaroya, B. Denby, and G. Chollet, "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 593–596.

[8] A. Jaumard-Hakoun, K. Xu, C. Leboullenger, P. Roussel-Ragot, and B. Denby, "An Articulatory-Based Singing Voice Synthesis Using Tongue and Lips Imaging," in *Proc. Interspeech*, San Francisco, CA, USA, 2016, pp. 1467–1471.

[9] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *Proc. ICASSP*, New Orleans, LA, USA, 2017, pp. 2971–2975.

[10] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3672–3676.

[11] T. Grósz, G. Gosztolya, L. Tóth, T. G. Csapó, and A. Markó, "F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 291–295.

[12] L. Tóth, G. Gosztolya, T. Grósz, A. Markó, and T. G. Csapó, "Multi-Task Learning of Phonetic Labels and Speech Synthesis Parameters for Ultrasound-Based Silent Speech Interfaces," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 3172–3176.

[13] E. Moliner and T. G. Csapó, "Ultrasound-based silent speech interface using convolutional and recurrent neural networks," *Acta Acustica united with Acustica*, vol. 105, no. 4, pp. 587–590, 2019.

[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] T. G. Csapó, M. S. Al-Radhi, G. Németh, G. Gosztolya, T. Grósz, L. Tóth, and A. Markó, "Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 894–898.

[16] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, dec 2017.

[17] M. Janke and L. Diener, "EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, dec 2017.

[18] L. Diener, G. Felsch, M. Angrick, and T. Schultz, "Session-Independent Array-Based EMG-to-Speech Conversion using Convolutional Neural Networks," in *13th ITG Conference on Speech Communication*, 2018.

[19] N. Shah, N. Shah, and H. Patil, "Effectiveness of Generative Adversarial Network for Non-Audible Murmur-to-Whisper Speech Conversion," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 3157–3161.

[20] S. Stone and P. Birkholz, "Silent-speech command word recognition using electro-optical stomatography," in *Proc. Interspeech*, San Francisco, CA, USA, 2016, pp. 2350–2351.

[21] A. Ephrat and S. Peleg, "Vid2speech: Speech Reconstruction from Silent Video," in *Proc. ICASSP*, New Orleans, LA, USA, 2017, pp. 5095–5099.

[22] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi, "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*, Berlin, Germany, 2018, pp. 581–593.

[23] T. G. Csapó, G. Németh, and M. Cernak, "Residual-Based Excitation with Continuous F0 Modeling in HMM-Based Speech Synthesis," in *Lecture Notes in Artificial Intelligence*, A.-H. Dediu, C. Martín-Vide, and K. Vicsi, Eds. Budapest, Hungary: Springer International Publishing, 2015, vol. 9449, pp. 27–38.

[24] T. G. Csapó, G. Németh, M. Cernak, and P. N. Garner, "Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in *Proc. EUSIPCO*, Budapest, Hungary, 2016, pp. 1338–1342.

[25] B. P. Tóth and T. G. Csapó, "Continuous Fundamental Frequency Prediction with Deep Neural Networks," in *Proc. EUSIPCO*, Budapest, Hungary, 2016, pp. 1348–1352.

[26] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "Time-Domain Envelope Modulating the Noise Component of Excitation in a Continuous Residual-Based Vocoder for Statistical Parametric Speech Synthesis," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 434–438.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *CoRR*, vol. abs/1609.0, 2016.

[28] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 3617–3621.

[29] G. Olaszy, "Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai [Development and services of a Hungarian precisely labeled and segmented, parallel speech database] (in Hungarian)," *Beszédkutatás 2013 [Speech Research 2013]*, pp. 261–270, 2013.

[30] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[31] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, 1994, pp. 1043–1046.

[32] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, 2013.

[33] T. Drugman and Y. Stylianou, "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1230–1234, 2014.

[34] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: a Self-Gated Activation Function," *ArXiv e-prints*, Oct. 2017. [Online]. Available: https://arxiv.org/abs/1710.05941v1

[35] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. ICASSP*, Victoria, Canada, 1993, pp. 125–128.

[36] "ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality," 2001.