13th edition of the Speech Synthesis Workshop
24-27 August 2025, Leeuwarden, The Netherlands

Exploring Language Dependency in Ultrasound-to-Speech Synthesis

Ibrahim Ibrahimov[1], *Csaba Zainkó*[1], *Gábor Gosztolya*[2,3]

[1]Department of Telecommunications and Artificial Intelligence,
Budapest University of Technology and Economics, Budapest, Hungary
[2]HUN-REN–SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[3]University of Szeged, Institute of Informatics, Szeged, Hungary
ibrahim@tmit.bme.hu, zainko@tmit.bme.hu, ggabor@inf.u-szeged.hu

Abstract

Articulation-to-speech synthesis using ultrasound tongue imaging is a promising approach for Silent Speech Interfaces. However, its effectiveness is hindered by challenges such as session and speaker dependency, dataset scarcity and language variability. This study explores the language dependency of an ultrasound-to-speech synthesis system, consisting of a 2D-CNN to map ultrasound tongue images to mel spectrograms and a HiFi-GAN vocoder. The CNNs were trained on Azerbaijani recordings collected from three native speakers, each recorded in a single session containing both Azerbaijani (L1) and English (L2) sentences, and were then used to generate mel spectrograms for both languages. While the CNNs showed language dependency with lower mean squared error on L1, the mel-cepstral distortion of the synthesized speech did not reflect this, revealing the language bias of the vocoder. These results demonstrate the importance of considering language-specific factors in silent speech synthesis.
Index Terms: language dependency, ultrasound tongue imaging, silent speech synthesis

1. Introduction

Speech production is a process which involves several human body parts to co-operate for generating an audible speech [1]. However, there might be cases where audible speech is not possible to be produced (speech disorders, etc.) or is not desired due to the environment or some specific purpose (military, etc.). The need of an alternative real-time communication system in research has attracted interest towards data acquisition of the parts from speech production and its processing towards silent speech interfaces (SSI) [2, 3]. Movements of articulatory organs (tongue, lips etc.) can be collected using different techniques such as ultrasound tongue imaging (UTI) [4, 5], magnetic resonance imaging (MRI) [6, 7], electromagnetic midsagittal articulography (EMA) [8, 9] and lip video recording [10, 11].

UTI is a non-invasive, cost-efficient technique which is used in research for biosignal collection of tongue motion during speech. UTI is portable, easy to use, and allows real-time data collection, making it a promising option for SSI. Ultrasound tongue image frame sequences (UTIF) can be collected either in coronal or midsagittal form and the latter is often preferred as it visualizes the tongue from its root till tip in the best case [12]. Obtained UTIF depicts the detailed visualization of tongue movements which can be used in clinical speech analysis and therapy [13], linguistics [14], as well as biosignal-based articulation-to-speech synthesis (ATS) [15].

Given the correlation between tongue movements and the acoustic speech signal, several methods have been explored for ATS using UTI. A typical ultrasound-to-speech synthesis (UTS) system consists of two steps: a mapping from UTIF to some intermediate representation of the speech signal (e.g. mel-spectrogram), and synthesizing speech using a neural vocoder from this generated representations.

Convolutional Neural Networks (CNNs) are renowned for their exceptional feature representation capabilities from sequential image data, which have established them as a dominant force in image processing. As UTIF is a two dimensional image sequence, applying CNNs for the analysis and processing of UTIF demonstrated promising results [16]. In the work of Saha et al. [17], Ultrasound2Formant Net three dimensional CNN based architecture was used to map UTIF to formants which were then used to synthesize continuous speech using Klaat synthesizer. Csapó et al. [18] proposed an UTS system in which the estimation of 80-dimensional mel-spectrogram from UTIF was done using CNNs and synthesized samples were obtained via a WaveGlow vocoder.

UTS system comes with its limitations as described in details by Xie et al. [19]. Speaker and session dependency of UTI is a barrier to focus on generalization of UTS systems [20]. This is due to the variability of human head shape, tongue anatomy and lack of standardization in ultrasound probe placement [21]. As a result of these limitations, UTS systems are usually speaker-specific. On the other hand, speech produced by a speaker has specific linguistic context which is also related to phonology of the utilized language and its structure. This raises the question of how transferable an UTS system is for the same speaker in a different linguistic context. To observe the effects of linguistic differences, in this work, we try to explore language dependency of UTS system.

Besides having data scarcity for processing of UTS systems, according to our best knowledge, all of openly accessible UTI datasets are single language oriented and the results of the related works on them could not showcase the effects of linguistic changes between languages uttered by same speaker [10]. In this paper, we present a newly collected bilingual UTI dataset from three speakers producing audible speech in Azerbaijani (L1) and English (L2) in the same session. The purpose of data collection was to prevent speaker and session dependency of UTI and solely focus on the effects of linguistic structure differences. We use two-dimensional CNNs for mapping the UTIF to 80 dimensional mel-spectrogram, which were trained on L1 recordings, but are utilized to generate this this intermediate

10.21437/SSW.2025-25

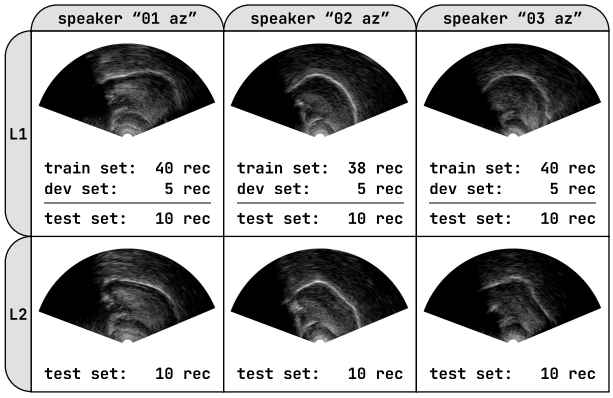

Figure 1: *Bilingual ultrasound tongue imaging dataset: Data splits (train, development, test) per speaker and language (L1: Azerbaijani, L2: English), with representative wedge representations.*

representation for recordings in both languages. Speech synthesis was done using Hifi-GAN vocoder for both languages from predicted mel-spectrograms. We utilized Mean Squared Error (MSE) and Mel-Cepstral Distortion (MCD) as objective measurements for this two-step system to evaluate the importance of language-specific factors in UTS towards SSI.

2. Dataset

Three male Azerbaijani speakers (average age of 25) with normal speaking abilities were recorded while producing sentences aloud in Azerbaijani and English languages. The subjects were native Azerbaijani and certified non-native English language speakers. Besides reading 50 Azerbaijani and 10 English sentences, speakers were asked to produce spontaneous speech while answering five situational questions in Azerbaijani. The read sentences were chosen from elicitation paragraph provided in [22], Harvard sentences [23, List 1], Rainbow passage [24, Chapter 8], TIMIT [25] and VCTK [26] corpus and translated to Azerbaijani. Sentences for the English recordings were taken from Rainbow passage solely.

Each subject was recorded individually in a quiet room while being fitted with an UltraFit stabilizing helmet which held a 2-4 MHz / 64 element 20mm radius convex ultrasound transducer and a Beyerdynamic TG H56c tan omnidirectional condenser microphone. The tongue movement was recorded in midsagittal orientation using a "Micro" ultrasound system of Articulate Instruments Ltd. at 81.5 fps. The data (tongue and audio) was recorded simultaneously utilizing the Articulate Assistant Advanced (AAA) software. Both the microphone signal and the ultrasound synchronization signals were digitized using an M-Audio – MTRACK PLUS external sound card at 22050 Hz sampling frequency. The ultrasound and the audio signals were synchronized using the frame synchronization output of the equipment with the AAA software.

Due to the weight of the stabilizing helmet, the duration of the recording sessions was limited to approximately 15 minutes. This resulted altogether 65 utterances per speaker (for the second speaker, two of L1 recordings were missing due to software malfunctioning). Participants were asked to read and agree to the consent form similar to standard version provided in [22] for further processing of collected dataset.

The Azerbaijani part of dataset was divided into train, development and test sets, while the ten English sentences were used for testing only. To ensure comparability across languages, the L1 and L2 test sets consisted of the same read sentences (the Rainbow Passage) for all speakers (Figure 1).

3. Methodology

3.1. Experimental setup

In our experiments, the scanline data of the ultrasound was used, after being resized to 64×128 pixels using bicubic interpolation. As the training set was quite small with the size of 38 (speaker "02az") and 40 recordings (speaker "01az" and "03az") (≈11 minutes), we utilized data augmentation techniques previously proposed in our work specifically for ultrasound tongue images [27]. According to their results, each of the presented augmentation method can have different performance impacts for different speakers. Based on this conclusion, we generated six data augmentation methods on each UTIF per speaker and randomly selected one of them per sentence to provide variability without relying on one specific augmentation technique. By this step, we doubled the size of the training set, while keeping the development and test set as they were, to depict real case scenarios.

All experiments were conducted on a server equipped with an Intel Core i7-4770 CPU (3.40 GHz, 8 cores, 16 threads) and an NVIDIA TITAN Xp GPU (12 GB VRAM) with the system of 32GB RAM. The experiments were performed on Ubuntu 18.04.6 LTS with Linux kernel 5.15.0-41-generic. For deep learning computations, we utilized TensorFlow 2.18.0 with CUDA 11.4 and NVIDIA driver 470.129.06.

3.2. Ultrasound-to-mel mapping

The recorded audio signal was converted to 80 dimensional mel-spectrogram representation; mapping from UTIF to mel was done using a CNN model for each speaker separately. Similarly to the presented structure by Csapó et al [18][1], we employed a model with a feed-forward architecture, which connected two 2D convolutional layers (kernel size: 13×13, number of filters: 30 and 60) followed by max-pooling with other two 2D convolutional layers with same kernel size (filters: 90 and 120) followed by max-pooling. After flattening was done on the resulting tensor shape, dense layer with 1000 and 80 neurons were employed to construct a 2D-CNN model to match the target dimension. Stochastic Gradient Descent optimizer were used to compile the presented model architecture with a manually selected learning rate of 0.1. The model was fitted on the L1 training set with a batch size 128 and an epoch number 50. Early stopping was used with a patience level of 3, while being validated on the L1 deveopment set with the MSE metric. Based on this structure, on average, for each speaker, training of 10 epochs took approximately 45 minutes.

3.3. Speech synthesis

L1-based speaker-specific mel-spectrogram generators were utilized to obtain predicted mel-spectrograms from UTIF of L1 and L2 test sets independently. For the step of speech synthesis from the generated mel-spectrogram, we chose high quality audio synthesizer and time-efficient HiFi-GAN neural vocoder [28]. We employed the first variation of the openly accessible pre-trained model on multi-speaker VCTK dataset

[1] https://github.com/BME-SmartLab/UTI-to-STFT

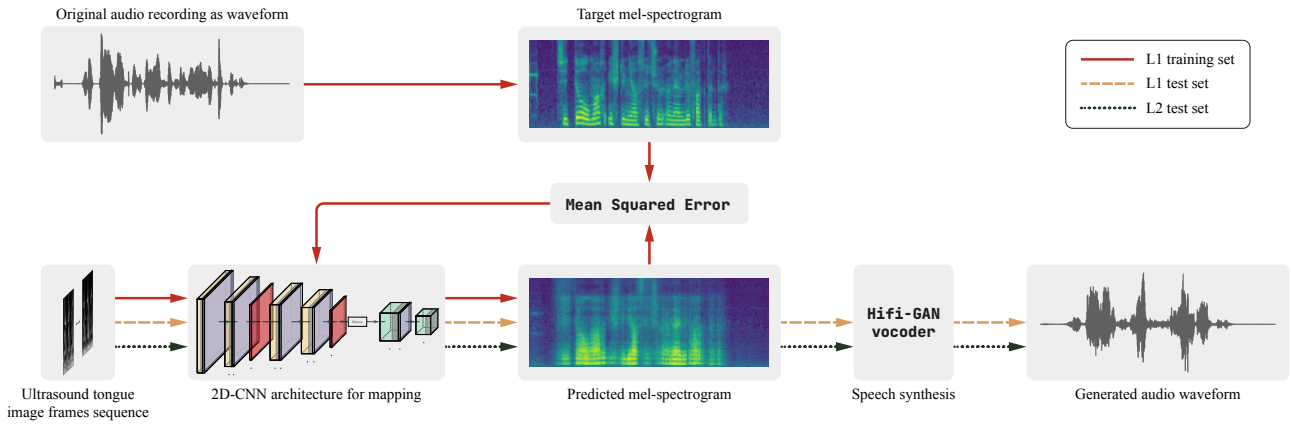

Figure 2: *Schematic diagram of the ultrasound-to-speech system pipeline, showing the 2D-CNN for ultrasound-to-mel mapping and the HiFi-GAN vocoder for speech synthesis. Training data (L1) flows through the solid arrow, while L1 and L2 test data flow through the dashed and dotted arrows after training phase was done, respectively. (L1: Azerbaijani, L2: English)*

Table 1: *Mean Squared Error (↓) results of mel-spectrogram predictions from L1 and L2 test sets.*

Language of recordings	**Speaker ID**		
	01az	02az	03az
L1 (Azerbaijani)	0.585	0.612	0.811
L2 (English)	0.774	0.878	0.851
Significance level (p)	0.015	0.002	0.280

Table 2: *Mel-Cesptral Distortion (↓) results of synthesized audio samples from L1 and L2 test sets.*

Language of recordings	**Speaker ID**		
	01az	02az	03az
L1 (Azerbaijani)	3.905	3.045	3.779
L2 (English)	3.942	2.882	3.793
Significance level (p)	0.481	< 0.001	0.123

(folder name: VCTK_V1)[2]. We selected the first variation of the HiFi-GAN generator due to its higher synthesized audio quality, regardless the longer running time as it does not affect the pipeline of our work. Mel spectrograms generated from L1 and L2 test sets were synthesized to audio waveforms separately as the final step of our UTS system.

The pipeline of our UTS system is demonstrated in Fig. 2.

4. Results

4.1. Preliminary consideration

It is important to note that the results of the UTS system should be analyzed separately for each speaker, as variations in the quality of the ultrasound tongue images can occur between individuals [29]. In this study, with three different speakers, we examine the tendencies in measurements for each speaker individually, allowing us to better understand the impact of language dependency on the performance of the system and to identify any speaker-specific trends in the results.

4.2. 2D-CNN mean squared error (MSE)

First, we report MSE values of the generated mel-spectrograms. That is, the speaker-specific 2D-CNN models (trained on L1 data) were employed to convert UTIF into mel-spectrograms, which were then compared to their original counterparts, and the MSE was computed for all recordings in each test set (L1 and L2). To assess whether any difference is statistically significant, we utilized the Mann-Whitney U test [30] on the (sentence-level) MSE scores for each speaker with a significance level of 0.05.

The average MSE values for each speaker and language are presented in Table 1. Clearly, the the system achieves lower MSE for the native language (Azerbaijani, L1) than for the second language (English, L2) for all three speakers, indicating better reconstruction accuracy for L1. This trend is particularly evident for speakers ”01az” and ”02az”; for these speakers, the difference is also statistically significant. In contrast, for speaker ”03az” the mean MSE value for Azerbaijani is noticeably higher than for the first two speakers, which is probably the reason why the difference between the two languages (L1 and L2) is not significant in this case.

4.3. Mel-cepstral distortion (MCD)

In addition to evaluating the mel-spectrogram generation, we analyze the quality of final synthesized speech samples using MCD. This metric quantifies the spectral distance between generated and original waveforms, with lower values indicating better synthesis quality. The MCD scores were computed following a publicly available implementation[3]. Table 2 presents the average MCD scores for each speaker in both test sets. The trends are not as clear-cut as they were in the MSE case: although the mean values for L1 are better than the L2 ones for two speakers (”01az” and ”03az”), the differences are quite small, not reaching the level of statistical significance. For the third speaker (”02az”), the mean MCD value is actually lower for English (i.e. L1) than for Azerbaijani (L2), representing a statistically significant difference, indicating an

[2]https://github.com/jik876/hifi-gan

[3]https://github.com/ttslr/python-MCD

Table 3: *Mel Cepstral Distortion (↓) results of synthesized speech using original mel-spectrograms from Azerbaijani (L1) and English (L2) test sets as input.*

Language of recordings	**Speaker ID**		
	01az	02az	03az
L1 (Azerbaijani)	1.549	1.371	1.466
L2 (English)	1.691	1.405	1.516
Significance level (p)	< 0.001	0.023	0.052

unexpected variation in the quality of the synthesized speech samples. Overall, by using MCD, the differences found via MSE vanished for two speakers, while for the remaining speaker it actually turned around. These findings suggest that while language context affects synthesis accuracy, its impact is not uniform across speakers.

4.4. Measuring the MCD of the vocoder

The previous two experiments differed not only in the evaluation metric used (i.e. MSE vs. MCD), but also on the level where the quality of the synthesis was evaluated: MSE was applied on the generated mel spectrograms, while MCD was calculated on the synthesized speech samples. Therefore, while the mean MSE values presented in Table 1 characterized the performance of the 2D-CNN models, the mean MCD scores in Table 2 summarized the performance of our whole workflow (i.e. the 2D-CNN *plus* the HiFi-GAN vocoder).

Due to this, in our last experiment we evaluated the performance of the HiFi-GAN VCTK_V1 generator model. This vocoder, widely used in speech synthesis, was selected after testing several available versions (with the VCTK dataset-trained variation providing the best results). However, as this generator was trained on a multi-speaker English dataset, analyzing its performance on Azerbaijani (an unseen language) is crucial. To assess this, we now used the *original* mel-spectrograms from the Azerbaijani and English test sets as input to generate audio waveforms. MCD was calculated to quantify synthesis accuracy; the results are shown in Table 3.

Surprisingly, across all speakers, synthesized samples from original mel-spectrograms in the Azerbaijani (L1) test set achieved lower MCD scores than those for English (L2), despite the fact that the vocoder was trained on English speech. This finding (that the vocoder also has language dependency), on one hand, explains why the L1 advantage over L2, found on the generated mel spectrograms, vanished when we investigated the synthesized speech samples. On the other hand, it also contradicts the expectation that the vocoder, trained on English, would perform better on English speech due to phonetic similarities. The difference in synthesized speech quality between the Azerbaijani and English test sets varied considerably across speakers, but in two cases it was statistically significant, while it remained at the edge of significance for the third speaker (”03az”).

Overall, we found that our, quite standard UTI-to-speech workflow was language-dependent on two distinct levels, as both the UTI-to-mel 2D-CNN network and the employed HiFi-GAN vocoder model showed statistically significant differences over the two languages for two out of the three speakers. This should be leveraged by using multilingual training data; by using an off-the shelf vocoder, however, this might not always be an available option.

5. Conclusions and Discussion

Addressing the open question of language dependency in UTS systems, this work investigated how the performance of UTS system gets affected by L2 speech when the system is trained exclusively on L1 data. A dedicated ultrasound tongue image and audio dataset was recorded to facilitate this analysis. A 2D-CNN and a HiFi-GAN vocoder were employed for the ultrasound-to-mel mapping and speech synthesis stages of the UTS system, respectively. Based on the objective evaluation using MSE and MCD metrics, we found that our, quite standard UTS system is language-dependent both on the UTI-to-mel and on the vocoder level. This underscores the necessity of considering linguistic context and individual pronunciation patterns in the development of ultrasound-based ATS systems.

Several factors could contribute to this finding. The inherent phonetic variations between English and Azerbaijani, particularly in vowel pronunciation, intonation, and the realization of certain consonants, may pose a challenge for the English-trained vocoder [31]. For example, English has a wider range of vowel sounds than Azerbaijani, potentially hindering the ability of the vocoder to accurately reproduce the nuances of English vowels. Similarly, intonation patterns differ significantly; English relies more heavily on intonation for conveying meaning, while Azerbaijani intonation is less prominent. These cross-linguistic phonetic differences, compounded by individual speaker characteristics like vocal tract morphology or speaking style, can significantly impact generalization performance.

The contrasting MCD scores of speakers ”01az” and ”02az” highlight the difficulties encountered in generalizing across speakers with varying accents and pronunciation, even when the target language is the same as the training language. The larger discrepancy in L1 and L2 MCD scores for speaker ”01az” suggests that the vocoder, despite its overall performance, may be particularly sensitive to subtle phonetic differences introduced by non-native pronunciation in certain speakers. This observation aligns with broader challenges in text-to-speech (TTS), where accent variation and cross-lingual generalization are active research areas. Studies in TTS have shown that even high-performing vocoders trained on one language or accent group can exhibit performance variations when applied to others [32].

The unexpected lower MCD for L2 test set of speaker ”02az” presents a unique puzzle. While a detailed analysis is beyond the scope of this paper, we hypothesize that English pronunciation of speaker ”02az”, despite being non-native, may coincidentally align more closely with specific acoustic patterns in the training data of the utilized vocoder.

For future work, we intend to expand our dataset by including additional bilingual speakers, particularly focusing on female voices, to increase diversity. Recognizing that spontaneous speech often reveals more nuanced linguistic characteristics than read speech, we will prioritize recording spontaneous utterances for both existing and new speakers. To mitigate the influence of linguistic factors in the synthesis stage, we plan to investigate fine-tuning the neural vocoder on Azerbaijani dataset. Finally, a detailed analysis of UTIF from both languages, considering specific linguistic factors, will be crucial for a deeper understanding of language dependency and the development of effective solutions.

For reproducibility, the full code and synthesized speech samples are available at https://doi.org/10.5281/zenodo.15808656.

6. Acknowledgements

This study was supported by the NRDI Office of the Hungarian Ministry of Innovation and Technology (grant TKP2021-NVA-09), and within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004) and the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI).

The authors gratefully acknowledge the subjects for their valuable participation in the data collection.

This work is dedicated to the memory of Dr. Tamás Gábor Csapó, whose inspiration and influence continue to guide us. May he rest in peace.

7. References

[1] P. B. Denes and E. N. Pinson, *The Speech Chain: The Physics and Biology of Spoken Language*, 2nd ed. New York: W. H. Freeman, 1993.

[2] 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.

[3] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. M. Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE access*, vol. 8, pp. 177 995–178 021, 2020.

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

[5] 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.

[6] R. Trencsenyi and L. Czap, "Ultrasound-and mri-based speech synthesis applying neural networks," in *2024 25th International Carpathian Control Conference (ICCC)*. IEEE, 2024, pp. 1–6.

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

[8] A. Wrench and W. Hardcastle, "A multichannel articulatory database and its application for automatic speech recognition," *5th Seminar on Speech Production: Models and Data*, 01 2000.

[9] Z. I. Skordilis, A. Toutios, J. Töger, and S. Narayanan, "Estimation of vocal tract area function from volumetric magnetic resonance imaging," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 924–928.

[10] 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.

[11] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 455–462.

[12] T. H. Shawker, B. C. Sonies, and M. Stone, "Soft tissue anatomy of the tongue and floor of the mouth: an ultrasound demonstration," *Brain and language*, vol. 21, no. 2, pp. 335–350, 1984.

[13] E. Sugden and J. Cleland, "Using ultrasound tongue imaging to support the phonetic transcription of childhood speech sound disorders," *Clinical linguistics & phonetics*, vol. 36, no. 12, pp. 1047–1066, 2022.

[14] B. Gick, B. Bernhardt, P. Bacsfalvi, I. Wilson, M. Zampini *et al.*, "Ultrasound imaging applications in second language acquisition," *Phonology and second language acquisition*, vol. 36, no. 6, pp. 309–322, 2008.

[15] W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov, and S. Lee, "Biosignal sensors and deep learning-based speech recognition: A review," *Sensors*, vol. 21, no. 4, p. 1399, 2021.

[16] E. M. Juanpere 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.

[17] P. Saha, Y. Liu, B. Gick, and S. Fels, "Ultra2speech-a deep learning framework for formant frequency estimation and tracking from ultrasound tongue images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 473–482.

[18] 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.

[19] Z. Xia, R. Yuan, Y. Cao, T. Sun, Y. Xiong, and K. Xu, "A systematic review of the application of machine learning techniques to ultrasound tongue imaging analysis," *The Journal of the Acoustical Society of America*, vol. 156, no. 3, pp. 1796–1819, 2024.

[20] 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.

[21] J. Cleland, "Ultrasound tongue imaging," in *Manual of clinical phonetics*. Routledge, 2021, pp. 399–416.

[22] S. Weinberger, "Speech accent archive," George Mason University, 2015. [Online]. Available: https://accent.gmu.edu/

[23] E. H. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[24] G. Fairbanks, *Voice and Articulation Drillbook*. Harper & Brothers, 1940.

[25] J. S. Garofolo, "Timit acoustic-phonetic continuous speech corpus," 1993.

[26] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.

[27] 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.

[28] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[29] L. Tóth, A. Honarmandi Shandiz, G. Gosztolya, and T. G. Csapó, "Adaptation of tongue ultrasound-based silent speech interfaces using spatial transformer networks," in *Interspeech 2023*, 2023, pp. 1169–1173.

[30] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[31] L. Safarova, "Comparative analysis of azerbaijani and english phonetic systems," *EuroGlobal Journal of Linguistics and Language Education*, vol. 1, pp. 17–25, 10 2024.

[32] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding," in *Interspeech 2019*, 2019, pp. 2105–2109.