Interspeech 2022
18-22 September 2022, Incheon, Korea

Data Augmentation for End-to-end Silent Speech Recognition for Laryngectomees

Beiming Cao[1,2], *Kristin Teplansky*[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 Speech, Language, and Hearing Sciences, 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 Neurology, Dell Medical School, University of Texas at Austin, USA
jun.wang@austin.utexas.edu

Abstract

Silent speech recognition (SSR) predicts textual information from silent articulation, which is an algorithm design in silent speech interfaces (SSIs). SSIs have the potential of recovering the speech ability of individuals who lost their voice but can still articulate (e.g., laryngectomees). Due to the logistic difficulties in articulatory data collection, current SSR studies suffer limited amount of dataset. Data augmentation aims to increase the training data amount by introducing variations into the existing dataset, but has rarely been investigated in SSR for laryngectomees. In this study, we investigated the effectiveness of multiple data augmentation approaches for SSR including consecutive and intermittent time masking, articulatory dimension masking, sinusoidal noise injection and randomly scaling. Different experimental setups including speaker-dependent, speaker-independent, and speaker-adaptive were used. The SSR models were end-to-end speech recognition models trained with connectionist temporal classification (CTC). Electromagnetic articulography (EMA) datasets collected from multiple healthy speakers and laryngectomees were used. The experimental results have demonstrated that the data augmentation approaches explored performed differently, but generally improved SSR performance. Especially, the consecutive time masking has brought significant improvement on SSR for both healthy speakers and laryngectomees.

Index Terms: silent speech recognition, silent speech interface, data augmentation, alaryngeal speech

1. Introduction

Laryngectomees are individuals who have their larynx partially or fully surgically removed (laryngectomy), to treat laryngeal cancer [1]. People with total laryngectomy lost the ability to produce audible speech. Currently, laryngectomees rely on alaryngeal speech for daily communication, which includes voice restoration options such as electro-larynx (EL) or tracheo-esophageal puncture (TEP) speech [2]. Both approaches generate unnatural-sounding speech, which increases communication apprehension and causes social isolation [3].

Silent speech interfaces (SSIs) are emerging technologies that convert biosignals derived from articulatory activity into audible speech. Such technologies have a strong potential to allow laryngectomees to produce natural-sounding speech [4, 5, 6]. There are currently two types of algorithmic designs in SSI: the recognition-and-synthesis design and the direct-synthesis design. The recognition-and-synthesis design consists of a silent speech recognition (SSR) model [7, 8, 9, 10, 11, 12] that converts the articulation to text, and a text-to-speech (TTS) [13, 14] model that converts the recognized text to speech. The direct-synthesis design directly maps articulatory information to speech [15, 16, 17, 18, 19]. The direct-synthesis design is relatively easier to implement and has low-latency; however, this approach requires synchronized or time-aligned articulatory and acoustic data from same [17] or different speakers [20]. Instead, the recognition-and-synthesis design allows an adoption of a well developed TTS model to convert the recognized text to audible speech [21]. Our previous study [21] indicated that the recognition-and-synthesis design may be more preferred than the direct-synthesis design for laryngectomees.

Although significant progress has been made in SSR [7, 8, 9, 10, 22, 11, 12], most related works were based on small-sized datasets (compared with acoustic data sets for acoustic speech recognition), as articulatory data collection requires more effort than audio speech data. Articulatory data from the alaryngeal speakers are especially limited [9, 23, 21, 24]. Meltzner et al. collected 980 phrases (sEMG data) from each of the eight laryngectomees participated [9], which is probably the largest articulatory data from laryngectomees, but still much smaller size than audio speech data (e.g., Librispeech corpus [25]).

Data augmentation [22] increases the data amount for training deep learning models by introducing variations to the existing data, which has been demonstrated effective in audio speech recognition [26]. Recently, Kimura et al. [22] has demonstrated the effectiveness of adopting data augmentation developed for audio speech recognition on the image-based silent speech recognition based on data from a single, healthy speaker. Data augmentation particularly for silent speech recognition has rarely been conducted, especially for laryngectomees.

In this study, we investigated multiple data augmentation approaches for SSR for laryngectomees, including consecutive and intermittent time masking, articulatory dimension masking, sinusoidal noise injection, and random scaling. These data augmentation approaches were directly applied to the raw articulatory signals. An end-to-end speech recognition model trained with the connectionist temporal classification (CTC) [27] was used. The performance of the SSR models were measured by phoneme error rates (PERs). The proposed data augmentation approaches were firstly investigated and compared

10.21437/Interspeech.2022-10868

on eight healthy speakers, then validated on two alaryngeal speakers (one uses electro-larynx, the other uses TEP speech), with different experimental setups (speaker-dependent, speaker-independent and speaker-adaptive [28]).

2. Dataset

2.1. Alaryngeal speech data

We collected alaryngeal speech data from two male American English speakers who have undergone total laryngectomy. One (age: 57) use tracheo-esophageal puncture (TEP) speech; the other (age: 71) uses electro-larynx (EL) speech. The phrase list of the stimuli was a list of 132 phrases, which were selected from a list of frequently used sentences designed for augmentative and alternative communication (AAC) devices [29] users. During the data collection session, the EL speaker read the list twice (264 recordings), the TEP speaker read it only once due to fatigue (132 recordings).

The electromagnetic articulography (EMA) data and audio were simultaneously recorded by the Wave system (Northern Digital Inc., Waterloo, Canada) [30]. Four sensors were attached to the tongue tip (TT), tongue back (TB), upper lip (UL), and lower lip (LL) of the speakers with dental glue (PeriAcryl 90, GluStitch) or tape. The 3D articulatory movements (superior-inferior, anterior-posterior, and left-right) of the sensors were collected in a sampling rate of 100 Hz. The trajectories of sensors have been filtered with a 20 Hz lowpass filter after recording. Only the superior-inferior and anterior-posterior dimension of the recorded data was used in this study [21].

2.2. Healthy speech data

The healthy speech dataset used in this study was the Haskins Production Rate Comparison [31] dataset, in which the EMA and audio data were recorded from 8 native American English speakers (4 males, 4 females). The stimuli is the 720 phonetically balanced Harvard sentences [32]. Each speaker read the list at least twice, one at a normal speaking rate, one at a fast speaking rate. They then read varying number of sentences at their habitual speaking rate. The fast speech data was excluded from the experiments in this study. We used all the remaining data (839 to 1019 recordings from each speaker).

The EMA data were also recorded with the NDI Wave system in a sampling rate of 100 Hz. A 20 Hz lowpass filtering was applied. The dataset provided 3D movement of 8 sensors: tongue tip (TT), tongue blade (TB), tongue rear (TR), upper lip (UL), lower lip (LL), mouth left (corner ML), jaw, and jaw left (canine) [31]. To match the collected alaryngeal speech dataset, we only use the superior-inferior and anterior-posterior dimensions of TT, TB, UL, and LL.

3. Methods

3.1. End-to-end silent speech recognition

End-to-end speech recognition skips the alignment stage in the conventional hidden Markov model-based hybrid ASR models, by leveraging the learning ability of the deep learning and the advanced techniques such as the connectionist temporal classification (CTC) loss function [27]. The end-to-end silent speech recognition (SSR) model in this study is a popular variation of the Deep Speech 2 [33, 34] model, which consists of two residual CNN blocks and two bidirectional-gated recurrent unit (BiGRU) layers, with an input CNN layer, a fully-connected (FC) layers between CNN and BiGRUs, and two FC layers in the end as the classifier (Table 1). CNNs were used as a feature extractor and BiGRUs were used for temporal modeling. The input of the model is the 2D movement tracks (in millimeters) of 4 sensors (TT, TB, UL, and LL), with their first and second order derivatives, therefore the dimension of inputs is 24 (4 sensors $\times$ 2D $\times$ 3 = 24). The output dimension of the SSR model is 41, which consist of the 39 phonemes from the CMU Pronouncing Dictionary, a silent phoneme, and a blank for CTC. The detailed experimental setup is shown in Table 1.

Similar to the Deep Speech 2, we use CTC loss function to train the model. The CTC loss function introduced an additional output dimension called "blank", which is a possible prediction of the model. With the predicted blanks, the output sequences could be collapsed into feasible or infeasible sequences. The collapsing works by removing the blank and merge the repeats in the output sequences [27]. During training, the CTC loss optimizes the model by maximizing the probability of predicting the feasible sequences [27]. In this study, our silent speech recognition models predicted phoneme sequences rather than character sequences. Since phonemes are a direct representation of the phonetics, it may be less challenging for SSR than the characters. The experimental results were measured using phoneme error rates (PERs), which were computed by the summation of the insertion, deletion and substitution errors divided by the total number of phoneme tested.

3.2. Data augmentation for silent speech recognition

The data augmentation in this study were performed in a on-the-fly manner. For each mini-batch during training, a certain ratio (augment ratio) of samples were randomly applied data augmentation transforms. All setup parameters (e.g., augmentation ratio and frequency of noise) for the data augmentation approaches were found in the preliminary experiments with the validation set of healthy speech data.

3.2.1. Consecutive time masking (CTM)

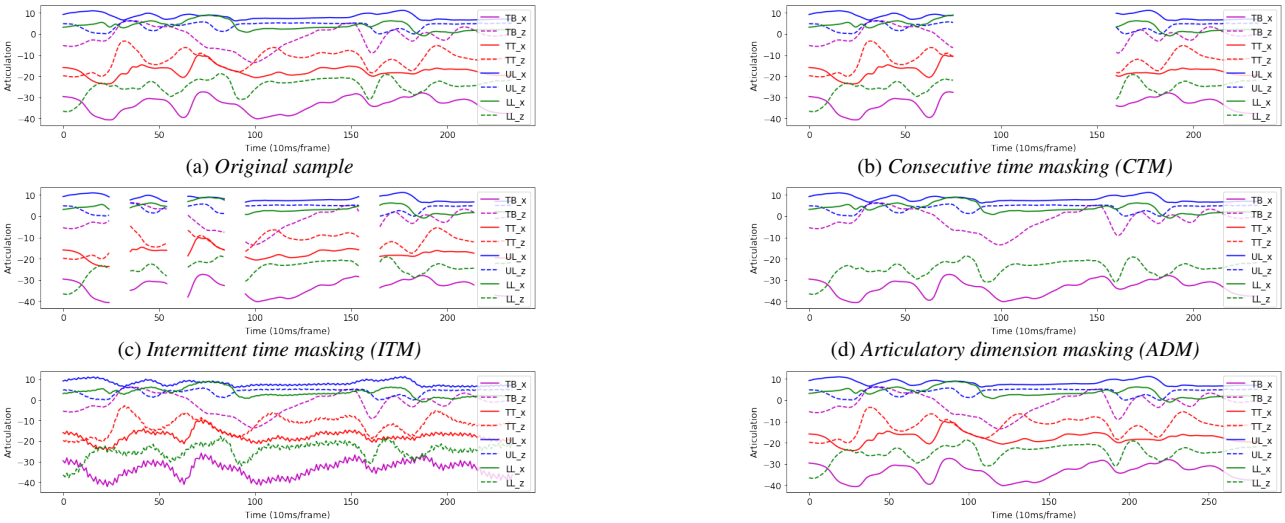
The consecutive time-masking data augmentation approach (Figure 1b) in this study was similar to the time masking in the SpecAugment [26]. Given an EMA speech sample, we masked a random number of consecutive frames to zeroes, the masking number was randomly selected from 0 to the maximum masking number. The starting point of masking was randomly chosen within the sample. The augment ratio for this approach was 0.8, which means 80% of the samples in the training set were masked during each training epoch. The maximum masking frame number was 80 (0.8s).

3.2.2. Intermittent time masking (ITM)

We also explored the intermittent time masking approach, in which the masking was performed on a few small intermittent segments, rather than a consecutive block (Figure 1c). A fixed number of starting points were randomly selected, then masked out fixed length of frames started from each of the starting points. It is possible that two masked blocks were consecutive, but no overlap between them. The augment ratio here was 0.7; the masking segment number was 5; the masked frame number in each segment was 10 (0.1s).

3.2.3. Articulatory dimension masking (ADM)

The articulatory dimension masking data augmentation approach (Figure 1d) was similar to the frequency masking in the



(a) *Original sample* (b) *Consecutive time masking (CTM)*

(c) *Intermittent time masking (ITM)* (d) *Articulatory dimension masking (ADM)*

(e) *Sinusoidal noise injection (SNI)* (f) *Random scaling (RS)*

Figure 1: *Examples of an EMA data sample before and after data augmentations.*

SpecAugment [26], in which we mask out a random consecutive number of articulatory movement (e.g., superior-inferior or anterior-posterior of tongue) from the articulatory dimension. The number of the masked dimensions was randomly chosen from 0 to the maximum masked dimension number, and the starting dimension of masking was randomly chosen from 0 to 24 (input dimension). The augment ratio was 0.7, the number of maximum masked dimension is 5 (out of 24-dim. input).

3.2.4. Sinusoidal noise injection (SNI)

We also explored using sinusoidal noise injection as a data augmentation. In this approach (Figure 1e), we added sinusoidal noise signals to each dimension of the articulatory movement. The amplitudes of the sinusoidal signals were computed using the mean amplitude of that dimension times a scaling factor. The frequency of the sinusoidal signal was a fixed number found during the preliminary exploration. The amplitude scaling factor was chosen as 0.05, the frequency of the noise was 40 Hz, phase was zero. For example, given a articulatory dimension with a mean amplitude of A, the noise signal added would be $I(t) = 0.05 \cdot A \cdot \sin(2\pi \cdot 40t)$. The augment ratio was 0.5.

3.2.5. Random scaling (RS)

A recent study [35] indicated that speaking rates are correlated with the performance of SSR. Therefore, we also explored a data augmentation approach of randomly scaling up or down the duration of the samples (Figure 1f). The scaling factor was chosen from a range of numbers. Varies of scaling ranges were explored as preliminary experiments, which led us to select a range from 0.8 to 1.2 for this approach. The augment ratio for this approach was 0.5.

4. Experimental Setup

We firstly validated the data augmentation approaches using the healthy speech data and then applied the best performed approach to the alaryngeal speech. Data from all speakers were separated into training, validation and testing sets. All the input articulatory data were z-score normalized with the mean and standard deviation calculated from their current training sets, which varied in different experiments.

To find the best data augmentation approach, for each of the healthy speakers, 50 sentences were used as testing set, the other 50 sentences were used as validation set, and the rest for training (739 to 919 sentences). Speaker-dependent (SD) and speaker-independent (SI) experiments were conducted for the healthy speech. The SI experiments for healthy speakers were leave-one-speaker-out-cross-validation, in which the models were trained/validated with the mixture of the training/validation sets from seven training speakers, then tested on the testing set of the left-out testing speaker.

Table 1: *Silent speech recognition model setups.*

Model	
Input Articulatory movement	**24-dim. vectors** 4 sensors × 2D = 8-dim. vectors + Δ + ΔΔ (24-dim.)
First CNN conv2d layer	**24-dim. (1-channel → 32-channel)** kernel = 3, padding = 1, stride = 2
Residual CNN × 2 layer norm + activation conv2d layer layer norm + activation	**24-dim. (32-channel → 32-channel)** activation: GeLU kernel = 3, padding = 1, stride = 1 activation: GeLU
Fully-connected layer linear layer	**24-dim. × 32-channel → 512-dim.** 768-dim. → 512-dim.
bidirectional GRU × 2 layer norm + activation Bidirectional GRU dropout	**512-dim. → 512-dim. × 2 (bidirectional)** activation: GeLU 512-dim. 0.3
Classifier linear layer 1 activation dropout linear layer 2	**512-dim. × 2 → 41-dim.** 1024-dim. → 512-dim. GeLU 0.3 512-dim. → 41-dim.
Output phoneme sequences	**41-dim. vectors** 39 phonemes + 1 silence + 1 blank
Loss function	**CTC loss (blank 41-dim.)**
Decoding	**Greedy deocoding**
Hyper-parameters	
batch size: 16 max Epochs: 80 early stop: True	optimizer: AdamW learning rate: 0.0005 patient: 10
Toolkit	Pytorch

Table 2: *PERs of the SSR experiments for healthy speech. CTM: Consecutive time masking; ITM: Intermittent time masking; ADM: articulatory dimension masking; SNI: sinusoidal noise injection; RS: random scaling.*

Healthy speech	Speaker-dependent PER (%)	p-value	Speaker-independent PER (%)	p-value
Baseline	39.31	—	60.06	—
CTM	**35.57**	**0.0008**	**53.71**	**0.0087**
ITM	37.13	0.146	58.35	0.233
ADM	36.07	0.158	57.66	0.0783
SNI	37.90	0.4472	58.46	0.3289
RS	37.38	0.2948	57.22	0.1386

To verify if the data augmentation approach selected by the previous experiment could be generalized to alaryngeal speech, the 132-sentence list was separated into a training list of 100 sentences a validation list of 12 sentences, and a testing list of 20 sentences. As introduced previously, the EL speaker read the sentence list twice. Therefore the EL speech has double amount of sentences in the training (200), validation (24), and testing (40) set than the TEP speech. The two alaryngeal speakers were measured and reported separately. Here, SD models were trained and tested on alaryngeal speech data. The SI models were trained and validated with all eight healthy speakers, then tested on the alaryngeal speakers. We also conducted speaker-adaptive (SA) experiments, in which the SSR models were first **pre-trained** with the eight healthy speakers, then **re-trained (fine-tuned, FT)** with the training sets from the alaryngeal speakers. This SA approach has been demonstrated to improve SSR performance in our previous study [21]. Experiments with or without data augmentations in these two training stages (i.e., pre-training and fine-tuning) were compared.

5. Results and Discussion

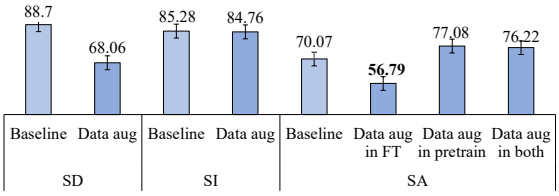
5.1. Healthy speech

Table 2 provides the average phoneme error rates (PERs) across the eight healthy speakers from the speaker-dependent (SD) and speaker-independent (SI) experiments. Two-tailed t-tests were performed to assess statistical significance of the PERs with and without data augmentation. In both SD and SI experiments, on average, the data augmentation approaches outperformed experiments without data augmentation. Additionally, the CTM showed a statistically significant improvement for both SD ($p < 0.001$) and SI setups ($p < 0.01$).

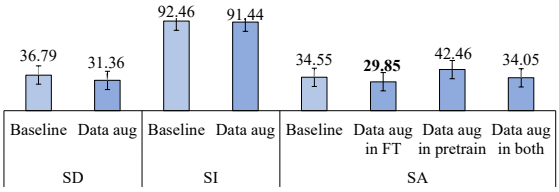
CTM outperformed other approaches possibly because it brought sufficient variation and maintained the articulation-to-text mapping in the samples. ITM introduced small intermittent masking, which could be compensated by the convolutions in CNN so that the variation was weakened. For ADM, the dependency among the articulation dimensions (e.g., tongue and lips) are not as strong compared to the temporal dependency in speech, which made it harder to learn the masked-out information. When injecting noise and random scaling, the variations were applied to the whole time-series samples, which might introduce excessively larger variations that impact the overall articulation-to-text mapping.

5.2. Alaryngeal speech

Figure 2 gives the experimental results on alaryngeal speech. Only the best performed data augmentation approach (CTM) was used here. The CTM improved the SD performance for the TEP and EL speakers by 20.64% and 5.43% respectively. However, this data augmentation approach only slightly improved the SI performance (trained on healthy speaker and tested on alaryngeal speakers). In speaker-adaptive (SA) experiments, applying CTM during the fine-tuning stage achieved the lowest PERs for alaryngeal speech in this study. When applying CTM on pre-training stage the performance decreased, yet slightly increased EL and decreased TEP performance when applying CTM on the both two stages.



(a) *Tracheo-esophageal puncture (TEP) speech*



(b) *Electro-larynx (EL) speech*

Figure 2: *PERs of SSR for alaryngeal speech.* ***Baseline****: no data augmentation.* ***Data aug in FT****: data augmentation in fine-tuning.* ***Data aug in Pretrain****: data augmentation in pretraining.* ***Data aug in both****: data augmentation in both pretraining and fine-tuning.*

Therefore, CTM appears to be an effective data augmentation for both healthy and alaryngeal speech, but not strong enough to compensate the variation gap between healthy and alaryngeal speech (SI). Applying data augmentation on alaryngeal speech data only is more effective in training SSR for alaryngeal speech (SD and SA). In this study, only data from one TEP speaker and EL speaker were used, and data size from each of them is small. Additional studies with a larger number of participants and data size are needed to verify these findings.

In this work, data augmentation strategies were applied on raw kinematic signals. It is unknown if they also work on extracted features or signals of other modalities (e.g., ultrasound and sEMG).

6. Conclusions and Future Work

In this study, we explored multiple data augmentation approaches for silent speech recognition (SSR). The experimental results demonstrated that the consecutive time masking (CTM) has brought higher improvement than other approaches, and CTM was effective for both healthy and alaryngeal speech. Additionally, applying data augmentation (CTM) on alaryngeal speech data only was more effective in training SSR for alaryngeal speech. Further studies with a larger number of alaryngeal speakers are needed to verify these findings.

7. Acknowledgements

This work was supported by the National Institutes of Health (NIH) under award number R01DC016621 (Wang). We also thank the volunteering participants.

8. References

[1] G. A. Gates, W. Ryan, J. Cooper Jr, G. F. Lawlis, E. Cantu, E. Lauder, R. W. Welch, and E. Hearne, "Current status of laryngectomee rehabilitation: I. results of therapy," *American journal of otolaryngology*, vol. 3, no. 1, pp. 1–7, 1982.

[2] B. J. Bailey, J. T. Johnson, and S. D. Newlands, *Head and Neck Surgery–Otolaryngology*. Lippincott Williams & Wilkins, 2006, vol. 1.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] S. Hahm and J. Wang, "Silent speech recognition from articulatory movements using deep neural network," in *Proc. of the International congress of phonetic sciences*, 2015, pp. 1–5.

[9] 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.

[10] 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 (TASLP)*, vol. 25, no. 12, pp. 2323–2336, 2017.

[11] S. Stone and P. Birkholz, "Cross-speaker silent-speech command word recognition using electro-optical stomatography," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7849–7853.

[12] C. Wagner, P. Schaffer, P. Amini Digehsara, M. Bärhold, D. Plettemeier, and P. Birkholz, "Silent speech command word recognition using stepped frequency continuous wave radar," *Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022.

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

[14] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[15] 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.

[16] J. A. Gonzalez, 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," *Proc. Interspeech 2017*, pp. 3986–3990, 2017.

[17] 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.

[18] 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.

[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. A. Gonzalez-Lopez, A. Gomez-Alanis, J. L. Pérez-Córdoba, and P. D. Green, "Non-parallel articulatory-to-acoustic conversion using multiview-based time warping," *Applied Sciences*, vol. 12, no. 3, p. 1167, 2022.

[21] B. Cao, N. Sebkhi, A. Bhavsar, O. T. Inan, R. Samlan, T. Mau, and J. Wang, "Investigating speech reconstruction for laryngectomees for silent speech interfaces," *Proc. Interspeech 2021*, pp. 651–655, 2021.

[22] N. Kimura, Z. Su, and T. Saeki, "End-to-end deep learning speech recognition model for silent speech challenge," in *INTERSPEECH*, 2020, pp. 1025–1026.

[23] A. Rameau, "Pilot study for a novel and personalized voice restoration device for patients with laryngectomy," *Head & Neck*, vol. 42, no. 5, pp. 839–845, 2020.

[24] J. M. Vojtech, M. D. Chan, B. Shiwani, S. H. Roy, J. T. Heaton, G. S. Meltzner, P. Contessa, G. De Luca, R. Patel, and J. C. Kline, "Surface electromyography–based recognition, synthesis, and perception of prosodic subvocal speech," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 6S, pp. 2134–2153, 2021.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[26] D. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," 09 2019, pp. 2613–2617.

[27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[28] X. Huang and K.-F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *IEEE Transactions on Speech and Audio processing*, vol. 1, no. 2, pp. 150–157, 1993.

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

[30] J. Berry, "Accuracy of the ndi wave speech research system," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 295–301, 2011.

[31] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017.

[32] Institute of Electrical and Electronics Engineers, "Ieee recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[33] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[34] B. F. Dossou and C. C. Emezue, "Okwugbe: End-to-end speech recognition for fon and igbo," *arXiv preprint arXiv:2103.07762*, 2021.

[35] L. Pandey and A. S. Arif, "Effects of speaking rate on speech and silent speech recognition," *CHI '22 Extended Abstracts*, pp. 1–8, 2022.