Regeneration of Ultrasound Tongue Images using Tongue Position Values towards Articulatory-to-Acoustic Mapping

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

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

Csaba Zainkó
Department of Telecommunications and Artificial Intelligence
Budapest University of Technology and Economics
Budapest, Hungary
zainko@tmit.bme.hu

***Abstract*—Silent Speech Interfaces (SSI) facilitate communication for individuals unable to speak aloud by leveraging articulatory data. Ultrasound Tongue Imaging (UTI) is a widely adopted, non-invasive method for capturing tongue movements in real time, making it valuable for Articulation-to-Speech (ATS) synthesis. However, challenges such as probe misalignment, incomplete tongue images, and inter-speaker variability can affect data quality. To address these issues, recent advancements in deep learning, including neural vocoders and Tacotron2-based models, have improved ATS synthesis by enhancing naturalness and intelligibility. This study explores an alternative approach by regenerating ultrasound tongue images from tongue position values using DeepLabCut, aiming to refine articulatory-to-acoustic mapping (AAM). Two different regeneration methods are investigated, and their effectiveness is assessed through visual and objective evaluations, including Structural SIMilarity (SSIM) and Mean Squared Error (MSE) metrics. While the regenerated images exhibit high SSIM scores, discrepancies in articulatory detail affect AAM performance, highlighting the need for further refinement. These findings contribute to the development of more robust UTI-based speech synthesis systems, with potential applications in assistive communication and articulatory training.**

***Index Terms*—regeneration, ultrasound tongue imaging, DeepLabCut, articulatory-to-acoustic mapping**

I. Introduction

Speech communication involves different parts of human body to co-operate together to produce an audible output. A Silent Speech Interface (SSI) is a system that generates audible speech by analyzing a person's articulatory movements—such as tongue and lip motions—without relying on actual vocal sounds. The main purpose of this system is to serve as an aid to a user who has either lost their ability to speak aloud or is in a place that is very noisy or silence should be provided.

Articulators such as the tongue, lips, and nose play a crucial role in speech production by shaping airflow and sound vibrations to create intelligible speech. Among these, the positioning of the tongue is particularly important, as it significantly influences the articulation of both vowels and consonants [1]. There are different acquisition techniques used in research for obtaining articulation data. Towards articulation-to-speech (ATS) reconstruction systems, electromagnetic articulography (EMA), permanent magnetic articulography (PMA), real-time magnetic resonance imaging (rt-MRI), ultrasound tongue imaging (UTI) and etc. are used by researchers [2]–[4].

Real time application of SSI systems is one of the main obstacles to overcome while progressing with the results of any acquisition techniques. Due to its safe and non-invasive nature and being relatively cost efficient, UTI is among the most researched techniques in this field. Ultrasound provides an acceptable means of studying tongue shapes and movements during speech production. The imaging speed of this technique is suitable for studying the dynamics of speech production. By positioning the ultrasound probe in horizontal or vertical direction under the chin, midsagittal or conoral view of the tongue movement can be obtained respectively.

Articulation-to-Speech (ATS) synthesis typically involves two main steps: articulatory-to-acoustic mapping (AAM) and speech synthesis using a vocoder. In previous research [5], the WaveGlow neural vocoder was tested on a Hungarian ultrasound and audio dataset of approximately 15 minutes. Objective results demonstrated that WaveGlow outperformed the baseline continuous vocoder, achieving 0.1 dB lower mel-cepstral distortion in predicting spectral features. For the AAM step, an adaptation of the Tacotron2-based text-to-speech (TTS) method was applied to improve synthesis quality using a limited database of roughly 200 sentences [6]. Leveraging a Hungarian pre-trained Tacotron2 TTS model, the approach yielded more natural speech, as confirmed by subjective evaluation results, which highlighted the benefits of incorporating the Tacotron2 component.

https://doi.org/10.3311/WINS2025-006

Recent studies have explored the use of ultrasound tongue images to enhance text-to-speech (TTS) systems. One study investigated predicting tongue motion from text by generating PCA-compressed ultrasound images synchronized with speech [7]. Using fully connected neural networks (FC-DNN) and long short-term memory (LSTM), researchers found that FC-DNNs performed better in small datasets, and the generated ultrasound videos closely resembled natural tongue movements. This approach shows promise for applications such as audiovisual speech synthesis and pronunciation training. Another study focused on articulation-to-speech synthesis, using ultrasound images alongside text input to generate speech [8]. The results showed that combining text and articulatory data improved the naturalness of synthesized speech, especially in scenarios with limited training data, although challenges like ultrasound transducer misalignment could negatively impact performance.

UTI presents several challenges during data collection, primarily due to its dependence on the speaker and the sensitivity of the recording setup. The headset, which holds the ultrasound probe and microphone, can shift during sessions due to its weight and variations in the speaker's head shape, leading to misalignment either within or between speakers. To analyze such transducer misalignments, a study employed the Mean Square Error (MSE) distance to quantify relative displacement between the chin and the transducer [9]. The findings indicated that extreme MSE values often correspond to recording issues such as transducer misalignment, insufficient gel application, or incomplete contact between the skin and the probe. Additionally, UTI's speaker-dependent nature poses challenges, as speakers vary in head shape, tongue movement patterns, and speech rate. To address this, another study proposed a technique using dynamic time warping (DTW) to align ultrasound tongue images between speakers, improving the consistency and usability of the articulatory data for further processing [10].

While the UTI technique offers valuable insights into tongue movement, it has certain limitations. Specifically, UTI cannot capture the palate due to the airway above the tongue, and it struggles to clearly capture the tongue tip and blade because of the air space beneath. These challenges often result in incomplete or unclear tongue images, which can negatively impact the quality of data used for training models. Given the noisy and incomplete nature of the data, recent studies have explored data augmentation techniques to enhance ultrasound images, providing clearer insights into tongue shape and increasing the amount of training data—beneficial for neural networks [11]. However, it has been shown that the most effective augmentation techniques are speaker-dependent, and not all augmentation methods improve the quality of training data equally.

This article aims to answer the question: What is the best possible representation of tongue shape that can be captured using ultrasound imaging? For this, the cartesian coordinate results of DeepLabCut (DLC) pose estimation software [12] on original ultrasound tongue images were used to regenerate tongue images [13]. These regenerated images were also utilized for articulatory-to-acoustic mapping purposes as part of this work.

II. Dataset

The Tongue and Lips (TaL) corpus from the UltraSuite repository is a multi-speaker dataset comprising audio recordings, ultrasound tongue images, and lip videos (https://ultrasuite.github.io/data/tal_corpus/). The corpus is divided into two subsets: TaL1, which consists of six recording sessions of a professional male voice talent who is a native English speaker, and TaL80, which includes recording sessions from 81 native English speakers without professional voice training. Data collection involved using an ultrasound probe placed under the speaker's chin to capture sagittal views of tongue movements and a camera positioned in front of the mouth to record frontal views of lip movements [14].

In this study, we utilize publicly available CSV files (https://github.com/rachelfbeeson/DeepLabCut_ASR_for_TaL) from that were derived from the original TaL80 corpus experiments using DLC [15]. Each CSV file contains three columns for each articulator: x position (in pixels), y position, and a confidence score representing the reliability of the prediction.

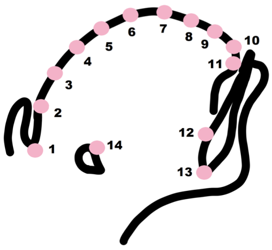

Fig. 1: The tongue points correspond to the vallecula (1), root (2-3), body (4-5), dorsum (6-7), blade (8-9), tip (10-11), short tendon (12), mandible (13), and hyoid (14). This figure and its caption were adapted from [15].

For this study, the initial experiments focused on data from a single speaker, identified as "01fi", a 33-year-old female native English speaker from Ireland. A total of 204 utterances produced by this speaker were analyzed. The CSV files provide tongue position coordinates for 14 anatomical points, as illustrated in Figure 1. However, for the purpose of ultrasound image regeneration, only 11 points were utilized. Points corresponding to the short tendon (12), mandible (13), and hyoid bone (14) were excluded due to their less direct relevance to the ultrasound tongue image reconstruction task.

III. Methods

The ultrasound tongue imaging technique produces a sequence of grayscale images captured frame-by-frame throughout the recording. These images initially appear in a wedge shape due to the configuration of the 64x842 ultrasound probe. In articulatory-to-acoustic mapping step during our work, the raw format of ultrasound images is utilized as input. Therefore, all wedge-shaped images must be converted into their raw form. Figure 2 depicts examples of the original ultrasound tongue image formats.

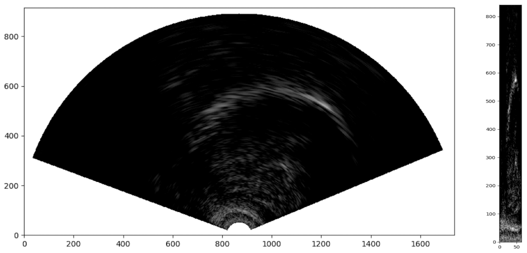

Fig. 2: 100th ultrasound tongue frame of the original 005_xaud utterence of speaker "01fi"; left - in wedge form, right - in raw form.

The CSV files obtained via DLC provide the positional data points from wedge form of images that serve as the foundation for our regeneration process. Two distinct methods were utilized for image regeneration, as detailed below. Examples will be shown using the "005_xaud" utterance from the speaker "01fi" ("The rainbow is a division of white light into many beautiful colours."), which consists of 651 frames in the original ultrasound tongue recording and 342 frames in the corresponding CSV file. Based on this, the 100th frame was selected for visualizing the processes. All of the selected values in this work were determined based on preliminary tests.

A. Regeneration process in two steps

In the first version of the regeneration task, the main focus was on the shape of the ultrasound images from which the x,y position values were derived. The goal was to reconstruct the ultrasound image by using these position values.

Initially, we created a blank image with dimensions of 860x1684 pixels, filled with the color black [RGB(0, 0, 0)]. This size was based on the original wedge-shaped ultrasound image, which was transformed into a rectangular format to make it easier to work with. The transformation aimed to standardize the image for further processing, which involved mapping the extracted x,y position values onto this newly created blank image.

To achieve this, we applied normalization and scaling techniques. Normalization adjusted the coordinates so that they fit within the dimensions of the 860x1684 image, while scaling ensured that the proportions of the original data were preserved when mapping the points. Essentially, these steps allowed the extracted x,y position values to be correctly placed onto the blank image, aligning with the actual locations of the tongue points in the ultrasound recording.

Once the x,y positions were mapped onto the blank image, we proceeded to draw the anatomical points and the lines connecting them, using the color white [RGB(255, 255, 255)]. This step helped to visualize the tongue's shape and movement, based on the provided data points.

In the second part of this version, we aimed to obtain the raw form of the ultrasound images, which required resizing the image from the 860x1684 format to the original ultrasound resolution of 64x842 pixels. This resizing process adjusted the image to match the actual dimensions of raw ultrasound image used in the further processing, ensuring that the reconstructed images adhered to the correct scale and resolution.

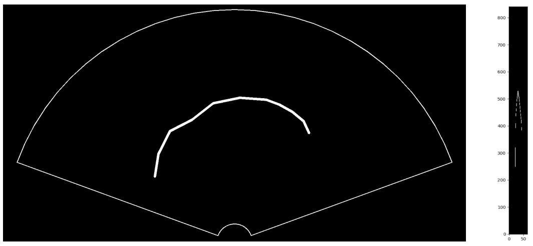

Fig. 3: 100th ultrasound tongue image frame "005_xaud" utterence of speaker "01fi" after the regeneration in two steps process; left - in wedge form, right - in raw form.

In Figure 3, the regenerated ultrasound images for the selected frame are shown, demonstrating the result of this regeneration process.

B. Regeneration process in one step

In the second version of the task, the order of the steps was reversed, compared to the first approach. The primary objective was to first convert the x, y position values of the 11 anatomical points into their raw form, and then plot these converted values onto a blank image that represented the raw ultrasound form. This shift in approach aimed to ensure that the conversion was done with minimal distortion and preserved the original structure of the ultrasound data.

The process began by resizing the original x, y position values to fit the target resolution of 64x842 pixels, which matched the raw form of the ultrasound image. This resizing step was crucial because it ensured that the scaled coordinates would align with the correct size and proportion.

Next, a blank image was created to serve as the canvas for the regeneration process. This image was also sized at 64x842 pixels, and was filled with the color black [RGB(0, 0, 0)] to represent the empty background before any points were added.

Once the blank image was prepared, the 11 anatomical points were plotted onto it using the resized x, y position values. After the points were placed, the lines connecting these points were drawn, using the color white [RGB(255, 255, 255)] for clarity.

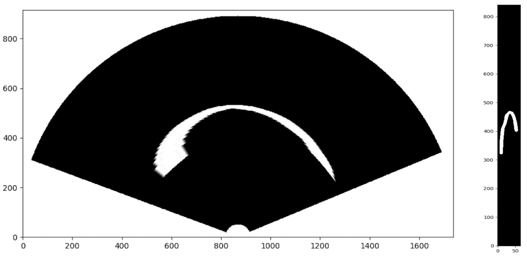

Fig. 4: 100th ultrasound tongue image frame ”005_xaud” utterence of speaker ”01fi” after the one-step regeneration process; left - in wedge form, right - in raw form.

Figure 4 shows the regenerated ultrasound images for the selected frame, illustrating the outcome of the one-step regeneration process.

For the conversion between raw form and wedge form, an adapted version of publicly available ultrasound tools was used (https://github.com/UltraSuite/ultrasuite-tools). These tools were customized to handle the specific requirements of the data and facilitate the accurate transformation between the two formats, ensuring consistency and preserving the integrity of the ultrasound images during the conversion process.

C. Articulatory-to-Acoustic Mapping

After regenerating the ultrasound tongue images, the articulation data was mapped to mel-spectrogram obtained from parallely recorded audio using a 2D Convolutional Neural Network (2D-CNN) towards articulation-to-speech synthesis. The 2D-CNN was chosen because it effectively captures spatial patterns in sequential data, making it ideal for processing the ultrasound image sequences that represent the dynamic tongue movements during speech.

The model was trained to map these image sequences to the corresponding mel-spectrogram. By learning the relationship between tongue movements and speech sounds, the 2D-CNN model links the articulatory gestures to their acoustic outputs, facilitating a deeper understanding of speech production [5].

IV. Results and Discussion

During processing, we observed that regenerating ultrasound images in a single step produced a fully connected line (tongue contour) in the raw form that closely resembled the structure of the original files. Based on this observation, the one-step regeneration results were used as input for the AAM task. During evaluation, these results were assessed both visually and objectively, as detailed below.

A. Visual Evaluation of Regeneration Process

To observe the complete flow of ultrasound tongue image sequences, ultrasound kymograms were utilized. Each frame in the kymogram was represented by the midline extracted from the wedge-shaped form of the frames, arranged sequentially to provide a comprehensive view of the entire utterance duration. More detailed information about ultrasound kymograms can be found in [11].

In Figure 5, the original and regenerated ultrasound kymograms are presented. These representations clearly reveal a misalignment between the two kymograms beginning around the 70th frame, which impacts the correspondence between the ultrasound sequences. Additionally, the peaks and troughs in the regenerated kymogram are shorter in duration compared to the original, resulting in less distinct and less observable outcomes.

The recordings were converted into video format from the original and regenerated ultrasound tongue image sequences using ultrasound tools to represent the final view of the entire utterance.[1]

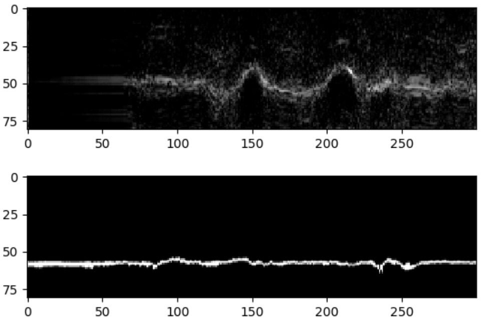

Fig. 5: Ultrasound kymogram representation of original (top) and regenerated form of ”005_xaud” of speaker ”01fi”.

B. Objective Evaluation of Regeneration Process

After visually examining the entire sequence of regenerated ultrasound tongue images, an objective metric was employed to quantify the similarity between the original and regenerated images. The choice of metric considered the specific characteristics of medical images. To this end, the Structural SIMilarity (SSIM) index was used, a widely accepted method for evaluating image similarity [16]. SSIM assesses perceptual differences by analyzing luminance, contrast, and structural features, producing a similarity score ranging from 0 (completely different) to 1 (identical).

Using a publicly available implementation of the SSIM index (https://github.com/andyj1/ssim_index), comparisons were performed on both the wedge form and raw form of

[1] Videos can be accessed at https://drive.google.com/drive/folders/1lB69XybzuGe-CmcMiEsAvCXXsdLJKTuU?usp=sharing

the original and regenerated images. These analyses approximated similarity with the scores of 0.926 0.926 and 0.973, respectively as shown in Table I.

Speaker	SSIM index for wedge form	SSIM index for raw form
01fi	0.926	0.973

TABLE I: Structural SIMilarity (SSIM) index values for wedge and raw form of ”005_xaud” utterence of speaker ”01fi”.

C. Objective Evaluation of Articulatory-to-Acoustic Mapping

The mean squared error (MSE) metric on the validation set (V-MSE) was utilized to compare the results obtained from the regenerated and original images for the speaker ”01fi”. Since the training model is identical to that used in the reference paper [5], the original validation MSE result for this speaker was adopted from that work.

Speaker	Original V-MSE	Regenerated V-MSE
01fi	0.212	1.392

TABLE II: Comparison of Original and Regenerated V-MSE values for speaker ”01fi”.

The final results are presented in Table II. During training for AAM, the regenerated ultrasound tongue images performed approximately 7 times worse than the original version.

D. Discussion

The findings in this study appear contradictory, for this reason, in this section, discussion about the results is given. Visual analysis of the regenerated ultrasound images reveals significant differences in the tongue's placement and shape compared to the original images. However, the calculated objective metric, the SSIM index, suggests a high degree of similarity between the original and regenerated images, with a near-identical index score.

To further investigate the calculated SSIM indices, SSIM difference maps for the wedge and raw forms were generated and are presented in Figures 6 and 7, respectively. These maps were scaled to a range of 0–255 for compatibility with standard grayscale image formats. In this representation, a value of 255 indicates the highest similarity (brightest regions), while a value of 0 represents the lowest similarity (darkest regions).

The difference maps clearly highlight that the SSIM indices fail to capture the actual similarity between the images, and the results are misleading. This is particularly concerning, as previous studies comparing medical images have found this indexing method to yield logical and reliable results [17].

When mapping the regenerated ultrasound images to their corresponding mel-spectrograms, the training results clearly indicated that the regenerated images failed to provide sufficient information for the convolutional neural network. This shortcoming resulted in a V-MSE that was significantly higher than that of the original version.

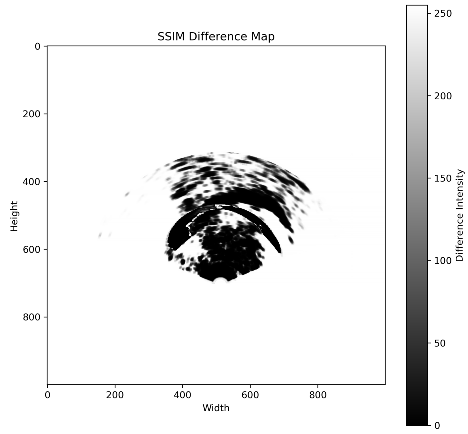

Fig. 6: Structural SIMilarity (SSIM) difference map for wedge form of ultrasound tongue frame.

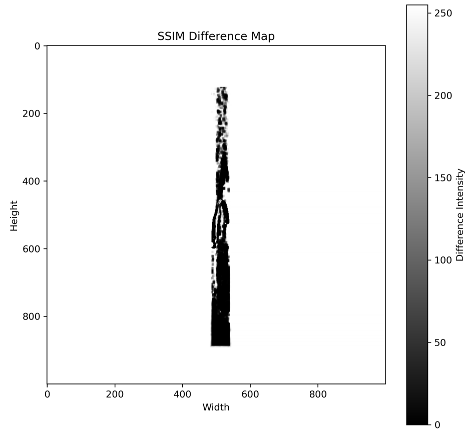

Fig. 7: Structural SIMilarity (SSIM) difference map for raw form of ultrasound tongue frame

This limitation could be attributed to the clarity of the regenerated images, which appear as nearly perfect black-and-white representations, unlike the more nuanced grayscale characteristics of actual ultrasound recordings. Additionally, the absence of certain frames from the original recording in the CSV file further reduced the number of frames regenerated, ultimately restricting the neural network to a limited dataset during training.

V. Conclusions and Future work

This study examined the regeneration of ultrasound tongue images from tongue position values obtained using the DeepLabCut pose estimation software, aiming to better represent tongue movements for articulatory-to-acoustic mapping. Visual and objective evaluations of the regenerated images revealed limitations, resulting in higher V-MSE scores during AAM training.

These findings highlight areas for improvement in regeneration methods. Ensuring the regenerated images closely resemble the original by incorporating noise or adjusting the tongue contour coloring could enhance accuracy. Additionally, the use of incomplete frame data may have contributed to the high V-MSE scores. This issue could be addressed by reprocessing the recordings in their entirety with DeepLabCut to ensure all frames are included and properly aligned.

Acknowledgment

We would like to dedicate this work to the memory of Dr. Tamás Gábor Csapó, whose pioneering contributions to ultrasound-to-speech synthesis have left an enduring legacy in the field. His innovative ideas, tireless dedication, and unwavering passion for advancing research were a constant source of inspiration. Beyond his remarkable professional achievements, his guidance, kindness, and encouragement were invaluable to us, particularly in shaping this work and inspiring many future research endeavors.

This study was supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) and 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).

References

[1] J. S. Perkell, "A physiologically-oriented model of tongue activity in speech production." 1974.

[2] B. Cao, N. Sebkhi, T. Mau, O. T. Inan, and J. Wang, "Permanent magnetic articulograph (ema) 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.

[3] A. Toutios, T. Sorensen, K. Somandepalli, R. Alexander, and S. S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data." in *Interspeech*, 2016, pp. 1492–1496.

[4] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–685.

[5] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based articulatory-to-acoustic mapping with waveglow speech synthesis," in *Interspeech 2020*, 2020, pp. 2727–2731.

[6] C. Zainkó, L. Tóth, A. H. Shandiz, G. Gosztolya, A. Markó, G. Németh, and T. G. Csapó, "Adaptation of tacotron2-based text-to-speech for articulatory-to-acoustic mapping using ultrasound tongue imaging," in *11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 54–59.

[7] T. G. Csapó, L. Tóth, G. Gosztolya, and A. Markó, "Speech synthesis from text and ultrasound tongue image-based articulatory input," in *11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 31–36.

[8] T. G. Csapó, "Extending text-to-speech synthesis with articulatory movement prediction using ultrasound tongue imaging," in *11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 7–12.

[9] T. G. Csapó, K. Xu, A. Deme, T. E. Gráczi, and A. Markó, "Transducer misalignment in ultrasound tongue imaging," in *Proceedings of the 12th International Seminar on Speech Production*, 2021, pp. 166–169.

[10] T. Csapó, "Is dynamic time warping of speech signals suitable for articulatory signal comparison using ultrasound tongue images?" in *1st Workshop on Intelligent Infocommunication Networks, Systems and Services*, 2023, pp. 65–70.

[11] I. Ibrahimov, G. Gosztolya, and T. G. Csapo, "Data augmentation methods on ultrasound tongue images for articulation-to-speech synthesis," in *12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 230–235.

[12] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[13] A. Wrench and J. Balch-Tomes, "Beyond the edge: Markerless pose estimation of speech articulators from ultrasound and camera images using deeplabcut," *Sensors*, vol. 22, no. 3, p. 1133, 2022.

[14] 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)*. IEEE, 2021, pp. 1109–1116.

[15] R. Beeson and K. Richmond, "Silent speech recognition with articulator positions estimated from tongue ultrasound and lip video," in *Interspeech 2023*, 2023, pp. 1149–1153.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] V. Mudeng, M. Kim, and S.-w. Choe, "Prospects of structural similarity index for medical image analysis," *Applied Sciences*, vol. 12, no. 8, p. 3754, 2022.