Silent speech recognition with articulator positions estimated from tongue ultrasound and lip video

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Silent Speech Recognition with Articulator Positions Estimated from Tongue Ultrasound and Lip Video

Rachel Beeson[1], *Korin Richmond*[2]

[1]University of Edinburgh, Scotland
[2]Centre for Speech Technology Research, University of Edinburgh, Scotland
R.Beeson@sms.ed.ac.uk, korin.richmond@ed.ac.uk

Abstract

We present a multi-speaker silent speech recognition system trained on articulator features derived from the Tongue and Lips corpus, a multi-speaker corpus of ultrasound tongue imaging and lip video data. We extracted articulator features using the pose estimation software *DeepLabCut*, then trained recognition models with these point-tracking features using *Kaldi*. We trained with voiced utterances, then tested performance on both voiced and silent utterances. Our multi-speaker SSR improved WER by 23.06% when compared to a previous similar multi-speaker SSR system which used image-based instead of point-tracking features. We also found great improvements (up to 15.45% decrease in WER) in recognition of silent speech using fMLLR adaptation compared to raw features. Finally, we investigated differences in articulator trajectories between voiced and silent speech and found that speakers tend to miss articulatory targets that are present in voiced speech when speaking silently.

Index Terms: silent speech interfaces, silent speech recognition, articulator pose estimation, ultrasound imaging, lip reading

1. Introduction

A device to allow speech(-like) communication without an audible speech signal is known as a silent speech interface (SSI). An SSI is intended to serve as a communication aid in situations where a user has either lost their ability to speak or is in an environment that is very noisy, or conversely where silence must be maintained [1]. Differing approaches have been explored to achieve this. Many methods, for example, assume the ability to move the articulators but that voicing is not possible. The communication that is restored by an SSI can take on different forms, such as a vocoder to restore the speech itself [2], or silent speech recognition (SSR) systems [3, 4]. The latter approach is the subject of this paper.

Much of the previous work on SSR has relied upon either point-tracking or imaging-based articulography. Electromagnetic articulography (EMA) is a well-known example of a point-tracking articulography technique, whereby sensor coils are attached directly to the articulators and their movements are tracked using a set of alternating electromagnetic fields. Meanwhile, examples of imaging-based articulography techniques would be ultrasound tongue imaging (UTI) and standard video of the mouth area. In UTI, a standard B-mode ultrasound probe placed submentally can show the surface of the tongue as a bright edge in video sequences. Numerous studies have used both these types of data (e.g., EMA [5] and optical/ultrasound video [6, 3], respectively).

Both methods have distinct advantages and disadvantages. Since UTI requires only an external probe to produce an image [7], it in principle offers a real-time SSI system that is lightweight and more convenient than using sensor technology like EMA [2] which is costly and invasive. However, while EMA can track specific articulator points over time reliably, ultrasound images may have artefacts like multiple or split visible tongue contour edges, frame discontinuities, or shadows from the jaw anatomy, which may obscure parts of the tongue [7]. The image is also speckled, which may introduce noise depending on the use case of the ultrasound data. An advantage of using EMA sensor tracking is that it is cleaner and easier to extract articulator movements. Because the sensors track movement in real Cartesian space, it can also be easier to calculate metrics relating to the position of the articulators over time. Ultrasound and other image data conversely have no direct tracking of the points of anatomy, and typically require some extraction or transformation to be performed in order to make it usable for study. Various methods have been used for this, including using the images themselves in some transformed way as an input to a model, or extracting features from the images in a more sophisticated feature extraction network [6, 8, 9], for example. Image-processing methods like edge detection, where the tongue contour is discovered by looking for a point of high contrast in the ultrasound image, are useful for some metrics, but because they rely on a contrastive edge which is present in the image, it is not completely reliable due to the noise [7].

In this work, we evaluate an approach to articulatory feature representation which is meant to combine the best of both worlds. We use an "off-the-shelf" pose estimation neural network model *DeepLabCut* (DLC) [10] to track specified points in video sequences. DLC implements "markerless pose estimation" - it is software which allows the user to train a model to find points of interest (indicated by a hand-marked training set) in a series of video frames. Previous work has been done to fine-tune DLC on ultrasound images of the tongue and video images of the lips [11], who used it to mark 22 points of anatomy: 14 on the tongue and 8 on the lips (see Fig. 1). The outputs of this model are thus x,y coordinates for articulatory points of interest. We then use these x,y coordinates as input to a speech recognition system. This method differs from other tongue contour extraction methods in that it does not rely on a high-contrast edge in the image, and is less susceptible to noise interruption. These articulator point estimates have been shown by [11] to be closer to those made by human hand-labellers than edge detection methods, and are also correlated with EMA sensor positions. Since they are derived from UTI data, it is possible to track points which may be difficult to track with a physical sensor (e.g. hyoid). DLC thus allows us to make use of convenient and cheap ultrasound and image data while preserving the Cartesian nature of sensor data. These attractive attributes

10.21437/Interspeech.2023-1966

motivate our investigation of using DLC features for an SSR system.

2. Dataset: Tongue and Lips corpus

The Tongue and Lips corpus (TaL) [12] contains the synchronized tongue ultrasound and lip video data from 82 different speakers of English. An ultrasound probe placed under the chin of a speaker captures a sagittal view of the tongue, while a camera placed in front of the mouth captures a frontal view of the lips. These recordings are captured during different kinds of utterances: spoken silently, spoken aloud, spontaneous speech, whispered speech, and swallows. Some utterances are spoken in both silent and voiced modalities, and some utterances are unique to a modality. The transcriptions of the utterances and the audio itself are also included. Sentences for read-speech utterances are taken from a variety of sources, including the Rainbow Passage, the Harvard Sentences, the TIMIT Corpus, the VCTK corpus, and the Librispeech corpus.

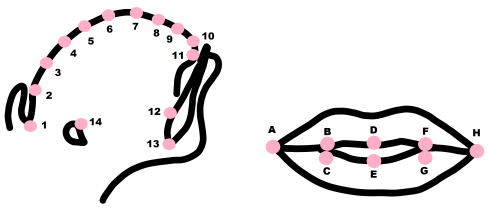
Most previous work on SSIs has resulted in speaker-dependent models due to the sort of challenge data available and the expense of collecting such data [8, 9, 5]. In contrast, the TaL corpus allows us to build and explore speaker-independent SSR systems. In Ribeiro et al. (2021) [3], a multi-speaker SSR system was built using data from the TaL corpus. The raw tongue and lip image data was used as input to a feature extraction network, where the targets were the time-aligned phone states from a monophone system trained using the audio data provided in TaL. A bottleneck layer was included in this extraction network, which was used as the input to an ASR system. In order to establish the overall improvement to a multi-speaker SSR using DLC features as opposed to bottleneck features derived from raw image data, we refer to their work as the basis for ours in making design decisions, and use their results as a baseline.

2.1. Silent speech versus normally-voiced speech

While the TaL corpus contains a large amount of data for voiced speech, it has less data for silent speech (11.06 hours compared to 2.34 hours). There is insufficient data to train a model on silent-speech data alone, and so we need to carefully consider the differences between silent and voiced speech and employ domain adaptation methods to obtain a stronger model.

Silent speech is defined by a lack of pulmonary airstream and laryngeal activity while maintaining articulatory activity [13]. Silent speech is therefore characterized by a lack of auditory feedback and a lack of intraoral pressure. When we speak aloud, there is evidence to suggest that we incorporate information about the auditory sensory information into corrective actions by articulators [14, 15]. Patients with cochlear implants will change their F0 when the implant is off versus on. The spectral characteristics of the vowels speakers produce when they lack auditory feedback changes as well. It has also been shown that speakers take different strategies with respect to speaking rate and articulatory space when producing silent speech compared to voiced speech [16, 17, 18, 19, 20]. These observations all indicate the TaL speakers are likely to have articulated differently when speaking without auditory feedback.

As a result of these different characteristics, speech recognition models which are trained on the articulator data of voiced speech may suffer performance losses when used to decode silent speech, because of the mismatches between the modalities. A systematic review of the TaL corpus has shown that overall, silent speech is hypoarticulated and produced at a slower rate [3]. However, that study also showed that these differences were not correlated with the WER of their speech recognition system. In order to gain insight into the nature of the articulator trajectory differences between modalities and how they affect our SSR system, we will analyse the articulator trajectories from corresponding silent and voiced utterances from the TaL corpus.



(a) *Tongue* (b) *Lip*

Figure 1: *Points of anatomy marked by DLC. The tongue points correspond to the vallecula (1), root (2-3), body (4-5), dorsum (6-7), blade (8-9), tip (10-11), short tendon (12), mandible (13), and hyoid (14). The midline of the lip (D,E) corresponds to the middle of the philtrum- the other points are midway between the midline and commissures (A,H).*

3. Data processing with DeepLabCut

The first step in building our speech recognition system is to extract articulator input features using DLC. DLC expects videos as input, whereas the UTI data in TaL exists as raw ultrasound scanline data. We used the UltraSuite Tools [21] to convert the raw TaL data to the required video format. As part of that process we downsampled the ultrasound data from 80Hz to 60Hz frame rate, which both matches the lip video frame rate and is also the frame rate DLC expects. Aligning the data streams at the same frame rate is a straightforward way to obtain a single feature vector corresponding to all articulator points at each time point. The tongue and lip videos were then run through DLC using the pre-trained articulator model [11]. The resulting output comprised two separate batches of csv files, one for the lip videos and one for the tongue videos. Each csv file contained three columns for each articulator: x position (in pixels), y position, and confidence about the prediction. The tongue and lip point data was then combined to give a single vector of all articulator coordinates per frame. These articulator points are illustrated in Fig. 1.

4. Experiments

We created a pipeline to train and test our features using Kaldi's nnet2 recipe [22, 23]. We followed a typical DNN-HMM pipeline for our models (see Fig. 2), similar to the one used by Ribeiro et al. (2021). We trained two DNN-HMM models for each condition, one trained on fMLLR-adapted features [24] and one trained on "raw" DLC articulator coordinate features. We chose fMLLR adaptation because it entails a model-based transformation of the features in terms of mean and variance. Therefore, fMLLR on the features of silent speech data when the features are the x,y coordinates of the articulator points is essentially a transform of all the articulator positions in space. Since previous work had shown differences in how silent speech behaves spatially (i.e. hypoarticulation), we believed this would make the silent features more like the voiced features. Our over-

all training process was as follows:

- Mean and variance normalization per utterance.
- Train an initial monophone model on these features and their transcriptions.
- Initialize a triphone system on the monophone alignments. Add delta values.
- Initialize a further triphone model on those alignments. The features this time were LDA+MLLT processed to reduce noise and normalize per speaker.
- Train a final triphone system on fMLLR features, initialized on the previous triphone features.
- These final triphone alignments were the gold labels for our DNN system. The DNN was trained on either fMLLR transformed features, or the unchanged feature vectors used on the initial monophone system.

These methods are commonly used in DNN-HMM systems to find the best state-frame alignments possible to be used as the gold labels for training the DNN system. As for our DNN system, we used 4 hidden layers of 1024 dimensions. Our input size was 44 * 4 frames on either side of the input. Our output size was 1832 states, for a total parameter size of 5.4M. We used a minibatch size of 128 and an initial learning rate of 0.01, final learning rate of 0.001. Decoding was done with a bigram language model. The probabilities of this language model were determined using all the possible sentences found in the TaL corpus. Likewise, the lexicon consisted only of words found in the TaL corpus. The language model was built using the SRILM toolkit [25], and then converted into FST format using Kaldi. The corresponding phones were determined using the BEEP[1] lexicon which is a British English lexicon. We chose this lexicon due to the majority of speakers having a British accent variety. All of our experiments were carried out on 2 NVidia TitanX GPUs with a runtime of between 12 and 18 hours for each model and corresponding test sets.

For our initial experiment, we trained our model using the x,y feature vectors as they were output from the DLC model. We did not include the confidence values and we did not remove or change any data based on confidence. We did not do any filtering of the articulator trajectories. This was in order to establish a baseline for DLC features to compare further results.

For our second experiment, we removed values from our feature vectors which were associated with a confidence value of less than 0.1. We determined this threshold by reviewing a small sample of videos, noting that when a body part is obscured by a shadow or is not present in the field of view, the confidence value drops below 0.1. Future experiments could tune the amount of confidence filtering as a hyperparameter. We then interpolated the articulator coordinate trajectories to fill in the gaps left behind by this removal, using a bi-directional linear method. Because of this data removal, some utterances had to be discarded, which occurred when there was not enough data left after confidence filtering to interpolate. Utterances removed in one test set were removed in the other in order to make consistent comparisons (Table 1).

In a third experiment, we similarly removed points which were more than 3 standard deviations away from the mean for a particular piece of anatomy for a particular utterance, and then interpolated over the missing values. This was meant to remove points which were discontinuous with the rest of the articulator behavior. We observed that generally, the points marked by DLC for a piece of anatomy were normally distributed.

[1]T. Robinson. (1996) Beep dictionary. Cambridge University. [Online]. Available: http://svr-www.eng.cam.ac.uk/comp.speech/ Section1/Lexical/beep.htm

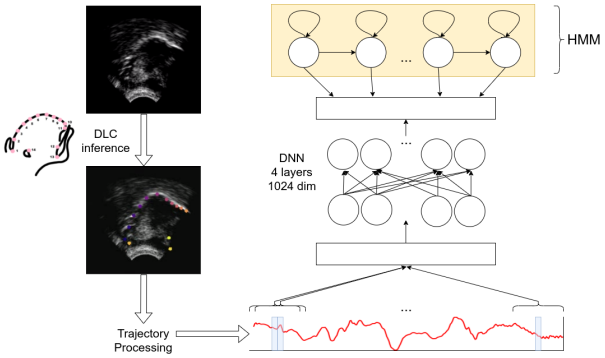


Figure 2: *Our processing pipeline and DNN-HMM system.*

	Train Utts	Test Utts	Speakers
Plain	13274	1205	81
+ Confidence filt	12803	1181	81
+ Outlier filt	12779	1179	81
+ Low pass filt	12779	1179	81

Table 1: *Test and training split counts for the four conditions.*

For a fourth experiment, we low-pass filtered the articulator trajectories at a rate of 20 Hz. We chose this value because the syllable production rate for adult speakers ranges from an average of around 3 syll/s towards a maximum possible 10 syll/s [26] and any motion shorter than this threshold would constitute noise introduced by the imprecision of the ultrasound image, or the DLC point estimation. We chose 20 Hz instead of 10 Hz as a safe buffer amount and to preserve any intra-syllabic effects which may be significant to determining phone identity (especially in the silent mode which has a greater number of articulatory sub-movements [18]). Low-pass filtering is a common technique in signal processing for de-noising purposes [27].

5. Results

The models are evaluated using Word Error Rate, calculated as:

$$WER = \frac{I + D + S}{N}$$

where I, D, and S correspond to the number of Insertion, Deletion and Substitution errors after decoding, and N is the total number of words in the gold transcription. Kaldi tries several different weights of the language model versus the acoustic model at decoding time. We report on the best WER found in each condition (Table 2).

We see a dramatic improvement in performance when filtering out and interpolating over articulator coordinates assigned low confidence compared to using DLC features unaltered. This suggests that while using the DLC model to label points on the articulators is a good start, filtering based on confidence values is necessary to get good performance, and that the low-confidence coordinates are very disruptive to the model's ability to learn the relationship between phone identity and articulator movement. In addition, filtering out outlier values and lowpass

	Voiced		Silent	
	Raw	fMLLR	Raw	fMLLR
Plain DLC	61.00	54.13	81.88	71.46
+ Confidence filt	35.72	28.17	63.64	48.53
+ Outlier filt	33.60	27.82	63.49	47.71
+ Low pass filt	33.64	**26.87**	62.60	**47.15**
Ribeiro et al. (2021)	39.34	39.79	77.79	70.21

Table 2: *%WER for our experiments and previous work.*

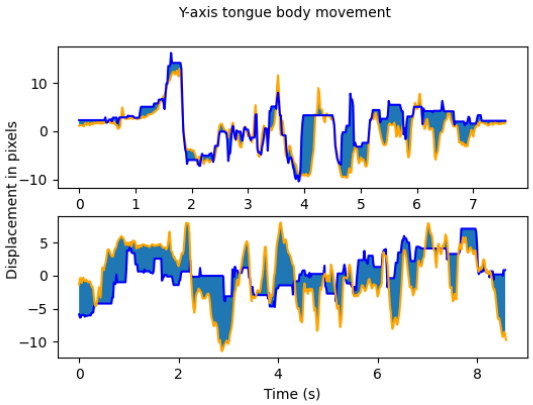


Figure 3: *Example of good (top) and bad (bottom) silent/voiced trajectory correspondence. Note the increased warp and fill area compared to the good correspondence example. Utterance: "When sunlight strikes raindrops in the air they act like a prism and form a rainbow."*

filtering the trajectories to reduce noise in the SSR system input provides further modest improvements to model performance. It is also noteworthy that the WER for our test sets outperforms the previous multi-speaker model of Ribeiro et al. once low-confidence data points are filtered out. It appears our model is able to learn more about how articulator trajectory features relate to phones as opposed to something abstracted from the raw image data. We believe the ease with which we can apply simple data conditioning techniques to the DLC-tracked features offers a distinct advantage over other articulatory representions.

We observe that fMLLR for DLC features proved more advantageous than fMLLR on the bottleneck features used in the previous study. We also note the fMLLR overall seems more helpful for silent speech than voiced speech. The greatest improvement in any condition for voiced speech due to fMLLR is 7.55%, while for silent speech it is 15.45%. This suggests that fMLLR is useful for domain adaptation, and is useful too for both modalities as a method of speaker adaptation.

Overall, the silent test set suffers from a worse WER than the voiced test set. This is expected due to domain mismatch. Performing a method of domain adaptation does dramatically improve WER but does not completely alleviate the problem. If fMLLR can reduce differences in how articulators are positioned in space, but the difference in WER between modalities is still relatively large, it suggests that there is more going on than just differences in articulatory space used.

Since audio feedback is important to driving corrective actions, we posit a lack of audio feedback may result in missed articulatory targets. We hypothesized that rather than effects like differences in speaking rate and articulatory space being responsible for the higher WER, perhaps speakers were not properly meeting targets while speaking silently. This would be difficult to ameliorate with an affine transformation method of domain adaptation, as fMLLR could essentially push peaks in the trajectory higher or lower, but it could not put them where they don't exist.

5.1. Analysis of silent versus normally-voiced speech

In order to compare the articulator trajectories between the two modalities, we applied dynamic time warping (DTW) to align the respective trajectories for the corresponding utterances of a given speaker [28]. We tested: i) warp distance as a function of WER; and ii) the area between trajectories as a function of WER. We reasoned that silent utterance trajectories which are already similar to the voiced counterpart utterances in both length and "shape" would need less warping. Similarly, a smaller area separating trajectories after DTW could represent articulators meeting their targets in a similar way (irrespective of differences in length or synchronicity before warping). Fig. 3 illustrates an example of this. We did this on trajectories which had first been adjusted to zero mean, so that effects like change in camera position would not be reflected in the difference in trajectories. We also used our low-pass filtered trajectories, as we were interested more in the overall shape of the trajectory rather than the finer details. We tested the relationship using linear regression, and used the difference in WER between corresponding utterances in the low-pass filtered condition. Warp distance and area between trajectories were averaged out among articulators for an utterance. We found that with $\alpha = .05$, warp distance was a function of WER ($R^2 = 0.060, F(1, 1179) = 74.66, p < .0001$), as well as the area between trajectories ($R^2 = 0.052, F(1, 1179) = 64.24, p < .0001$). This suggests that the more warping that is needed to get silent utterances to look like their modal counterparts, and the more dissimilar the final trajectories are, the higher WER we can expect. Since the peaks and valleys of the trajectories in the modal condition represent articulators moving to meet articulatory targets, essentially, if people speaking silently do not meet their articulatory targets in the same way (or at all) as they do when speaking aloud, then the WER of our model will be higher.

6. Conclusions

In this paper, we explored a multi-speaker ASR system for silent speech recognition. We used DeepLabCut to extract articulator coordinate features from tongue ultrasound and lip video sequences to feed as input to this model. We also used these features in an analysis of the differences in silent and voiced speech articulator trajectories. We noted that although the WER was relatively high for silent speech due to the domain mismatch, using fMLLR as a domain adaptation method greatly improved model performance. Analysing system performance in terms of differences in the overall trajectories of articulators between modalities using DTW, we found that trajectory mismatch was predictive of model performance. While fMLLR is a good candidate for a transformation based on differences in variance, other methods of adaptation would be required to address the more complex issues raised by our DTW analysis. Further kinematic analysis could be done with DLC data to deepen our understanding of the differences between silent and voiced speech.

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] J. Cai, B. Denby, P. Roussel-Ragot, G. Dreyfus, and L. Crevier-Buchman, "Recognition and real time performances of a lightweight ultrasound based silent speech interface employing a language model." in *Proceedings of INTERSPEECH*, 2011, pp. 1005–1008.

[3] 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*, 2021.

[4] L. Liu, Y. Ji, H. Wang, and B. Denby, "Comparison of dct and autoencoder-based features for dnn-hmm multimodal silent speech recognition," in *Proceedings of International Symposium on Chinese Spoken Language Processing*. IEEE, 2016, pp. 1–5.

[5] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical engineering & physics*, vol. 30, pp. 419–25, 2008.

[6] T. Hueber, E. 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, pp. 288–300, 2010.

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

[8] Y. Ji, L. Liu, H. Wang, Z. Liu, Z. Niu, and B. Denby, "Updating the silent speech challenge benchmark with deep learning," *Speech Communication*, vol. 98, pp. 42–50, 2018.

[9] N. Kimura, M. Kono, and J. Rekimoto, "Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proceedings of 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11.

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

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

[12] 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 *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*, 2021.

[13] 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.

[14] C. A. Niziolek, S. S. Nagarajan, and J. F. Houde, "What does motor efference copy represent? evidence from speech production," *Journal of Neuroscience*, vol. 33, no. 41, pp. 16 110–16 116, 2013.

[15] F. H. Guenther, J. S. Perkell, B. Maassen, R. Kent, and H. Peters, "A neural model of speech production and its application to studies of the role of auditory feedback in speech," in *Speech motor control in normal and disordered speech*, 2004, pp. 29–49.

[16] K. J. Teplansky, B. Y. Tsang, and J. Wang, "Tongue and lip motion patterns in voiced, whispered, and silent vowel production," in *Proceedings of International Congress of Phonetic Sciences*, 2019, pp. 1–5.

[17] K. J. Teplansky, A. Wisler, B. Cao, W. Liang, C. W. Whited, T. Mau, and J. Wang, "Tongue and lip motion patterns in alaryngeal speech." in *Proceedings of INTERSPEECH*, 2020, pp. 4576–4580.

[18] C. Dromey and K. Black, "Effects of laryngeal activity on articulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 2272–2280, 2017.

[19] L. Crevier-Buchman, C. Gendrot, B. Denby, C. Pillot-Loiseau, P. Roussel, A. Colazo-Simon, and G. Dreyfus, "Articulatory strategies for lip and tongue movements in silent versus vocalized speech," in *Proceedings of 17th International Congress of Phonetic Science (ICPhS)*, 2011, pp. 1–4.

[20] M. Janke, M. Wand, and T. Schultz, "Impact of lack of acoustic feedback in emg-based silent speech recognition," in *Proceedings of 11th Annual Conference of the International Speech Communication Association*, 2010.

[21] A. Eshky, M. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench, "Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions," in *Proceedings of INTERSPEECH*, 2018.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proceedings of IEEE Signal Processing Society*, 2011.

[23] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of dnns with natural gradient and parameter averaging," in *Proceedings of International Conference on Learning Representations*, 2015.

[24] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech Language*, vol. 12, no. 2, pp. 75–98, 1998.

[25] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Proceedings of International Conference on Spoken Language Processing (INTERSPEECH)*, 2002.

[26] S. Knuijt, J. Kalf, B. V. Engelen, A. Geurts, and B. de Swart, "Reference values of maximum performance tests of speech production," *International Journal of Speech-Language Pathology*, vol. 21, no. 1, pp. 56–64, 2019.

[27] D. D. Deliyski, H. S. Shaw, and M. K. Evans, "Adverse effects of environmental noise on acoustic voice quality measurements," *Journal of Voice*, vol. 19, no. 1, pp. 15–28, 2005.

[28] W. Holmes, *Speech synthesis and recognition*, 2nd ed. Taylor & Francis, 2002, ch. 8.