A Comparison of Data Augmentation Methods on Ultrasound Tongue Images for Articulatory-to-Acoustic Mapping towards Silent Speech Interfaces

Ibrahim Ibrahimov
Institute of Informatics
University of Szeged
Szeged, Hungary
ibrahimkhaliloglu@gmail.com

Gábor Gosztolya
ELKH-SZTE Research Group on Artificial Intelligence
Szeged, Hungary
ggabor@inf.u-szeged.hu

Tamás Gábor Csapó
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
csapot@tmit.bme.hu

***Abstract*—Silent Speech Interfaces (SSI), being a subfield of speech technology, break the limitations of automatic speech recognition when acoustic signals cannot be produced or clearly captured. SSI focuses on the articulation process of speech production in order to map articulatory data into acoustics. Ultrasound tongue imaging (UTI), a non-invasive, clinically safe technique to view the shape, position, and movements of the tongue, has recently become popular in the process of collecting articulatory data of the tongue movement. Despite advancements in the field of SSI, the majority of related research has been conducted using limited datasets due to challenges in acquiring additional information, which results in overfitting. It has already been shown that data augmentation can be helpful for solving the overfitting problem and improving the generalization ability of deep neural networks. In this paper, we discuss the preliminary implementation and comparison of data augmentation methods on Azerbaijani ultrasound and speech recordings that has been recorded by our team. These strategies include consecutive and intermittent time masking, sinusoidal noise injection, and random scaling. We explore the generation of new data samples using the provided methods on the dataset. We use mean-squared error validation loss as an evaluation metric to measure the performance of all the above data augmentation methods.**

***Index Terms*—data augmentation, silent speech interfaces, ultrasound tongue imaging, speech technology**

I. Introduction

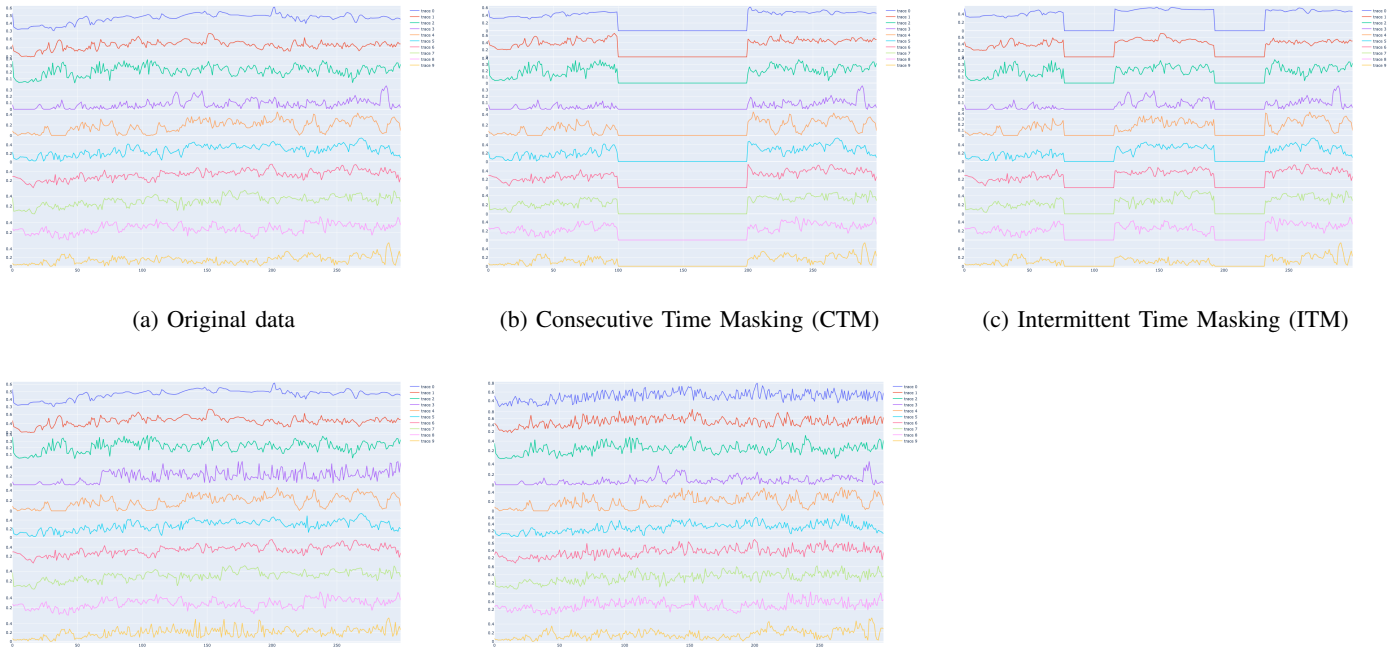
Silent speech interfaces (SSI) are systems that aim to provide speech communication when audible acoustic signals are not available. SSI is an assistive device to produce a digital representation of speech, which can be synthesized directly by obtaining articulatory data from elements of the speech production process (articulators, such as the tongue, lips, jaw, etc.) [1]. Since these systems have been shown to provide speech communication without acoustic signals, they offer a profoundly new way of restoring communication capabilities to those with speech impairments [2], [3]. Other than clinical use, potential applications of SSIs include improving oral communication in noisy environments and having private conversations over the phone in public places, as the articulators are mainly insensitive to noise.

There are two different types of algorithmic design in SSI [4]. One of the articulation-to-speech conversion designs converts articulations to text with silent speech recognition (SSR) and runs text-to-speech synthesis (TTS). The SSR+TTS model always causes a delay because SSR takes time to decode and TTS requires two separate text-processing and analysis stages [5]–[8]. An alternative method of SSI is direct-synthesis design, and in this type, the focus is on synthesizing speech signal directly from articulatory data input (which is called articulatory-to-acoustic mapping (AAM)). This algorithmic design relies on the theory that articulatory movements are directly linked with acoustic speech signal in the speech production process, typically using vocoders [9]–[12]. Although the quality of direct synthesis is not as good as TTS due to a lack of textual information, the speech output of articulatory-to-acoustics mapping has recently been improved to the point where it can be used in SSI because of its low latency and ease of implementation.

The core idea of SSI is recording the articulation organs, which are used in human speech production. In the area of AAM, several different types of articulatory acquisition methods have been used. In comparison to other techniques (e.g., PMA, EMA, X-ray, XRMB, and vocal tract MRI [13]–[16]), ultrasound tongue imaging (UTI) has gained popularity because it is a clinically safe and non-invasive method for tracking tongue movement and provides us with a clearly visible tongue surface. The analysis of ultrasound tongue images has illustrated that useful, reliable information about several parameters of the tongue gesture can be seen in the data that is extracted from UTI [17]–[19].

Generally, collecting articulatory data necessitates more effort than audio speech data [20]. Although significant developments have been made in SSI, due to additional data collection difficulties, the vast majority of related studies have been based on relatively small datasets compared to acoustic data sets for acoustic speech recognition.

Deep learning (for example, convolutional neural networks,



(a) Original data

(b) Consecutive Time Masking (CTM)

(c) Intermittent Time Masking (ITM)

(d) Sinusoidal Noise Injection (SNI)

(e) Random Scaling (RS)

Fig. 1: Examples of UTI example sample's ultrasound pixels over time before and after data augmentations. (Details: x-axis: time (12ms/frame) ; y-axis: pixel intensity)

or CNNs) has been successfully applied to SSI using UTI; however, deep learning models tend to overfit easily and require large dataset. Therefore, data augmentation plays a crucial role in SSI with UTI by providing additional training examples to prevent overfitting and enhance the performance of deep learning models, given the limited amount of available data. Data augmentation has been proposed as a method to generate additional training data for end-to-end SSR on EMA datasets: Cao and his colleagues applied data augmentation strategies to raw kinematic signals [21].

As a result of the recent effectiveness of augmentation, in this paper, we implemented data augmentation techniques on our dataset, which consists of ultrasound tongue images. In this study, we investigated several approaches for AAM towards SSI. These data augmentation approaches (consecutive time masking, intermittent time masking, sinusoidal noise injection, and random scaling) were directly applied to the ultrasound tongue images. The methods have been compared based on mean-squared error validation loss, which is used as an evaluation metric.

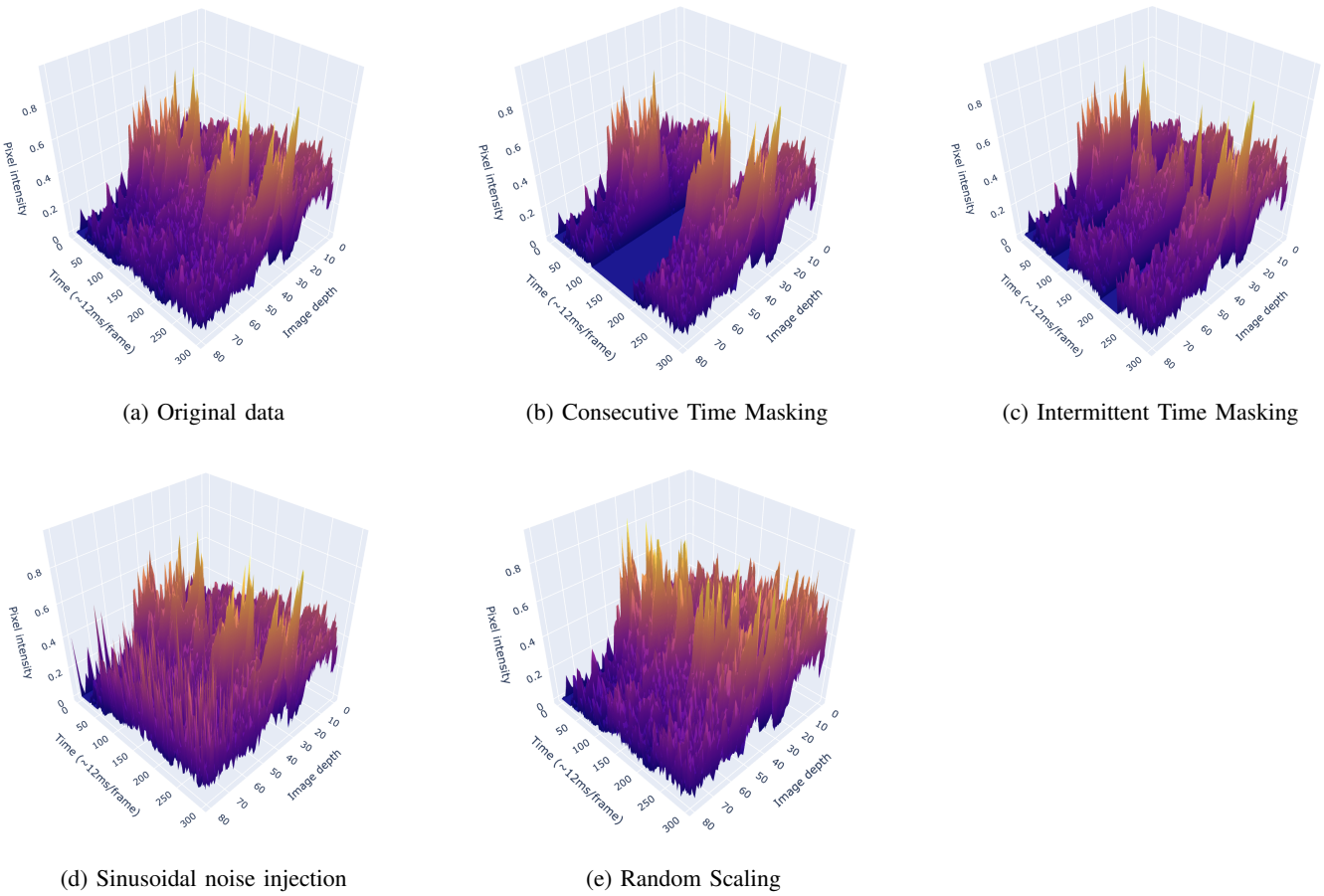
II. Dataset

In articulatory data acquisition, one Azerbaijani male subject (the first author of this study) was recorded while reading sentences aloud (154 sentences). 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 Beyer-dynamic 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 (c.f. [22, Fig. 2] in the current WINS proceedings), after being resized to 64×128 pixels using bicubic interpolation. More details about the recording set-up and articulatory data can be found in [23]. The total duration of the recordings was about 15 minutes, which was partitioned into training, validation and test sets in a 80-10-10 ratio.

III. Methods

A. *WaveGlow neural vocoder (baseline)*

In the baseline vocoder, during analysis, the mel-spectrogram was estimated from the Azerbaijani speech recordings (digitized at 22 kHz). Similarly to the original WaveGlow paper [24], 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



(a) Original data

(b) Consecutive Time Masking

(c) Intermittent Time Masking

(d) Sinusoidal noise injection

(e) Random Scaling

Fig. 2: Examples of UTI example sample's 3D visualisation of ultrasound pixels over time before and after data augmentations.

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 [24].

B. Data augmentation for articulatory-to-acoustic mapping

Deep learning models have been successfully applied to articulatory-to-acoustic mapping for silent speech interfaces. However, improving the generalization ability of these models is one of the essential challenges to be handled. Generalizability refers to the performance difference of a model between training data and testing data that has never been seen before by a model. With poor generalizability, it has been shown that models have overfitted the training data. It is important to build a model in which the validation error decreases with the training error.

Data augmentation methods are used for improving baseline model's robustness, by representing more comprehensive set of possible data points. As a result, augmentation techniques seek to reduce and minimize the distance between the training and validation sets to the greatest extent possible.

Data augmentation can be thought of as a way for computer algorithms to simulate the process of imagination, similar to how humans use their imaginations to understand and interpret the world around them. As a result, these algorithms are able to more accurately identify patterns and make predictions based on the augmented dataset. Techniques (such as consecutive time masking, intermittent time masking, sinusoidal noise injection, and random scaling) create variations of ultrasound tongue images based on their existing knowledge, thereby improving their understanding of the UTI data.

The neural network is exposed to a new set of variations of the input image on each epoch by applying different modifications to the image, which enriches the learning process. 2D ultrasound tongue images samples from various speakers are shown in [22, Fig. 1], in the current WINS proceedings.

From the 2D(space) + 1D(time) ultrasound data, several pixels were chosen and plotted as a function of time. Examples of transformations are illustrated in Fig. 1 where an original sample's ultrasound pixels over time is also demonstrated. For clarity of visualization, the [0:300] time range is chosen. In our dataset, per frame, the duration of the sample is approximately 12 ms.

As ultrasound tongue images are not simple two-dimensional data, we also provide the illustration of a data sample's ultrasound pixels over time as a 3D figure for each of the original (Fig. 2a) and transformed versions to visualize the differences between before and after applying data augmentation approaches.

1) Consecutive time masking (CTM): The consecutive time masking (CTM) data augmentation technique used in this study aims to increase the robustness of a model. In this approach (Fig. 1b), certain number of consecutive frames are selected from an UTI sample and set to zero. The idea behind this is that by masking a portion of the sample, the model is forced to learn more robust features and to rely less on specific frames of the data. This technique is similar to SpecAugment [25], it is a time-domain data augmentation technique that was designed to improve the robustness of automatic speech recognition models. The main difference is that, in SpecAugment, the masking is applied in the frequency domain instead of the time domain (Fig. 2b). The starting point for the masking process is manually chosen to be within one-third to two-thirds of the way into the sample; this is to ensure that the masking process is applied in the same parts of the arrays and not in a random manner.

2) Intermittent time masking (ITM): In addition to the CTM approach, we also investigated a technique called "intermittent time masking" (ITM). This method (Fig. 1c) involves masking small segments of the data rather than a continuous block. The process is illustrated in Figure 1c. In this approach, a fixed number of starting points were manually chosen from specific portions of the pixels (specifically $[\frac{1}{6} : \frac{1}{3}], [(\frac{1}{3} + \frac{1}{6}) : (2 \cdot \frac{1}{3})], [(2 \cdot \frac{1}{3} + \frac{1}{6}) : 1]$), then a fixed number of frames were masked out from each of these starting points. The goal of this technique is to expose the model to a variety of masked segments, rather than a single, continuous block of masked frames. This way, the model is exposed to different variations and can learn to generalize better. In the 3D figure of this method (Fig. 2c), we can see the demonstrated segments that are masked out.

3) Sinusoidal noise injection (SNI): The sinusoidal noise injection (SNI) method is particularly useful for tasks that involve signals with cyclic patterns, such as speech recognition and audio processing, as it can help the network better understand the underlying patterns in the data [26]. SNI involves adding noise in the form of sinusoidal waves to the input data. In this approach (Fig. 1d), this is done by applying a sinusoidal function to the pixel values of ultrasound images. The amplitude of the sinusoidal waves was determined by taking the average amplitude of that specific dimension and then multiplying it by a coefficient (scaling factor) to increase or decrease the amplitude. The frequency of the sinusoidal waves per unit of time was a predetermined constant that was identified during the initial analysis. The amplitude scaling factor was set at 0.02, the number of oscillations per second (Hz) of the noise was 40, and the phase was set to zero. For example, if the mean amplitude of an articulatory dimension was A, the sinusoidal noise that was added would be represented by $I(t) = 0.02 \cdot A \cdot sin(2\pi \cdot 40t)$ which is a product of the mean amplitude, the scaling factor, and the sine function with frequency (40). The difference in the three-dimensional figure after using the sinusoidal noise injection technique on our ultrasound tongue image is clearly visible in Fig. 2d. The value of the ultrasound pixels has increased after adding sinusoidal waves to the original dimension, as shown in the figure.

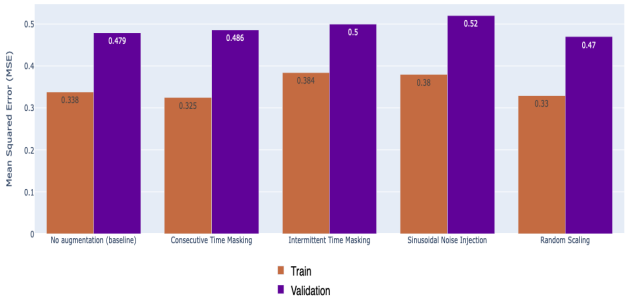


Fig. 3: Mean Squared and Validation Mean Squared Errors of UTI example sample before and after data augmentation methods.

4) Random scaling (RS): A recent study [21], has shown the implementation of random scaling over EMA signals. With this in mind, we decided to explore a data augmentation strategy that involved altering the duration of the samples by randomly stretching or shrinking them on our ultrasound tongue image dataset (Fig. 1e). To achieve this, we selected a scaling factor from a range of numbers. Through a series of preliminary experiments, we found that a range of 0.8 to 1.4 produced the best results. We applied this technique to all samples. This approach not only improves the robustness of the model but also emulates the variability found in real-world scenarios (Fig. 2e).

IV. Results and Discussion

Based on the results shown in the bar chart (Fig. 3), it is clear that the data augmentation techniques we applied had a slight impact on the performance of our model. On our dataset, we used mean squared error as an evaluation metric to compare each augmentation method. The methods are compared using two errors that are tested in the training and validation datasets.

Without data augmentation, the baseline model has given results of 0.338 and 0.479 as mean squared error (MSE) and validation mean squared error (V-MSE), respectively. The initial goal of applying these augmentation methods was to minimize the error rate as much as possible; however, the results on our Azerbaijani dataset did not satisfy our target. (These numbers have been taken after implementing 15 epochs over the dataset.) It is worth noting that after applying CTM to our dataset, the smallest but most remarkable reduction in mean squared error by 0.013 was observed among all

techniques, however, no improvement in the validation set was observed after applying consecutive time masking. According to both error rates, applying random scaling to the dataset could help us achieve minimizations on error rates (0.33 MSE and 0.47 V-MSE).

Although the data augmentation techniques that have been proposed in this paper are powerful on an audio signal, EMA signal, applying them to ultrasound tongue images is not an easy task as ultrasound data contains more data and in a slightly different format in itself. Using augmentation techniques and observing ultrasound pixel changes over time is a must-do to be sure if the method could reach the target of our initials. The observed results and changes in this paper demonstrated that dealing with ultrasound tongue images in the sense of augmenting them to obtain additional data to overcome overfitting to our dataset is only marginally beneficial.

V. Conclusions and Future work

This study examined various methods of enhancing data for articulatory-to-acoustic mapping towards silent speech interfaces through data augmentation. The results of this initial feasibility study showed that random scaling (RS) was the most effective approach, leading to greater improvements than other methods; but clearly, mode advanced analysis of the results will be necessary. Additionally, it was found that applying CTM ultrasound data resulted in a lower MSE than other techniques. We plan to apply these methods to a larger dataset or on other modalities (e.g., vocal tract MRI [27]), as well as observe these method combinations separately on a sample. Further experiments and studies are needed to verify these results – but we are happy to get inspiration from colleagues of speech technology / linguistics / neuroscience at the WINS 2023 workshop.

Acknowledgment

The research was partially funded by the National Research, Development and Innovation Office of Hungary (FK 142163 grant). T.G. Cs.'s research was supported by the Bolyai Research Fellowship of the Hungarian Academy of Sciences, and by the ÚNKP-22-5-BME-316 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. G. G.'s research was supported by the NRDI Office of the Hungarian Ministry of Innovation and Technology (grant no. TKP2021-NVA-09), and within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004). The Titan X GPU used was donated by NVIDIA. We would like to thank the ex-MTA-ELTE Lingual Articulation Research Group for providing the equipment necessary for the articulatory recordings and for Beiming Cao and his colleagues at the University of Texas at Austin, USA for the excellent research idea [21].

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] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical engineering & physics*, vol. 30, no. 4, pp. 419–425, 2008.

[3] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 12, pp. 2386–2398, 2017.

[4] 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, 2017.

[5] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, 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.

[6] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2971–2975.

[7] L. Tóth, G. Gosztolya, T. Grósz, A. Markó, and T. G. Csapó, "Multi-task learning of speech recognition and speech synthesis parameters for ultrasound-based silent speech interfaces." in *INTERSPEECH*, 2018, pp. 3172–3176.

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

[9] 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 *ISCA Interspeech 2016*, vol. 2016, 2016, pp. 1467–1471.

[10] N. Shah, N. J. Shah, and H. A. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion." in *INTERSPEECH*, 2018, pp. 3157–3161.

[11] A. Ephrat and S. Peleg, "Vid2speech: speech reconstruction from silent video," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5095–5099.

[12] L. Diener, G. Felsch, M. Angrick, and T. Schultz, "Session-independent array-based emg-to-speech conversion using convolutional neural networks," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[13] B. Cao, N. Sebkhi, T. Mau, O. T. Inan, and J. Wang, "Permanent magnetic articulograph (pma) vs electromagnetic articulograph (ema) in articulation-to-speech synthesis for silent speech interface," in *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, 2019, pp. 17–23.

[14] P. Badin, G. Bailly, L. Reveret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[15] 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, 2017.

[16] 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, 2017.

[17] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 455–501, 2005.

[18] M. Stone, B. C. Sonies, T. H. Shawker, G. Weiss, and L. Nadel, "Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system," *Journal of Phonetics*, vol. 11, no. 3, pp. 207–218, 1983.

[19] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone, "Prospects for a silent speech interface using ultrasound imaging," in *2006 IEEE International*

Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1. IEEE, 2006, pp. I–I.

[20] K. Richmond, Z. Ling, and J. Yamagishi, "The use of articulatory movement data in speech synthesis applications: An overview—application of articulatory movements using machine learning algorithms—," *Acoustical Science and Technology*, vol. 36, no. 6, pp. 467–477, 2015.

[21] B. Cao, K. Teplansky, N. Sebkhi, A. Bhavsar, O. T. Inan, R. Samlan, T. Mau, and J. Wang, "Data augmentation for end-to-end silent speech recognition for laryngectomees," in *Proceedings of the Interspeech*, 2022, pp. 3653–3657.

[22] T. G. Csapó, "Is Dynamic Time Warping of Ultrasound Tongue Images Suitable for Articulatory Signal Comparison?" in *WINS 2023*, 2023.

[23] 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 2017*, 2017, pp. 3672–3676.

[24] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[26] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[27] T. G. Csapó, "Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 2722–2726.