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

How Silent Are Silent Speech Interfaces? Speech Reconstruction From Whispered and Silent Ultrasound Tongue Images

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

[1]University of Szeged, Institute of Informatics, Szeged, Hungary
[2]HUN-REN–SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[3]Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics, Budapest, Hungary

ggabor@inf.u-szeged.hu, ibrahim@tmit.bme.hu, zainko@tmit.bme.hu

Abstract

In the silent speech interfaces (SSI) area the aim is to restore or recognize speech whenever normal verbal communication is not possible or desirable. SSI systems use some non-acoustic biosignal of the body (e.g. tongue or lip movement) as input, and they are typically trained on data where real speech was produced, implicitly assuming that during silent (i.e. whispered or silently articulated) speech production the articulatory organs move similarly as they do during normal speaking. In this study we test this hypothesis in practice: we train our speech restoration DNNs on ultrasound tongue images recorded during audible speech, and synthesize speech from images recorded during whispering and two types of articulated-only speech. We found that synthesized speech for these silent "speaking" styles is significantly less intelligible than for audible speech, suggesting a difference in the articulatory movements, which should be considered when training silent speech restoration models.

Index Terms: ultrasound tongue imaging, silent speech synthesis, articulated speech, whispered speech

1. Introduction

Human speech production requires the coordination of various articulatory organs [1]. In certain situations, however, producing audible speech may not be possible (e.g. after laryngectomy) or it is undesirable (e.g. extreme background noise, specific military applications). Owing to these factors, there is a growing interest in processing "speech", when the speaker actually produces only soundless articulatory movements, which area is called silent speech interfaces (SSI, [2, 3]). These systems vary both in their input (electromagnetic articulography (EMA, [4]), ultrasound [5, 6, 7], surface electromyograpy (EMG, [8]), magnetic resonance imaging (MRI, [9]) or multimodal [10, 11, 12]) and in their aim (silent speech recognition [11, 13] or synthesis [5, 9, 14]).

Although the motivation behind silent speech interfaces in general, and silent speech synthesis in particular, is to handle situations when the user is articulating without producing any sound (or, at most, whispers), it is straightforward to use recordings where the articulatory movement (e.g. ultrasound tongue images (UTI) or EMG time series) and the speech signal are recorded in parallel [15, 16]. This way the training targets of the underlying machine learning model (e.g. DNN) can directly be derived from the corresponding speech utterances both for silent speech recognition (SSR, [16, 17]) and for silent speech synthesis (or reconstruction, [15]), while the input of the model can be derived from the (time-aligned) articulatory movements.

Of course, articulatory movements do differ in normal and in silent and/or whispered speech [18, 19]. Therefore, if one decides to use recordings of modal speech for training an SSI system, there will be a mismatch between the (vocalized) training data and the (silent) input of the model during application. This mismatch and its effect to system performance was investigated in several studies. For example, Janke et al. investigated EMG-based silent speech recognition of audible, whispered and silent speech using a HMM-based workflow [20]. Wand et al. developed a spectral mapping method to reduce the discrepancy between the EMG signals of audible and silent speech, but still reported consistently larger WER values for silent speaking mode than for audible recordings [21]. Petridis et al. utilized normal, whispered and silent recordings on visual speech recognition (i.e. lip videos), and found that the effect of this mismatch is significant, silent speech differing the most from the other two speaking styles [22].

Regarding UTI, Ribeiro et al. investigated recognition of silent, whispered and modal speech using both ultrasound tongue images and lip videos [11], using only modal utterances for model training. They found that word error rates were significantly higher for silent speech than for the other two speaking styles. They also found differences in utterance durations and in the articulatory space (i.e. intensity of tongue movements) between modal and silent speech.

What these studies have in common, though, is that they all focus on silent speech *recognition*. This task is simpler than silent speech synthesis in the sense that it is easier to obtain training targets even for articulatory movement data recorded during silent speech, as we need only to map a finite set of phonetic classes to each input feature vector (e.g. ultrasound image). Evaluation is also quite simple, requiring only the transcript of the silently articulated word or word sequence. However, to the best of our knowledge, no study yet investigated the effect of whispered and silent speech in silent speech *synthesis*.

In this study we investigate how silent and whispered speech deteriorates the quality of the synthesized speech samples. To do this, we utilize a standard UTI-based speech synthesis workflow, consisting of a two-dimensional convolutional neural network (2D-CNN) to map each ultrasound image to the corresponding mel spectral vector, and a HiFi-GAN vocoder to produce the final speech signal from the mel spectrogram. We train the neural network component on recordings containing vocalized speech, and evaluate it on ultrasound tongue image sequences recorded in four styles (vocalized, whispered and two types of silent articulation). The intelligibility of the synthesized samples is measured by a phone-level ASR system.

10.21437/SSW.2025-24

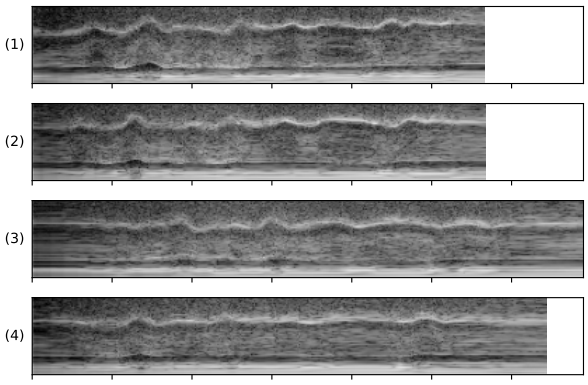

Figure 1: *A sample set of kymograms, containing the same sentence (#7) for the same speaker (ID 103). (1): normal speech, (2): whispered speech, (3): silent hyperarticulated speech, (4): silent speech with normal articulation.*

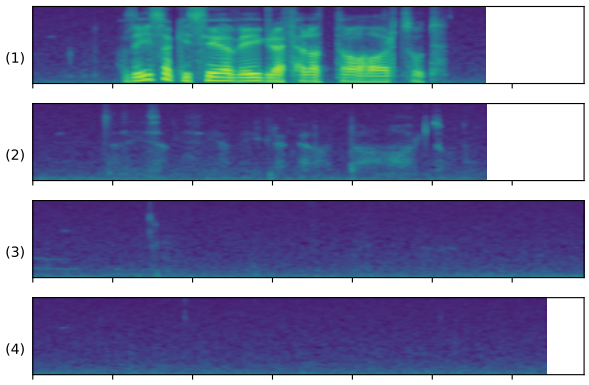

Figure 2: *The mel spectrograms of the same recordings as shown in Fig. 1. (1): normal speech, (2): whispered speech, (3): silent hyperarticulated speech, (4): silent speech with normal articulation.*

2. Data

Several Hungarian male and female subjects with normal speaking abilities were recorded while reading 200 sentences aloud, then two female and one male speakers were selected for the experiments (with speaker IDs 048, 049 (female) and 103 (male)). The average duration of the recordings was about 15 minutes per speaker. The speech signal and the ultrasound articulatory data were recorded in parallel, and they were synchronized using the software provided with the UTI equipment. The tongue movement was recorded in a midsagittal orientation using the 'Micro' ultrasound system of Articulate Instruments Ltd. at 81.67 fps. The speech signal was recorded with a Beyerdynamic TG H56c omnidirectional condenser microphone. (For more details on the recording set-up, see [23].)

In the experiments we used the raw scanline data of the ultrasound as the input of an articulatory-to-acoustic mapping CNN after resizing it to 64×128 pixels. The intensity range of the images was min-max normalized to the [-1, 1] interval. The speech signals were converted to mel spectrograms with 80 spectral components. These 80-component spectral vectors served as the training targets for our neural networks, after standardizing to zero mean and unit variance. The 200 read sentences were used for training the UTI-to-Mel neural network, with a random 190-10 division for training and development.

Besides the (vocalized) recordings used for model training, the speakers read the 9-sentence long Hungarian version of the short tale "The North Wind and the Sun". These sentences were read in four variations:

(1) **Normal**: The sentences were read aloud.

(2) **Whispered**: The speakers read the sentences whispering.

(3) **Silent (hyperarticulation)**: The sentences were read silently, but with articulation. The speakers were asked to move their tongue and lips in an exaggerated manner.

(4) **Silent (normal articulation)**: The sentences were read silently, but with articulation. This time the speakers were asked to refrain from hyperarticulation.

The ultrasound tongue images of the utterances of "The North Wind and the Sun" were used as the basis of speech synthesis.

Fig. 1 shows the kymograms (the change of the middle vertical line in the ultrasound image over time) for the 7th sentence of speaker 103 in the four variations produced ("Az Északi Szél végre feladta a harcot.", i.e. "At last the North Wind gave up the attempt."). The four kymograms have different durations, as they belong to different "utterances". The bumps (indicating sudden vertical tongue movements) are also present at different locations, and they are slightly less intense in cases (3) and (4), which correspond to silent articulation.

The corresponding mel spectrograms can be seen in Fig. 2. These spectrograms appear as expected: while the formants are clearly visible on the audible speech (i.e. (1)), they are a lot less apparent in the whispered variation (i.e. (2)). The relation between these two mel spectrograms and the corresponding kymograms is clear, as the sentence begins slightly earlier in whispered utterance than in the normally uttered one, and the two more intensive tongue movements (i.e. bumps) at the beginning are in synch with this. As expected, the two silent variations ((3) and (4)) have empty mel spectrograms.

3. Experimental setup

We applied a straightforward 2D convolutional (2D-CNN) network that transforms one ultrasound image to one mel-frequency spectral vector [24]. This network consisted of 4 convolutional layers with 30-60-90-120 filters, and had a max-pooling layer after every second layer. This was followed by a fully connected layer of 1000 neurons, and the linear output layer consisting of 80 neurons. All the hidden layers used the Swish (or silu) activation function, while the output layer had linear neurons. To reduce overfitting, we used dropout layers ($p = 0.2$) after each convolutional layer and also after the only dense layer. The network was trained to minimize MSE with a batch size of 128 and an initial learning rate of 1e-4, which was regulated using the AdamW optimizer. Early stopping was employed based on the MSE value of the validation set (i.e. 10 sentences). Similarly to most works in UTI-based speech synthesis [25, 15, 26], we trained speaker-dependent models, i.e. a separate 2D-CNN network was trained for each speaker.

For generating the synthesized speech samples from the mel spectrograms, we chose a high-quality and efficient HiFi-GAN neural vocoder [27]. We employed the first variation of the openly accessible pre-trained model on the multi-speaker VCTK dataset (VCTK_V1)[1]. We selected the first variation of the HiFi-GAN generator due to its higher synthesized audio quality, despite its slightly higher running time.

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

Table 1: *Phonetic error rates (PER%) measured for the three speakers and the four different speaking styles.*

Speaking style	Speaker ID		
	spk048	spk049	spk103
Original recording	21.6%	18.8%	19.5%
Normal (audible)	64.9%	64.8%	51.5%
Whispered	88.8%	86.0%	91.7%
Silent (hyperarticulated)	96.9%	89.6%	96.9%
Silent (normal articulation)	96.7%	96.0%	96.9%

Table 2: *Recognized phonetic ratios, i.e. number of phones recognized divided by the total number of phones.*

Speaking style	Speaker ID		
	spk048	spk049	spk103
Original recording	100.9%	100.9%	100.3%
Normal (audible)	71.8%	68.4%	83.1%
Whispered	26.6%	32.8%	20.0%
Silent (hyperarticulated)	7.9%	29.5%	4.5%
Silent (normal articulation)	6.4%	11.9%	4.3%

3.1. Evaluation

In a standard articulatory-to-acoustic mapping set-up, objective metrics such as MSE and mean cepstral distortion [28] can be used to quantify the quality of synthesized speech. However, these metrics compare the synthesized audio with the original one, which is clearly not an option when the speaker actually uttered no speech (or only whispered it), while we expect our system to output voiced speech. Another option might be to compare all four synthesized recording variations to the original, audible speech recording; however, this was also unfeasible since the four utterances significantly differed in their timing (see Fig. 1 again).

Due to the above difficulties, we applied an automatic speech recognition (ASR) system to rate the quality of the synthesized recordings [29, 30]. (We are aware that this procedure does not measure naturalness, but only intelligibility; however, in this work our focus was the intelligibility of the generated speech samples.) Our first attempt was to use Whisper [31], but it was hampered by the frequent hallucinations of the model [32]. Owing to this, we employed a traditional phone-level HMM/DNN system, trained on 240 hours of noise-augmented spontaneous speech from the Hungarian BEA dataset [33], and a simple phone bigram as the language model. From the output of the ASR system we calculated phonetic error rate (PER%) scores for all nine sentences altogether, which served as a proxy for intelligibility. (Exploiting the properties of Hungarian orthography, the phonetic transcription of the sentences was created automatically, in a rule-based way.)

4. Results

Table 1 shows the measured phonetic error rates for the three speakers and the four speaking styles. For the original recordings we measured PER% values in the range 18.8%. . .21.6%, which characterize the accuracy of our phonetic ASR system, and indicate the glass ceiling for the PER value for the synthesized speech samples. The fact that the values of the three speakers are quite close to each other indicate that the ASR system had no significant speaker dependency.

Compared to these values, the phonetic error rates for the synthesized samples from the normal (audible) recordings are significantly higher: they lay between 51.5% (speaker 103) and 64.9% (speaker 48). This shows that the employed workflow to reconstruct the speech samples from the ultrasound tongue images deteriorates the intelligibility of the resulting speech signals, even when both the training and the test UTI frames were recorded during audible speech.

Regarding whispered speech, the phonetic error rates are even higher: they lay in the range 86.0%. . .91.7%, showing a clear tendency (but also reflecting a speaker-wise variation). This level of lower intelligibility clearly indicates that the tongue, the lips, and any other possible organ used during speech production which is visible on the ultrasound tongue images move differently during whispered speech production than they do during normal (audible) speech. By listening to these generated samples, we could acknowledge that they usually contained some longer intelligible parts. We can also observe a significant amount of speaker-dependency: for speakers 048 and 049, the phonetic error rates increased from cca. 65% to 86-89%, while for the third speaker (ID 103) there was a larger gap (from 51.5% to 91.7%).

The results for the two silent styles show that the generated speech samples were even of a lower quality than those synthesized from the whispered UTI recordings: in five cases out of six, the phonetic error rate values reached 96%. The only exception was the hyperarticulated variation of speaker 049, where the 89.6% PER score is still quite high, but falls quite close to the values of the whispered recordings (i.e. 86.0%. . .91.7%). This outlier value probably reflects the behaviour of the particular speaker instead of some biological phenomenon; that is, how she articulated when following the instruction "move your tongue and lips in an exaggerated manner".

4.1. Recognized phonetic ratios

By investigating the results of phone-level recognition, the high number of phonetic deletion errors was quite apparent, which probably reflects the (observed) low volume of the synthesized speech samples from the silent ultrasound images. To express this, we calculated the total number of phones in the output of the phonetic ASR system, and divided it by the total number of phones in the transcription. The lower this ratio is, the more silence is present in the (synthesized or original) speech sample.

Table 2 shows these recognized phonetic ratios (expressed in percentages for the sake of better readability). Clearly, the speech samples generated from the normal (i.e. audible) recordings contained more speech (and/or more distinguishable phones) than the synthesized samples from the whispered input: the former recordings had 68. . .83% of the original phones, while the latter recordings contained only 20. . .33%. The speech samples reconstructed from the two types of silent utterances have even lower values, indicating that the ASR output contained mostly silence phones (being present at the start and end of each utterance). The exception is again speaker 103 with the "silent (hyperarticulated)" recordings, where the amount of phones recognized (29.5%) roughly matched that of the whispered recordings (32.8%).

4.2. The generated mel spectrograms

Fig. 3. shows the generated mel spectrograms for sentence #7 from speaker 103. (The kymograms and the original mel spec-

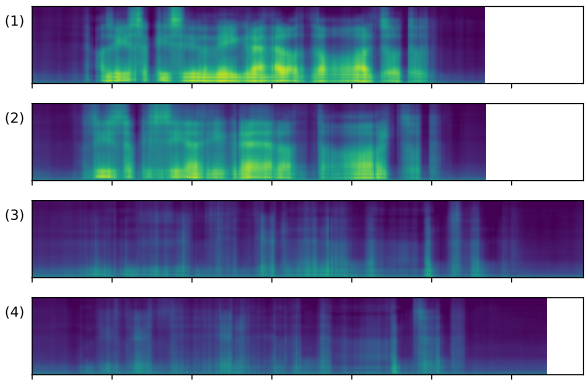

Figure 3: *The mel spectrograms of the synthesized samples for the same sentence as in Fig. 4 (sentence #7) for speaker 103. (1): normal speech, (2): whispered speech, (3): silent hyperarticulated speech, (4): silent speech with normal articulation.*

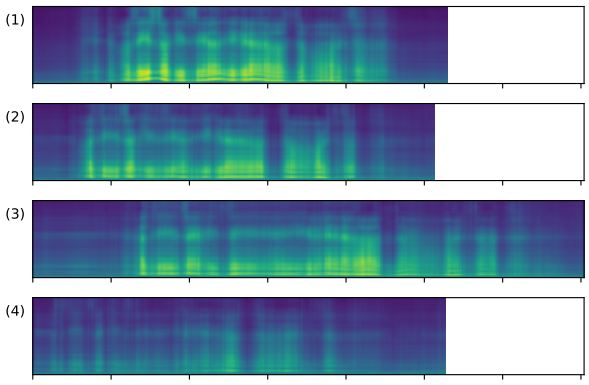

Figure 4: *The mel spectrograms of the synthesized samples for the same recordings as shown in Fig. 4 (sentence #7) for speaker 049.*

trograms for these utterances can be seen in Fig. 1. and Fig. 2, respectively.) As it can be seen, the mel spectrogram corresponding to the whispered utterance (i.e. (2)) is quite similar to the one generated for the recording with normal speaking style (1), but the timing is slightly different, and the formants are more blurred. This latter property is probably the reason why the ASR system distinguished a lower number of independent phones in this synthesized speech sample than in the first, normally uttered one. From the silent utterances, however, quite low-intensity "speech" recordings were synthesized; the ASR module recognized both displayed samples simply as silence.

Fig. 4., in contrast, shows the generated mel spectrograms of the same sentences for speaker 049. Although the one corresponding to the normally articulated silent recording (i.e. (4)) appears to be similarly silent as for the previous speaker, the hyperarticulated speech sample (3) is significantly louder. (Also, it is significantly longer than the other three samples, which phenomenon was reported for silent speech articulation [11].) This, in our opinion, demonstrates that it is possible to synthesize audible speech samples from ultrasound tongue images even when the speaker omits no sound. However, it most likely requires the subject to hyper-articulate, and, based on our results, it can be significantly speaker-dependent.

4.3. Utterance durations

Finally, we take a look at the duration of utterances produced by the three speakers in the four speaking styles. The mean durations of one phone (i.e. total utterance duration divided by the number of phones present in the transcription) are plotted in Fig. 5. Although there is indeed a visible speaker-wise variation, it is obvious that silent hyper-articulated speaking style, in average, led to longer-articulated phones than the three other speaking styles. It was already reported that silent articulation leads to increased duration (e.g. by Ribeiro et al. [11]). However, in our case we found this only for the hyper-articulated recordings, while the mean phonetic duration of the silent, but normally (i.e. not over-) articulated utterances was similar to those of the whispered and modal speech recordings.

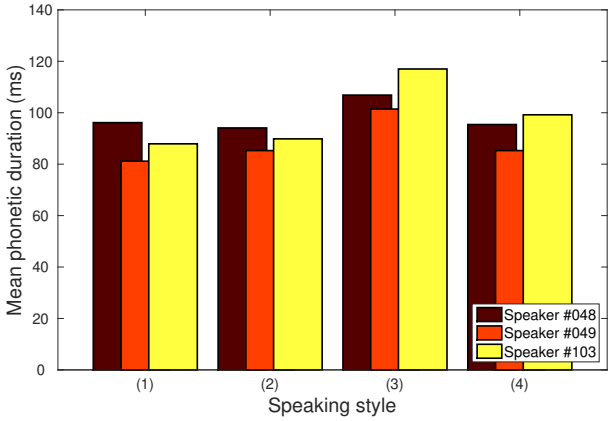

Figure 5: *Mean phonetic durations for the three speakers and four speaking styles. (1): normal speech, (2): whispered speech, (3): silent hyperarticulated speech, (4): silent speech with normal articulation.*

5. Conclusions

In this study, motivated by works from the silent speech recognition area, we investigated how whispered and silent speech affects a standard silent speech *synthesis* system. For this aim we trained speaker-dependent CNNs on ultrasound tongue imaging (UTI) frames, which were recorded during normal (audible) speech, then we synthesized speech samples from recordings of normal, whispered and two types of silently articulated speaking styles. By our results, synthesizing audible speech from whispered recordings is possible, but the generated speech samples are not always intelligible. However, ultrasound images from the silent speaking styles led to (at least partly) intelligible speech only for one speaker, and only when she was hyper-articulating. This indicates that the speech production organs (or at least those captured by an ultrasound transducer in midsagittal orientation) move differently during silent and whispered speech than they do during normal speech production, to the extent that this effect harms the quality of the output of a standard UTI-based speech synthesis system. Regarding the only speaker where we generated audible speech, we believe that the key was *how* she followed the instructions to hyper-articulate. Of course, this hypothesis has to be verified with more speakers, which we plan to do in the near future. Another solution towards really silent speech synthesis could be to collect more training data and solve the issue of missing training targets, which is another path we intend to follow.

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 would also like to thank Tamás Gábor Csapó for the recordings and all the inspiration. 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. González-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] J. Wang, A. Samal, J. R. Green, and F. Rudzicz, "Whole-word recognition from articulatory movements for silent speech interfaces," in *Proceedings of Interspeech*, Portland, OR, USA, Sep 2012, pp. 1327–1330.

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

[6] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone, "Prospects for a silent speech interface using ultrasound imaging," in *Proceedings of ICASSP*, vol. 1. IEEE, 2006, pp. I–I.

[7] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer, Speech & Language*, vol. 36, pp. 274–293, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2015.03.005

[8] L. Diener, M. R. Vishkasougheh, and T. Schultz, "CSL-EMG_Array: An open access corpus for EMG-to-speech conversion," in *Proceedings of Interspeech*, Shanghai, China, Oct 2020, pp. 2958–1796.

[9] Y. Otani, S. Sawada, H. Ohmura, and K. Katsurada, "Speech synthesis from articulatory movements recorded by real-time MRI," in *Proceedings of Interspeech*, Dublin, Ireland, Aug 2023, pp. 127–131.

[10] 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 *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 1467–1471.

[11] M. S. Ribeiro, A. Eshky, K. Richmond, and S. Renals, "Silent versus modal multi-speaker speech recognition from ultrasound and video," in *Proceedings of Interspeech*, Brno, Czech Republic, Sep 2021, pp. 641–645.

[12] R. Beeson and K. Richmond, "Silent speech recognition with articulator positions estimated from tongue ultrasound and lip video," in *Proceedings of Interspeech*, Dublin, Ireland, Aug 2023, pp. 2958–2962.

[13] I. Salomons, E. del Blanco, E. Navas, and I. Hernáez, "Spanish phone confusion analysis for EMG-based silent speech interfaces," in *Proceedings of Interspeech*, Dublin, Ireland, Aug 2023, pp. 1179–1183.

[14] P. Wu, T. Li, Y. Lu, Y. Zhang, J. Lian, A. W. Black, L. Goldstein, S. Watanabe, and G. K. Anumanchipalli, "Deep speech synthesis from MRI-based articulatory representations," in *Proceedings of Interspeech*, Dublin, Ireland, Aug 2023, pp. 5132–5136.

[15] T. G. Csapó, L. Tóth, G. Gosztolya, and A. Markó, "Speech synthesis from text and ultrasound tongue image-based articulatory input," in *Proceedings of SSW*, Budapest, Hungary, Aug 2021, pp. 31–36.

[16] J. Wu, Y. Zhang, L. Xie, Y. Yan, X. Zhang, S. Liu, X. An, E. Yin, and D. Ming, "A novel silent speech recognition approach based on parallel inception convolutional neural network and mel frequency spectral coefficient," *Frontiers in Neurobotics*, vol. 16, 2022.

[17] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer, Speech & Language*, vol. 39, no. Sep, pp. 67–87, 2016.

[18] C. Dromey and K. M. Black, "Effects of laryngeal activity on articulation," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 25, no. 12, pp. 2272–2280, 2017.

[19] K. J. Teplansky, B. Y. Tsang, and J. Wang, "Tongue and lip motion patterns in voiced, whispered, and silent vowel production," in *Proceedings of ICPhS*, Melbourne, Australia, Aug 2019, pp. 1–5.

[20] M. Janke, M. Wand, and T. Schultz, "Impact of lack of acoustic feedback in EMG-based silent speech recognition," in *Proceedings of Interspeech*, Makuhari, Chiba, Japan, Sep 2010, pp. 2958–1796.

[21] M. Wand, M. Janke, and T. Schultz, "Investigations on speaking mode discrepancies in EMG-based speech recognition," in *Proceedings of Interspeech*, Florence, Italy, Aug 2011, pp. 601–604.

[22] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *Proceedings of ICASSP*, Calgary, AB, Canada, Apr 2018, pp. 6219–6223.

[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 *Proceedings of Interspeech*, Stockholm, Sweden, Aug 2017, pp. 3672–3676.

[24] I. Ibrahimov, G. Gosztolya, and T. G. Csapó, "Data augmentation methods on ultrasound tongue images for articulation-to-speech synthesis," in *Proceedings of SSW*, Grenoble, France, Aug 2023, pp. 230–235.

[25] A. H. Shandiz and L. Tóth, "Voice activity detection for ultrasound-based silent speech interfaces using convolutional neural networks," in *Proceedings of TSD*, Olomouc, Czech Republic, Sep 2021, pp. 499–510.

[26] L. Tóth, A. H. Shandiz, G. Gosztolya, and T. G. Csapó, "Adaptation of tongue ultrasound-based Silent Speech Interfaces using Spatial Transformer Networks," in *Proceedings of Interspeech*, Dublin, Ireland, Aug 2023, pp. 1169–1173.

[27] 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.

[28] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Proceedings of SLTU*, Hanoi, Vietnam, May 2008, pp. 63–68.

[29] R. Peinl and J. Wirth, "Quality assurance for speech synthesis with ASR," in *Proceedings of IntelliSys*, Amsterdam, The Netherlands, Sep 2022, pp. 739–751.

[30] D. Alharthi, R. S. Sharma, H. Dhamyal, S. Maiti, B. Raj, and R. Singh, "Evaluating speech synthesis by training recognizers on synthetic speech," in *Proceedings of Interspeech*, Kos Island, Greece, Sep 2024, pp. 66–70.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of ICML*, Hawaii, USA, Jul 2023, pp. 28 492–28 518.

[32] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane, "Careless Whisper: Speech-to-text hallucination harms," in *Proceedings of FACCT*, Rio de Janeiro, Brazil, Jun 2024, pp. 1672–1681.

[33] T. Neuberger, D. Gyarmathy, T. E. Gráczi, V. Horváth, M. Gósy, and A. Beke, "Development of a large spontaneous speech database of agglutinative Hungarian language," in *Proceedings of TSD*, Brno, Czech Republic, Sep 2014, pp. 424–431.