INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Investigating Speech Reconstruction for Laryngectomees for Silent Speech Interfaces

Beiming Cao[1,2], *Nordine Sebkhi*[3], *Arpan Bhavsar*[3], *Omer T. Inan*[3], *Robin Samlan*[4], *Ted Mau*[5], *Jun Wang*[2,6]

[1]Department of Electrical and Computer Engineering, University of Texas at Austin, USA
[2]Department of Neurology, Dell Medical School, University of Texas at Austin, USA
[3]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
[4]Department of Speech, Language, and Hearing Sciences, University of Arizona, USA
[5]Department of Otolaryngology, UT Southwestern Medical Center, USA
[6]Department of Speech, Language, and Hearing Sciences, University of Texas at Austin, USA
jun.wang@austin.utexas.edu

Abstract

Silent speech interfaces (SSIs) are devices that convert non-audio bio-signals to speech, which hold the potential of recovering quality speech for laryngectomees (people who have undergone laryngectomy). Although significant progress has been made, most of the recent SSI works focused on data collected from healthy speakers. SSIs for laryngectomees have rarely been investigated. In this study, we investigated the reconstruction of speech for two laryngectomees who either use tracheoesophageal puncture (TEP) or electro-larynx (EL) speech as their post-surgery communication mode. We reconstructed their speech using two SSI designs (1) real-time recognition-and-synthesis and (2) directly articulation-to-speech synthesis (ATS). The reconstructed speech samples were measured in subjective evaluation by 20 listeners in terms of naturalness and intelligibility. The results indicated that both designs increased the naturalness of alaryngeal speech. The real-time recognition-and-synthesis design obtained higher intelligibility in electro-larynx speech as well, while the ATS did not. These preliminary results suggest the real-time recognition-and-synthesis design may have a better potential for clinical applications (for laryngectomees) than ATS.

Index Terms: silent speech interfaces, alaryngeal speech

1. Introduction

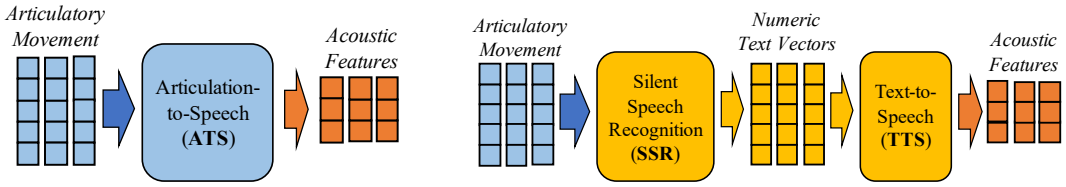
Silent speech interfaces (SSIs) are devices that enable speech communication when audible speech signals are unavailable [1, 2, 3], which have the potential of recovering quality speech for people who are unable to produce speech sounds but can still articulate (e.g., laryngectomees). Currently, laryngectomees primarily use three types of alaryngeal speech for their daily communication: electro-larynx (EL) [4, 5], tracheo-esophageal puncture (TEP) speech [6], and esophageal speech [7]. An electro-larynx (EL) is a battery-powered device, which adopts an electro-mechanical device to produce either pharyngeal or oral cavity vibrations [5]. TEP speech requires an additional surgery that makes a valve between the trachea and esophagus, which allows airflow from trachea to drive the vibration of throat wall [6]. Esophageal speech involves ingesting air into the esophagus and then expelling it to drive throat wall vibration to produce sound [8], which is difficult to learn [4, 8]. Therefore, TEP and EL are more commonly used. These approaches allow laryngectomees for daily communication but generate unnatural-sounding speech, which may discourage the willingness of communication and cause social isolation [9].

Currently, there are two typical designs of an SSI. The first is "recognition-and-synthesis" [10, 11, 12], which consists of two components: silent speech recognition (SSR) [13, 14, 15, 16] and text-to-speech synthesis (TTS) [17, 18]. SSR converts the articulatory information to text, and then the TTS component synthesizes the recognized text to speech. The second design is articulation-to-speech (ATS) synthesis, which directly maps articulatory information to speech [19, 20, 21, 22, 23]. ATS has a few advantages over the traditional SSR+TTS approach including easy implementation and low-latency. Traditional SSR+TTS takes the whole sentence as input rather than frame-level processing in ATS [10, 11, 12]. Thus, SSI researchers mainly focused on the ATS design recently [19, 24, 20, 22, 25]. ATS using multiple types of articulatory input have been shown to be able to generate natural and intelligible speech for healthy people [19, 24, 20, 22].

However, despite the studies of SSR for alaryngeal speech [15, 13]. SSI models (with speech synthesis) have rarely been studied for laryngectomees, due to lack of natural-sounding speech data from them. With the objective of generating natural and intelligible speech for laryngectomees, reconstructing their speech with other healthy speech data could be a solution [26]. The SSR+TTS design is known to be able to generate quality speech without requiring audio data from the users [10, 11, 12], thus could be more suitable for clinical applications if its high-latency issue could be resolved.

In this study, we proposed a real-time "recognition-and-synthesis" model (we call it RT-SSR+TTS) design and evaluated it with alaryngeal speech. This design was implemented in a frame-streaming manner to maintain a low-level latency. The main idea is to map the input articulatory data to frame-level numeric text vectors, and then the recognized text frames will be used for speech synthesis. In this design, the TTS model was trained with healthy speech, and the SSR model was trained or adapted with the alaryngeal speech. We also implemented ATS and compared it with the proposed design. The ATS was trained with healthy speech and directly tested with alaryngeal speech. The data used in this study include the electromagnetic articulograph (EMA) data collected from two alaryngeal speakers (one use TEP speech, the other use EL speech), and the mngu0 EMA dataset (healthy speech) [27]. The synthesized speech samples of laryngectomees were measured subjectively with the

http://dx.doi.org/10.21437/Interspeech.2021-1842



(a) *Articulation-to-speech synthesis* (b) *Real-time silent speech recognition and text-to-speech*

Figure 1: *ATS design (a) and the proposed RT-SSR+TTS design (b)*

comparative mean opinion scores (CMOS) on naturalness and intelligibility compared with the original alaryngeal speech by multiple listeners. The potential of the two types of SSI designs for laryngectomees was discussed.

2. Dataset

2.1. Alaryngeal speech data

The alaryngeal speech data was collected from two male English speakers, both of them undergone total laryngectomy. One uses tracheo-esophageal puncture (TEP) speech (age: 57), the other uses electro-larynx (EL) speech (age: 71). During the data collection, the EL speech subject read a list of 132 sentences twice. The TEP speech subject read it only once due to fatigue. The phrase list was selected from the frequently used sentences of AAC devices [28] users.

The electromagnetic articulograph (EMA) data was recorded simultaneously with the audio data (22050 Hz) by the Wave system (Northern Digital Inc., Waterloo, Canada). The audio data was downsampled to 16000 Hz for the experiments. The 3D articulatory movements (up-down, anterior-posterior, and left-right) of tongue tip (TT), tongue back (TB), upper lip (UL) and lower lip (LL) were tracked by the sensors attached to them. The sampling rate of EMA data is 100 Hz. Only the up-down and anterior-posterior dimension of the recorded data was used in this study. The EMA data were upsampled to 200 Hz to match the healthy speech data (mngu0) [27].

2.2. Healthy speech data

The mngu0 dataset is a corpus of articulatory data of different forms acquired from one male English speaker [27]. In this paper, we use the EMA subset of mngu0 that consists of audio and EMA data with 1,354 sentences recorded using Carstens EMA AG500 [29]. The total length of the speech data is about 67 mins [27]. The raw EMA data of mngu0 dataset tracks 12 sensor coils in 3D space with two angles of rotation [27]. In this study, we used 2D movement tracks (up-down, anterior-posterior) of 4 sensors attached to the same articulatory flesh points as laryngectomees (TT, TB, UL and LL). The sampling rate of EMA data is 200 Hz. The audio data was recorded in a sampling rate of 16000 Hz [27].

3. Methods

3.1. Articulation-to-speech synthesis (ATS)

The articulation-to-speech (ATS)-based SSI design is shown in Figure 1(a). ATS has been shown to be effective in SSI applications on healthy speech by multiple studies [20, 21, 22, 19]. The main advantages of ATS are easy-implementation and low-latency. The frame-level latency allows ATS to be a real-time system. Given enough training data, the synthetic speech could achieve high intelligibility [30].

3.2. The proposed design (real-time SSR+TTS)

Figure 1(b) is the proposed real-time "recognition-and-synthesis" design based on an SSR and a TTS component. This model was implemented in a frame-streaming manner to maintain a low-latency model. Rather than traditional SSR that coverts articulatory to text [11, 10], the real-time SSR (RT-SSR) convert articulation movement frames into numeric textual frames (the one-hot encoding of phonemes). Then the TTS component directly converts the recognized text frames to speech. In this design, the speech (audio) data from the users are not indispensable. The main challenge is developing an RT-SSR models for the user, then a TTS for other speakers could be used for speech generation. Therefore, as discussed, this SSI design may be more suitable for people who already lost their voice (laryngectomees).

3.3. Data preparation

Acoustic features: ATS and TTS have acoustic features as the output. In this study we use World vocoder [31] to convert the

Table 1: *LSTM models.*

Acoustic Feature	187-dim. vectors
Mel-generalized cepstrals (MGCs)	(60-dim. vectors) + $\Delta + \Delta\Delta$ (180-dim.)
Band aperiodicities (BAPs)	(1-dim. vectors) + $\Delta + \Delta\Delta$(3-dim.)
Fundamental frequency on log scale (log-f0)	(1-dim. vectors) + $\Delta + \Delta\Delta$ (3-dim.)
Voiced/unvoiced (V/UV) label	(1-dim.)
Sampling rate	16000 Hz
Windows length	25 ms
Articulatory movement	24-dim. vectors (4 sensors) $\times$ (2-dim. vectors) + $\Delta + \Delta\Delta$
Text Labels	48-dim. vectors One-hot encoding of 48 phonemes
LSTM Topology	
Input	ATS: 24-dim articulatory SSR: 24-dim. articulatory TTS: 48-dim. text
Output.	SSR: 48-dim. text (softmax layer) ATS: 187-dim. acoustic TTS: 187-dim. acoustic
Hidden-dim.	256
Depth	SSR: 4 hidden layers ATS: 3 hidden layers TTS: 3 hidden layers
Batch size	Sentence lengths
Dropout	SSR: 0.5 (after LSTM)
Max Epochs	Fine-tune SSR: 40 Other models: 50
Early stop	True (patient 3)

predicted acoustic features to speech samples. Same as in the Merlin speech synthesis toolkit [32], the acoustic features were extracted from audio data in a step size of 5ms, including 1-dimensional logarithm of the fundamental frequency (log-F0), 1-dimensional band aperiodicities (BAP) [33], 60-dimensional Mel-generalized cepstrals (MGCs) (warping factor: $\alpha = 0.41$), and 1-dimensional voiced/unvoiced label (V/UV). The first- and second-order derivatives of log-F0, BAP, and MGCs were concatenated to them as the output. Therefore, the dimension of model output is 187 ($(1 + 1 + 60) \times 3 + 1 = 187$).

Articulatory movement is the input of ATS and SSR, which is 2D movement tracks (in millimeter) of 4 sensors (TT, TB, UL and LL), with their first and second order derivatives, therefore the dimension of ATS and SSR inputs is 24 ($4 \times 2 \times 3 = 24$). Before the experiments, Procrustes matching was applied to remove the locational and rotational effects in these articulatory flesh point data [15].

Text labels are the output of SSR and the input of TTS. For mngu0 data and the EL speech data, the alphabetic phoneme labels of the sentences were aligned to their audio samples using the Festival speech synthesis system [34]. The articulatory data was synchronously recorded with the audio, thus each data frame (acoustic and articulatory) was labeled with a phoneme. There are 48 phonemes represent the textual information includes: the 39 English phonemes in the CMU dictionary, silence ("sil"), and extra phonemes: [*"ax", "axr", "dx", "el","em", "en","hv", "nx"*], which are same to that used in the Merlin and HTS toolkits [32, 35]. The phoneme frame labels were finally converted to 48-dimensional numeric vectors which indicate the phoneme of current frames (one-hot encoding, one for the current phoneme, zeros for the other 47 phonemes).

For the TEP speech data, due to the speech quality, the auto-alignment method generated severe shifts between the real and the aligned time stamps. Therefore, the aligned textual labels of TEP speech were generated by manually tagging in a phoneme level.

3.4. Experimental Setup

The ATS, SSR, and TTS models were implemented with long short-term memory-recurrent neural networks (LSTM-RNNs). Table 1 summarized the detailed setup of the LSTM models. The mngu0 dataset was separated into a training set of 1,226 sentences, a validation set of 63 sentences, and a testing set of 65 sentences. For alaryngeal speech, 100 sentences of the 132-sentence list were used for training, 12 sentences for validation and 20 for testing. As mentioned, the EL subject read the list twice thus the sentence numbers were doubled in the training (200), validation (24) and testing (40) set of EL speech. All the articulatory data was z-score normalized with the mean and standard deviation from their training sets.

The experiments were conducted as following: 1) developing SSI models in both of the two designs with the training and validation set of the healthy speech (mngu0) data, then test the models with the mngu0 testing set. 2) Directly testing the ATS and RT-SSR+TTS models (trained with mngu0 data) with the testing sets of alaryngeal speech data. 3) Improving the RT-SSR+TTS model in 2) by using the training and validation set of alaryngeal speech data. Two SSR training strategies were validated here: training with alaryngeal speech only, and using alaryngeal speech to fine-tune the SSR trained with mngu0.

Table 2: *Objective and subjective measures of SSIs for the mngu0 dataset (N: naturalness, I: intelligibility). The TTS results are the TTS with the true textual input as a reference.*

	MCD (dB)	BAP (dB)	F0 (Hz)	V/UV (%)	MOS (N)	MOS (I)
ATS	5.21	0.16	9.70	13.98	3.98	3.97
RT-SSR +TTS	5.82	0.17	10.69	15.88	3.43	3.54
TTS	4.89	0.14	9.46	7.55	-	-

3.5. Outcome Measures

The synthesized speech samples from healthy speech (mngu0) were measured both objectively and subjectively. The objective measures are the distortions or errors of the acoustic features prediction include: Mel-cepstral distortion (MCD) [36], distortion of band aperiodicities (BAPs), RMSE of the fundamental frequencies (F0-RMSE), and voiced/unvoiced error rate (V/UV). We also used the accuracies of SSR as an intermediate measure of the RT-SSR+TTS model. Here, the accuracies of SSR are the number of correctly predicted textual frames divided by the total number of frames tested. The subjective measurements are the mean opinion scores (MOS) in naturalness and intelligibility on 20 of the tested sentences, from 20 listeners. The responses were a 5-point scale of "excellent" 5, "good" 4, "fair" 3, "poor" 2, and "bad" 1. The number of total testing trials is: 20 samples $\times$ 2 SSI designs $\times$ 20 listeners = 800.

For laryngectomee speech, we only measured the output speech samples subjectively, since no original healthy speech samples are available for objective measurement. For the RT-SSR+TTS model here, same as the mngu0 data, we used the accuracies of SSR as an intermediate measure of RT-SSR+TTS in addition to the subjective evaluation. In subjective evaluation, the same listeners compared the synthesized speech samples with the original alaryngeal speech by giving comparative mean opinion scores ("definitely better" +2, "better" +1, "same" 0, "worse" -1, and "definitely worse" -2). The number of total comparison pairs is: 20 samples $\times$ 2 SSI designs $\times$ 2 alaryngeal speech $\times$ 20 listeners = 1600.

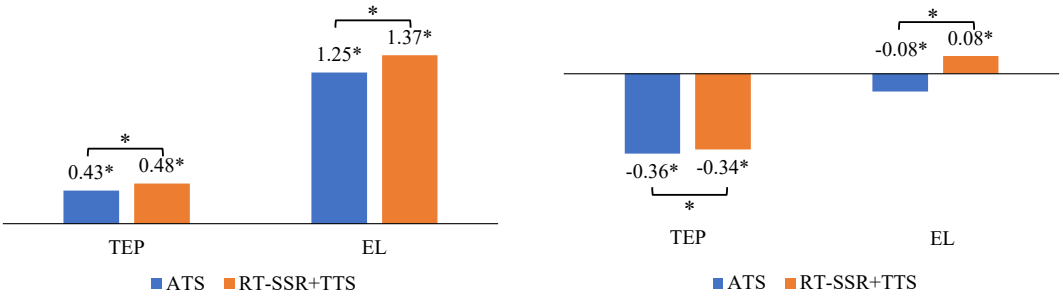
4. Results and Discussion

4.1. Results for the healthy subject

Table 2 has shown the results of SSI for the mngu0 data. The second row is the RT-SSR+TTS with an SSR that achieved an accuracy of 69.06%. The left four columns are the objective measures and the right 2 columns are the subjective measures. The ATS (first row) outperformed RT-SSR+TTS in all measurements (objectively and subjectively), which is expected due to the misrecognitions of SSR. The subjective testing results indicated that the synthetic samples from the both SSIs achieved satisfaction between "fair" (3) and "good" (4) in naturalness and intelligibility. The third row in Table 2 (TTS) shows the performance of the TTS given true textual input, which is significantly higher than ATS in objective measures. Therefore, we think the RT-SSR+TTS has a potential of outperforming ATS model if SSR accuracies could be improved.

4.2. Results for the alaryngeal speech

Table 3 gives the accuracies of SSR for alaryngeal speech using different training strategies. The first row gives the accuracies of SSR that was trained with the mngu0 data set. The testing accuracies are 18.23% and 13.43% for EL and TEP speech, re-



(a) *Naturalness* (b) *Intelligibility*

Figure 2: *CMOS test result of comparing the SSI synthesized to alaryngeal speech in terms of naturalness and intelligibilities. Asterisks indicated the significance of difference ($p < 0.01$). Positive values mean better than the alaryngeal speech.*

Table 3: *The frame-level accuracies of RT-SSR trained with different datasets and tested with alaryngeal speech.*

Training data	Test EL (%)	Training data	Test TEP (%)
mngu0	18.23	mngu0	13.43
EL	27.41	TEP	12.23
mngu0+EL	45.84	mngu0+TEP	28.47

spectively. The second row are the accuracies of SSR trained with alaryngeal speech only. Here the EL testing accuracy was increased to 27.41%, and the TEP accuracy was decreased to 12.23%. The best performance was obtained when training the SSR models with both mngu0 and alaryngeal speech data (45.84% for EL and 28.47% for TEP). The SSR models were firstly trained with mngu0 data, then fine-tuning on the alaryngeal speech. These SSR models (with the highest accuracies) and a TTS trained with only mngu0 were used to generate speech samples for the subjective listening tests. The EL has higher SSR accuracies than the TEP, which may due to the larger dataset from the EL subject. Also their similarities to the mngu0 data in articulation may affect the SSR accuracies.

Figure 2 shows the subjective listening test results of the alaryngeal speech. Two-tailed t-tests were performed between the CMOS scores and zeros (no preference), also between the two types of SSIs. The results with significant difference were marked with asterisk stars ($p < 0.01$). For naturalness, both of the two SSIs in this study significantly improved the naturalness of alaryngeal speech (Figure 2a). The intelligibilities were decreased for TEP speech by both SSI models, while the EL speech intelligibility was increased when using RT-SSR+TTS design (Figure 2b). When comparing the two types of SSIs, the RT-SSR+TTS design outperformed ATS for both two alaryngeal speech (in both naturalness and intelligibility).

4.3. Discussion

The experimental results indicated that both types of SSI improved EL speech more than TEP speech. TEP speech typically is more intelligible and natural-sounding than the EL speech [9]. Therefore, outperforming TEP speech might be a more challenging task for SSIs. The intelligibility was improved for EL speech by RT-SSR+TTS, but not for TEP speech. As shown in table 3, the SSR accuracies of TEP speech are much lower than EL speech, which may led to the lower performance of RT-SSR+TTS for TEP speech. In this early stage investigation, to maintain the real-time implementation, the RT-SSR was a simplified speech recognition model of frame-level phoneme classification. Other ASR components such as language models have not be integrated, which could further improve the SSR performance. More studies are needed in improving the SSR with the real-time implementation.

ATS outperformed RT-SSR+TTS for healthy speech (Table 2), which is expected. However, RT-SSR+TTS has shown higher performance than ATS for alaryngeal speech. Also the RT-SSR+TTS design could be further improved by collecting larger alaryngeal articulatory data to improve the SSR. This characteristics made the RT-SSR+TTS model more tractable and has better potential of recovering laryngectomee speech than the ATS model. In addition, the TTS component could be further improved by using a few pre-surgery voice samples (if available) from laryngectomees to improve the voice identity [37]. Selected speech samples of this study are available at [38].

Limitations: The dataset used in this study is small. Only one healthy and two alaryngeal speakers were used in this study. The speaker variation in articulation across the speakers may affect the results when generalizing to other subjects.

5. Conclusions and Future Work

In this study, we investigated two types of silent speech interface (SSI) designs for alaryngeal speech. Their performance of reconstructing speech for two alaryngeal speech subjects were evaluated and compared. The results have demonstrated that both of the two SSI designs generated more natural-sounding speech than alaryngeal speech. Also the intelligibility of electro-larynx speech was increased by the proposed real-time recognition-and-synthesis (RT-SSR+TTS) design. The RT-SSR+TTS design has shown higher performance than articulation-to-speech (ATS) for alaryngeal speech users, also better potential. Therefore, although ATS have shown high performance in healthy speech reconstruction recently, the RT-SSR+TTS design may have better potential for alaryngeal speech. Future works include validating these findings with more subjects (both healthy and alaryngeal) and larger dataset.

6. Acknowledgements

This work was supported by the National Institutes of Health (NIH) under award numbers R03DC013990 and R01DC016621 and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We also thank Avery Singson for helping with data preprocessing, and the volunteering participants.

7. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent Speech Interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based Spoken Communication: A Survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[3] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín-Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent Speech Interfaces for Speech Restoration: A Review," *IEEE Access*, pp. 177 995 – 178 021.

[4] H. Liu and M. L. Ng, "Electrolarynx in Voice Rehabilitation," *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332, 2007.

[5] R. Kaye, C. G. Tang, and C. F. Sinclair, "The Electrolarynx: Voice Restoration after Total Laryngectomy," *Medical Devices (Auckland, NZ)*, vol. 10, pp. 133–140, 2017.

[6] M. I. Singer and E. D. Blom, "An Endoscopic Technique for Restoration of Voice after Laryngectomy," *Annals of Otology, Rhinology & Laryngology*, vol. 89, no. 6, pp. 529–533, 1980.

[7] H. Nijdam, A. Annyas, H. Schutte, and H. Leever, "A New Prosthesis for Voice Rehabilitation after Laryngectomy," *Archives of oto-rhino-laryngology*, vol. 237, no. 1, pp. 27–33, 1982.

[8] J. P. Gleysteen, D. A. Elliott, and D. R. Clayburgh, "40 - Advanced Larynx Cancer," in *Oral, Head and Neck Oncology and Reconstructive Surgery*, R. B. Bell, R. P. Fernandes, and P. E. Andersen, Eds. Elsevier, 2018, pp. 818–829.

[9] T. L. Eadie, D. Otero, S. Cox, J. Johnson, C. R. Baylor, K. M. Yorkston, and P. C. Doyle, "The Relationship between Communicative Participation and Postlaryngectomy Speech Outcomes," *Head & neck*, vol. 38, no. S1, pp. E1955–E1961, 2016.

[10] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[11] J. Wang, A. Samal, and J. R. Green, "Preliminary Test of a Real-Time, Interactive Silent Speech Interface Based on Electromagnetic Articulograph," *Proceedings of the 5th SLPAT*, pp. 38–45, 2014.

[12] J. Wang, "Demo of a Real-time, Interactive Silent Speech Interface," Jul 2015. [Online]. Available: https://www.youtube.com/watch?v=23RxvetISac

[13] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent Speech Recognition as an Alternative Communication Device for Persons with Laryngectomy," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 12, pp. 2386–2398, 2017.

[14] S. Hahm, J. Wang *et al.*, "Silent Speech Recognition from Articulatory Movements Using Deep Neural Network," in *Proc. of the International congress of phonetic sciences*, 2015, pp. 1–5.

[15] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-Independent Silent Speech Recognition from Flesh-point Articulatory Movements Using an LSTM Neural Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2323–2336, 2017.

[16] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (Silent) Speech Recognition System for Patients Following Laryngectomy," *Medical engineering & physics*, vol. 30, no. 4, pp. 419–425, 2008.

[17] R. W. Sproat and J. P. Olive, "Text-to-Speech Synthesis," *AT&T technical journal*, vol. 74, no. 2, pp. 35–44, 1995.

[18] M. S. Heiga Zen, Andrew Senior, "Statistical Parametric Speech Synthesis Using Deep Neural Networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7962–7966.

[19] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-Based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis," *Proc. Interspeech*, pp. 2727–2731, 2020.

[20] J. Gonzalez Lopez, L. A. Cheah, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Evaluation of a Silent Speech Interface based on Magnetic Sensing and Deep Learning for a Phonetically Rich Vocabulary," in *Proceedings of Interspeech*. ISCA, 2017, pp. 3986–3990.

[21] C. T. Kello and D. C. Plaut, "A Neural Network Model of the Articulatory-Acoustic forward Mapping Trained on Recordings of Articulatory Parameters," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2354–2364, 2004.

[22] B. Cao, M. Kim, J. R. Wang, J. Van Santen, T. Mau, and J. Wang, "Articulation-to-Speech Synthesis Using Articulatory Flesh Point Sensors' Orientation Information," in *Proceedings of INTERSPEECH*, vol. 2018, 2018, pp. 3152–3156.

[23] N. Kimura, M. Kono, and J. Rekimoto, "SottoVoce: an Ultrasound Imaging-based Silent Speech Interaction Using Deep Neural Networks," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11.

[24] 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, pp. 67–87, 2016.

[25] L. Diener, G. Felsch, M. Angrick, and T. Schultz, "Session-Independent Array-based EMG-to-Speech Conversion Using Convolutional Neural Networks," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[26] T. Dinh, A. Kain, R. Samlan, B. Cao, and J. Wang, "Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency," *Proc. Interspeech*, pp. 4781–4785, 2020.

[27] K. Richmond, P. Hoole, and S. King, "Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus," in *Interspeech*, 2011, pp. 1505–1508.

[28] D. R. Beukelman, P. Mirenda *et al.*, *Augmentative and Alternative Communication*. Paul H. Brookes Baltimore, 1998.

[29] M. Stella, A. Stella, F. Sigona, P. Bernardini, M. Grimaldi, and B. G. Fivela, "Electromagnetic Articulography with AG500 and AG501." in *Interspeech*, 2013, pp. 1316–1320.

[30] B. Cao, B. Y. Tsang, and J. Wang, "Comparing the Performance of Individual Articulatory Flest Points for Articulation-to-Speech Synthesis," *ICPhS*, pp. 3041–3045, 2019.

[31] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[32] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *SSW*, 2016, pp. 202–207.

[33] M. Morise, "D4C, a Band-Aperiodicity Estimator for High-Quality Speech Synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[34] A. Black, P. Taylor, R. Caley, and R. Clark, "The Festival Speech Synthesis System," 1998.

[35] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based Speech Synthesis System Applied to English," in *IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.

[36] R. Kubichek, "Mel-Cepstral Distance Measure for Objective Speech Quality Assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.

[37] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting New Speakers based on a Short Untranscribed Sample," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3683–3691.

[38] B. Cao, "Demo of Speech Reconstruction for Laryngectomee for Silent Speech Interfaces," June 2021. [Online]. Available: https://beimingcao.github.io/IS2021demo/