Towards Cross-Speaker Articulation-to-Speech Synthesis using Dynamic Time Warping Alignment on Speech Signals

Ibrahim Ibrahimov
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
ibrahim@tmit.bme.hu

Gábor Gosztolya
HUN-REN–SZTE Research Group on Artificial Intelligence and University of Szeged, Insitute of Informatics
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) aim to provide a non-intrusive means of communication by decoding articulatory information directly from the speaker's silent gestures, such as tongue movements. However, existing SSI methods often face challenges related to speaker dependency, arising from the substantial variations in individual articulatory organ structures and speeds. This paper explores the integration of Dynamic Time Warping (DTW) alignment in the context of cross-speaker articulation-to-speech synthesis. The DTW is performed on the speech signals, which is in synchrony with the ultrasound tongue images (UTI). The alignment of UTI is done based on the calculated DTW distance. We tested cross-speaker articulation-to-speech synthesis with 4 subjects from the UltraSuite-TaL dataset. Through the utilization of aligned ultrasound data, we trained convolutional neural networks to predict mel-spectrogram from the UTI input, and finally synthesized speech with each speaker pair. The results underline the potential of DTW as a valuable tool in enhancing the applicability of SSI.**

***Index Terms*—dynamic time warping, ultrasound tongue imaging, silent speech interfaces**

I. Introduction

Speech technology research, i.e., the analysis, synthesis, and processing of speech and other speech-related signals can have applications in the field of human-computer interaction. Silent Speech Interfaces (SSI) aim to provide a non-intrusive means of communication by decoding articulatory information directly from the speaker's silent gestures, such as tongue movements [1]. In general, digital applications using speech technology could significantly help the everyday communication of people with speech or vision impairments. As a specific example, individuals may experience speech impairments due to damage to the organs responsible for articulation, such as the larynx (voice box) which can prevent them from producing audible speech; and Silent Speech Interfaces can be beneficial in this case. Besides, SSI might be useful in military applications or extremely noisy conditions.

A. Articulation-to-Speech synthesis

The methodology behind SSI is often referred to as articulatory-to-acoustic mapping (AAM), or articulation-to-speech synthesis (ATS). These focus on converting articulatory biosignal information into audible speech, nowadays mostly using deep neural networks (DNNs) [2]–[6]. Within the domain of Silent Speech Interfaces, the ATS or direct synthesis process stands out as a prominent approach for speech synthesis from articulatory data. In contrast to the sequential framework involving silent speech recognition followed by text-to-speech (TTS) procedures, ATS employs vocoders to directly transform articulatory data into speech signals. These vocoders utilize DNNs to predict spectral parameters based on the input articulatory information, generating synthesized speech [2]. While the direct synthesis approach encounters challenges compared to SSR+TTS technique due to the absence of textual information, recent advancements have markedly elevated the quality of speech output in articulatory-to-acoustic mapping [6]. This progress positions ATS as a compelling option within SSI applications, emphasized by its advantageous characteristics, including low latency and straightforward implementation.

Various techniques have been used to collect articulatory data as input for ATS, including electromagnetic articulography (EMA), surface electromyography (sEMG), permanent magnetic articulography (PMA), ultrasound tongue imaging (UTI) [7]. In case of UTI, most often the mid-sagittal orientation is recorded (for examples, see Fig. 1). The advantage of ultrasound over other articulatory recording techniques is that it is easy to use, non-invasive, affordable, and can be used to record at a relatively high resolution (up to 800 x 600 pixels) and high speed (up to 100–150 frames per second) [8], [9].

https://doi.org/10.3311/WINS2024-002

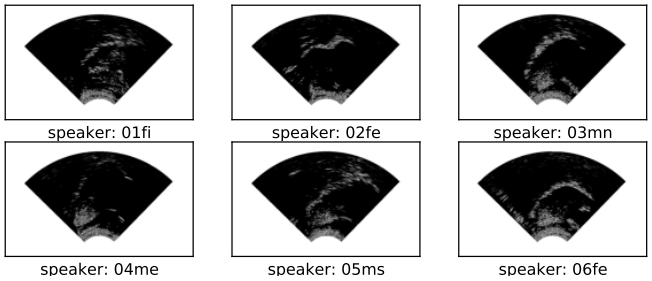

Fig. 1: Examples for the differences in image position and image quality across speakers of the UltraSuite-TaL80 database.

B. Session and speaker dependence of ultrasound

Examining cross-session and cross-speaker aspects became trivial for most speech technologies, but unfortunately, this is not the case for other biosignals like ultrasound tongue imaging. Standard methods developed for the speech signal typically cannot be used directly with articulatory data.

The image quality of the ultrasound tongue images may vary between sessions and speakers. By session, we mean that the probe-fixing headset is dismounted and mounted again onto the speaker [4]. The quality of the images is influenced by many factors, such as the anatomy of the speaker or the condition of the tissues of the articulatory organs. The variation between speakers may also be due to the fact that the ultrasound transducer is positioned differently (in different orientations) for different head sizes and shapes [10]. Due to this, the ultrasound probe cannot be positioned identically for different speakers, so the possibility of comparing speakers is limited because of the potentially unaligned orientations.

C. Dynamic time warping and biosignals

Dynamic Time Warping (DTW) is a long-established method for comparing speech samples of different lengths [11], and recently, it also has been used in the context of articulatory data [12]–[15]. DTW has been successfully applied earlier to 1) UTI and intra-speaker comparisons [12], 2) EMA for analyzing inter-speaker differences [13], 3) EMA and ECoG, also for inter-speaker comparisons [14]. All these experiments are summarized in our previous study [15]. However, there has been no detailed research on the application of DTW for cross-speaker ultrasound tongue image analysis. Ealier we performed an initial study that proposes the idea of such a DTW-based comparison of UTI, but only includes demonstration samples without real experiments or any kind of objective measurements [15].

D. Goal of the current study

In this research, we aim to produce a time alignment for a pair of UTI recordings, coming from different speakers. To achieve this goal, we will obtain the alignment on the basis of the parallel-recorded speech utterances, and verify the quality of the alignment for the UTI data, using articulation-to-speech experiments.

II. Data

We used the UltraSuite-TaL80 database [16], downloaded from *https://ultrasuite.github.io/data/tal_corpus/*. In this corpus the midsagittal movement of the tongue was recorded using the 'Micro' system (AAA software, Articulate Instruments Ltd.) with a 64-element, 20-mm radius convex ultrasound transducer at around 81.5 fps, with the help of a probe fixing metal headset. Speech was digitized at a sampling rate of 48 kHz using a Sennheiser HKH 800 p48 microphone. Synchronization of ultrasound data and speech signals was performed using a tool provided by Articulate Instruments Ltd. Lip video was also recorded, but this information was not used in the current study.

We used the data of the first four speakers (i.e. '01fi', '02fe', '03mn' and '04mn'; 2 females and 2 males). As for our alignment experiments we had to use data with the same spoken content, we restricted our experiments to the recordings with the 'shared audible read speech utterances' prompt ('xaud'); there were 24 such sentences for each speaker.

An example for the cross-speaker differences in the ultrasound tongue image recordings is shown in Fig. 1, displaying ultrasound images of 6 speakers. It can clearly be seen that ultrasound can visualize different sections of the tongue (e.g. '02fe' has a shorter, while '06fe' has a longer tongue), and also different visibility of the tongue contour (e.g. '01fi' has a blurred image, but '02fe' has a clear upper surface of the tongue).

The ultrasound tongue images were stored as 8-bit grayscale pixels in the raw ultrasound form of the "Micro" system (64×842 resolution). The audio was resampled to 22 050 Hz.

III. Cross-speaker articulatory alignment and articulation-to-speech experiments

To achieve cross-speaker articulatory alignment, we employ the DTW alignment method on distinct speakers' speech data in a cross-pairing arrangement. After that, we conduct cross-speaker articulation-to-speech experiments.

A. Time-alignment of Ultrasound Tongue Images, using Dynamic Time Warping of speech signals

Here, our aim is to obtain a time-alignment of two UTI recordings, with the same phonetic content, as proposed in [15]. Such alignment methods like Dynamic Time Warping require dividing the two sequences into fixed-size units, and defining a distance measure between these units. Although in the case of ultrasound tongue imaging, a division into small elements straightforwardly exists (as the recordings are sequences of images with the same resolution), an appropriate distance function is hard to find. In contrast, we would like to apply a method which focuses on the spoken content, and is insensitive to different sessions or speakers.

Due to this, we decided to obtain the time alignment of the two recordings based on the audio, since the UTI and speech signals are synchronized. For this, we calculated 80 MFCCs with a frame step in synch with the fps of the ultrasound videos (12 ms roughly equals 81.5 fps). Therefore, the best path

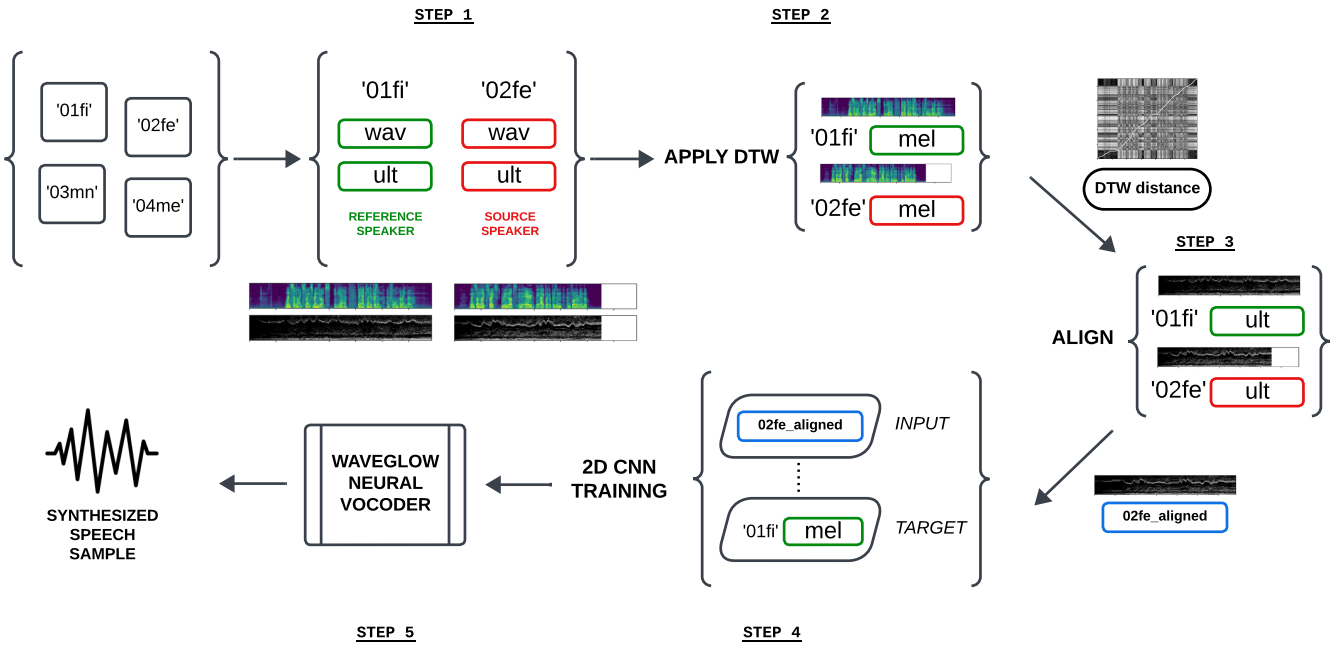

Fig. 2: Pipeline for aligning ultrasound tongue images (UTI) across different speakers using Dynamic Time Warping (DTW) for articulation-to-speech synthesis: Step 1 - pairing reference and source speakers' audio and UTI files; Step 2 - Applying DTW on the mel-spectrograms of audio files; Step 3 - Utilizing calculated DTW distance for aligning source speaker's ULT with reference speaker's ULT; Step 4 - Cross-speaker training; Step 5 - Cross-speaker speech synthesis.

returned by the DTW method over the MFCC frames matches that of the ultrasound images. Examples for the aligned speech spectrograms and aligned ultrasound sequences can be found in [15].

Next, we select the first four speakers' data from the Ultrasuite-Tal80 dataset, corresponding to "01fi", "02fe", "03mn", and "04me" for this experiment, as shown in Figure 2. By employing the speakers one by one as references for the remaining three source speakers, we obtain DTW-aligned versions of source speakers to the reference speaker in each step of the four-step process (as we have four speakers).

We also use a visualization technique called 'kymogram', which is a kind of 'articulatory signal over time': the middle slices (midline) of the ultrasound tongue images (roughly corresponding to the middle of the tongue) are plotted as a function of time, similarly to a spectrogram [17]. In Figure 3, the ultrasound tongue images acquired after DTW-alignment are presented in their kymogram representations (2nd-4th rows) alongside with the original ultrasound tongue image of reference speakers (1st row) for the '006' speech sample. In the rotating structure, each reference speaker serves as a source for the remaining speakers, and vice versa, as depicted in the provided figure.

Upon employing the alignment process and compiling the aligned source ultrasound data that corresponds to the reference audio file, we could create new aligned datasets of 24 sentences. In order to check the usefulness of these, we conducted articulation-to-speech experiments, which will be detailed in the next subsection.

B. Cross-speaker articulatory-to-acoustic mapping: Training and Synthesis

During the experimental phase, we employ 2D Convolutional Neural Networks (CNN) for articulatory-to-acoustic mapping, similarly to [6]. As illustrated in Figure 2, the deep neural network is trained on ultrasound tongue images, with each step involving the ingestion of aligned ultrasound images from source speakers as input. The objective is to predict the audio mel-spectrogram of the reference speaker. The structure of the CNN is detailed in [6].

The dataset used to train the network comprises a modest 24 sentences. While past experiments in the field commonly used UltraSuite-TaL dataset with its full availability for each speaker (almost 200 sentences per speaker), our study faced a challenge with only 24 sentences, which is notably low for training CNN. This limited data reflects the available resources for shared sentence recordings, making this research a feasibility study under these constraints.

To create a training and test set, three predefined speech samples ("006", "015", and "021") are intentionally reserved for testing, while the remaining 21 samples constitute the training set. The selection of test samples is deliberate and focuses on varied structural characteristics.: "015" - short ("Other men have tried to explain the phenomenon physically."), "006" - mid-length ("These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon."), and "021" - long sentence ("If the red of the second bow falls upon the green of the first, the result is to

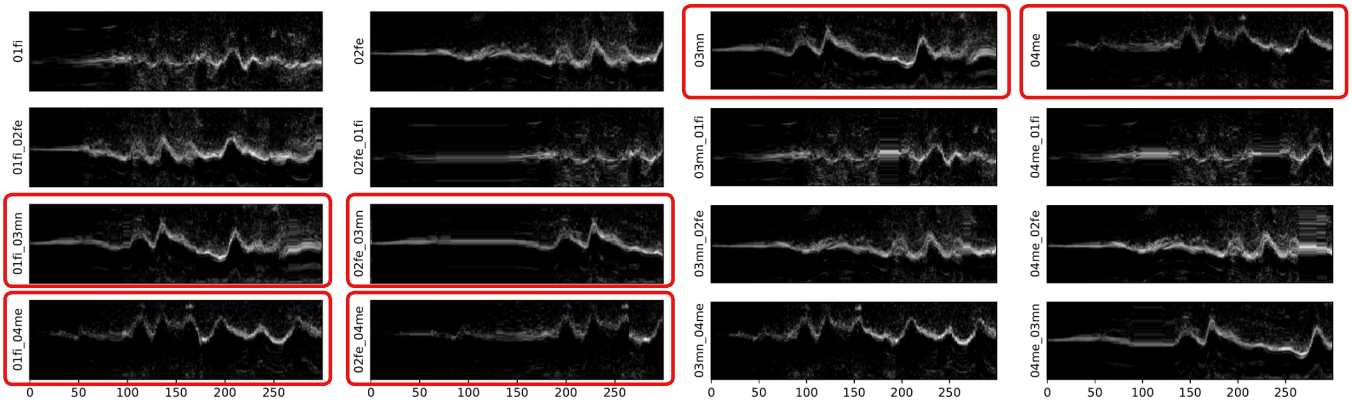

Fig. 3: Kymogram representation of cross-speaker DTW alignment for corresponding ”006” sentence ultrasound tongue images. 1st row: original data; 2nd-4th rows: DTW-aligned data. (Label structure: [reference speaker]_[source speaker])

give a bow with an abnormally wide yellow band, since red and green light when mixed form yellow.”). This thoughtful selection aims to provide a nuanced evaluation of the system post-training. Given the relatively small size of the dataset, a non-randomized approach to selecting the test set enhances result comparability.

After training the ultrasound-to-mel-spectrogram network, speech synthesis from mel-spectrogram is conducted using the WaveGlow neural vocoder [18]. WaveGlow takes the mel-spectrogram representation as input and produces synthesized speech samples. The mel-spectrogram used for synthesis is obtained from training on aligned ultrasound images of source speakers, aligning with the reference speaker's mel-spectrogram. Thus, the study adopts a cross-speaker articulation-to-speech synthesis approach. For a detailed exploration of the neural vocoder structure utilized in this study, please refer to Section 3.3 in [19].

C. Cross-speaker articulatory-to-acoustic mapping: Results

To assess the efficacy of the cross-speaker articulation-to-speech synthesis process, we measure Mean Squared Error (MSE) values during the training phase and Mel-Cepstral Distortion (MCD) values for the synthesis stage. These metrics are systematically computed on the designated test set which is outlined in the previous section.

To establish a benchmark for comparison, we utilize the original ultrasound tongue images along with their corresponding mel-spectrogram representations as a baseline method. This synthesis process adheres to the same train-test ratio as employed for the proposed model. By computing the mean values of these metrics across the test set, comprising three representative samples, we gain a comprehensive insight into the overall performance of the proposed system.

In Table I, we present the mean MSE values for the four speakers on obtained mel-spectrograms from the test set after training involved in this study. These values are collected after excluding epochs beyond the patience threshold, set to 3 for early stopping during training. The limited size of our dataset is evident in the fact that almost half of the Mean MSE results surpass 0.6. This highlights the difficulties arising from the scarcity of data. In comparison, the previous study, which utilized all the available dataset, reported Validation MSE values consistently below 0.3 [19].

TABLE I: Mean MSE values on the test set for each speaker and speaker pairs.

		Reference speaker			
		01fi	02fe	03mn	04me
	Baseline	0.531	0.547	**0.388**	**0.479**
Source spk.	01fi	—	0.655	0.787	0.716
	02fe	0.611	—	0.767	0.754
	03mn	**0.511**	**0.528**	—	0.713
	04me	**0.502**	**0.485**	0.515	—

Upon examining Table I, intriguing deviations from this trend are observed, particularly when speakers ”01fi” and ”02fe” serve as reference speakers for source speakers ”03mn” and ”04me”. For instance, in the case of speaker ”01fi” as the reference, the baseline's mean MSE results are approximately 0.02 and 0.03 higher than when ”03mn” and ”04me” act as source speakers, respectively. Similarly, for speaker ”02fe” as the reference, the baseline yields mean MSE values 0.02 and 0.06 higher than those obtained with ”03mn” and ”04me” as source speakers, respectively. Opposite to these cases, a noteworthy trend emerges when speakers ”03mn” and ”04me” are the reference speakers, with the baseline consistently exhibiting superior MSE results.

Due to its widespread application in the field of speech synthesis, MCD serves as the chosen metric for evaluation of the synthesized samples of the articulation-to-speech synthesis. The calculation of MCD values between synthesized speech samples and the original audio files is facilitated using the *pymcd* package in Python, initially introduced in [20]. Among the available options in the package, the ”plain” mode is employed for its simplicity, by taking into account that DTW alignment is already done in our proposed structure.

TABLE II: Mean MCD values on the test set for each speaker and speaker pairs.

		Reference speaker			
		01fi	02fe	03mn	04me
	Baseline	10.934	11.316	**10.267**	**9.881**
Source spk.	01fi	—	12.726	14.726	12.466
	02fe	11.791	—	14.632	13.304
	03mn	**10.455**	**10.960**	—	12.240
	04me	**9.948**	**10.538**	11.196	—

In Table II, the mean MCD values for test speech samples across different speakers are presented. The observed trends in MCD values align with those identified in the mean MSE results in Table I. Particularly, when speakers "01fi" and "02fe" are designated as reference speakers for "03mn" and "04me" source speakers, the MCD results exhibit superior performance compared to the baseline. Conversely, when "03mn" and "04me" serve as reference speakers, the baseline yields synthesized samples with better mean MCD results.

The observed trends in both evaluation metrics underlines the importance of a thorough investigation to pinpoint the factors contributing to instances where the proposed system yielded better results compared to the baseline. This analysis is further detailed in the subsequent section.

IV. Discussion

The results of the cross-speaker articulatory-to-acoustic experiments in the previous section highlight that certain pairs of speakers perform better than the baseline system. To understand why, we decided to investigate the alignment of ultrasound tongue images. Figure 3 shows kymograms of the speakers' "006" sentence ultrasound images, until the 300th frame. We focused on specific pairs [01fi-03mn], [01fi-04me], [02fe-03mn], [02fe-04me] (marked in red shapes) that demonstrated better results in terms of validation MSE (Table I) and MCD (Table II).

Despite the challenges in ultrasound images, such as gray scale representations and potential artifacts, a visual inspection indicates noticeable differences. Taking speaker "01fi" as an example, the original kymogram has a noisy area between time frames [100-150]. This noise increases when "02fe" is aligned with "01fi." However, in the aligned kymograms of [01fi-03mn] and [01fi-04me], the same time frame range appears clearer than in the original version. This suggests that the alignment process has improved the clarity of the tongue trajectory (brighter areas in the kymograms) for these specific pairs, providing a potential explanation for the unexpected improvement in results.

Moreover, anatomical differences of the speakers' articulators (see Section I-B) can cause that the tongue images vary in quality, as illustrated in Figure 1. Specifically, when examining the case of speaker "03mn," the wedge-shaped representation of the ultrasound tongue image in Fig. 1 highlights a notably clearer and brighter tongue contour. This consistency is reflected in Figure 3, particularly in the third column, where fewer artifacts are noticeable in the original kymogram of speaker "03mn" as compared to those of other speakers.

V. Conclusions

This paper explored the integration of speech-based DTW alignment in the context of cross-speaker articulation-to-speech synthesis. The alignment of UTI was done based on the calculated DTW distance. We tested cross-speaker synthesis with 4 subjects from the UltraSuite-TaL dataset. Using aligned ultrasound data, we did articulation-to-speech experiments, synthesized speech outputs with each speaker pair, and compared the MSE and MCD errors. We found that for 2 speakers, the DTW-aligned input ultrasound images resulted in lower errors than using the original ultrasound data of the same speaker, indicating that anatomical differences or speckle noise in ultrasound images are an important factor. Overall, the results underlined the potential of DTW as a valuable tool in enhancing the applicability of SSI.

Continuing this study, a thorough investigation of the proposed structure becomes imperative. While kymograms offer a potential explanation for the observed results, a more in-depth and objective evaluation is necessary to enhance the accuracy of our findings. Broadening the pool of speakers employed for cross-speaker alignment allows us to assess both the objective and, on a larger scale, the subjective aspects of the results. In order to apply these findings to a more extensive dataset, future efforts could involve employing cross-validation techniques. Furthermore, we plan to integrate the proposed system into the data augmentation pipeline, building upon our previous work [19]. Additionally, we aim to explore the feasibility of training all aligned articulatory data, alongside the original data, to develop a speaker-independent articulation-to-speech synthesis, which will be the focus of our future work.

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-23-5-BME-440 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.

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. 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.

[3] 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.

[4] 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 Polytechnica Hungarica*, vol. 17, no. 7, pp. 109–124, 2020.

[5] B. Cao, A. Wisler, and J. Wang, "Speaker Adaptation on Articulation and Acoustics for Articulation-to-Speech Synthesis," *Sensors*, vol. 22, no. 16, p. 6056, 2022.

[6] 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.

[7] M. Stone, B. Sonies, T. 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, pp. 207–218, 1983.

[8] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical Linguistics and Phonetics*, vol. 19, no. 6-7, pp. 455–501, jan 2005.

[9] D. H. Whalen, J. Kang, R. Iwasaki, G. Shejaeya, B. Kim, K. D. Roon, K. Mark, Tiede, J. Preston, E. Phillips, T. McAllister, and S. Boyce, "Accuracy assessments of hand and automatic measurements of ultrasound images of the tongue," in *Proc. ICPhS*, Canberra, Australia, 2019, pp. 542–546.

[10] L. Spreafico, M. Pucher, and A. Matosova, "UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 1517–1520.

[11] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[12] C. Yang and M. Stone, "Dynamic programming method for temporal registration of three-dimensional tongue surface motion from multiple utterances," *Speech Communication*, vol. 38, no. 1-2, pp. 201–209, sep 2002.

[13] S. M. Jayanthi, L. Menard, and C. Laporte, "Divide-and-warp temporal alignment of speech signals between speakers: Validation using articulatory data," in *Proc. ICASSP*, New Orleans, LA, USA, 2017, pp. 5465–5469.

[14] G. Le Godais, "Decoding speech from brain activity using linear methods," Ph.D. dissertation, Université Grenoble Alpes, 2022.

[15] T. G. Csapó, "Is Dynamic Time Warping of speech signals suitable for articulatory signal comparison using ultrasound tongue images?" in *Workshop on Intelligent Infocommunication Networks, Systems and Services (WINS 2023)*, 2023.

[16] M. S. Ribeiro, J. Sanger, J.-X. Zhang, A. Eshky, A. Wrench, K. Richmond, and S. Renals, "Tal: A synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 1109–1116.

[17] S. M. Lulich, K. H. Berkson, and K. de Jong, "Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion," *Journal of Phonetics*, vol. 71, pp. 410–424, 2018.

[18] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," 2018.

[19] I. Ibrahimov, G. Gosztolya, and T. G. Csapo, "Data Augmentation Methods on Ultrasound Tongue Images for Articulation-to-Speech Synthesis," in *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 230–235.

[20] Q. Chen, Y. Li, Y. Qi, J. Zhou, M. Tan, and Q. Wu, "V2c: Visual voice cloning," 2021.