INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Tongue and Lip Motion Patterns in Alaryngeal Speech

Kristin J. Teplansky[1], Alan Wisler[1], Beiming Cao[1,2], Wendy Liang[1], Chad W. Whited[3], Ted Mau[4], Jun Wang[1,2]

[1]Department of Speech, Language, and Hearing Sciences, University of Texas at Austin, USA
[2]Department of Neurology, Dell Medical School, University of Texas at Austin, USA
[3]Austin Ear, Nose, and Throat Clinic, USA
[4]Department of Otolaryngology, UT Southwestern Medical Center, USA

kristin.teplansky@austin.utexas.edu, alanwisler@utexas.edu, beiming.cao@utexas.edu, wendy.liang@austin.utexas.edu, cwhited@austinent.com, ted.mau@utsouthwestern.edu, jun.wang@austin.utexas.edu

Abstract

A laryngectomy is the surgical removal of the larynx which results in the loss of phonation. The aim of this study was to characterize tongue and lip movements during speech produced by individuals who have had a laryngectomy. EMA (electromagnetic articulography) was used to derive movement data from the tongue and lips of nine speakers (four alaryngeal and five typical). The kinematic metrics included movement duration, range, speed, and cumulative path distance. We also used a support vector machine (SVM) to classify alaryngeal and healthy speech movement patterns. Our preliminary results indicated that alaryngeal articulation is longer in duration than healthy speakers. Alaryngeal speakers also use larger lateral tongue movements and move the tongue back at a slower speed than healthy speakers. The results from the SVM model also indicates that alaryngeal articulatory movement patterns are distinct from healthy speakers. Taken together, these findings suggest that there are differences in articulatory behavior that occur after the removal of the larynx. It may be helpful to consider the distinct articulatory motion patterns of alaryngeal speech in clinical practice and in the development of technologies (e.g., silent speech interfaces) that assist to provide an intelligible form of speech for this patient population.

Index Terms: speech kinematics, support vector machine (SVM), alaryngeal speech

1. Introduction

A laryngectomy is the surgical removal of the larynx due to the treatment of laryngeal cancer [1], which results in the patients' loss of normal laryngeal function and the ability to phonate. The loss of normal verbal communication has profound consequences for patients as voice plays an important role in speech intelligibility, personal identity, and social interaction. Current voice restoration options for communication post-laryngectomy include electrolarynx, tracheoesophageal prosthesis (TEP), esophageal voice, or whisper [2]. An electrolarynx is an external source that serves as a sound source by inducing vibrations through the neck tissues [3]. TEP speech uses a voice prosthesis to redirect pulmonary air through the pharyngoesophageal segment (PES) when the stoma (a surgical opening on the anterior neck) is sealed [4]. In esophageal speech, air from the esophagus sends the PES into vibration [5]. Daily communication remains a struggle as current methods used for verbal communication post-laryngectomy results in poor voice quality and often results in social isolation [6].

The vast majority of research on alaryngeal speakers has focused on investigating voice quality. Prior investigations comparing TEP speech to laryngeal speech found TEP speech to have greater variability in fundamental frequency [7], higher formant frequencies [8], increased intensity, reduced duration of phonation and speech rate [9], higher jitter and shimmer [10], and a smaller acoustic vowel space area [11]. Altered formant frequencies may be explained by a shortened vocal tract length as a result of the laryngectomy [12]. Although not directly investigated, it has been suggested that modifications to articulatory behavior also influence the formant frequencies in alaryngeal speech [5]. Moreover, reduced control of the voice source (i.e., neoglottis), increased resistance to pulmonary airflow from the neoglottis, and air leakage at the stoma site may all contribute to reduced duration of phonation and a slower speaking rate in alaryngeal speech [3]. While vocal quality is an important area of research, remarkably few studies have been designed to directly investigate articulatory movements derived from alaryngeal speakers. Advancing current understanding of tongue and lip kinematics in disordered speech can lead to improved rehabilitative treatment plans in clinical practice.

Recent speech kinematic research on healthy speakers has shown significant differences in tongue and lip movements depending on laryngeal activation (i.e., voiced versus silent) during the production of phrases [13], vowels and consonants [14], [15]. Specifically, silent articulation is longer in duration, shows a reduced peak speed [13], [14] and an increased number of articulatory sub-movements during the production of phrases [13]. Silently produced vowels also show less distinct tongue movement patterns indicated by a reduced articulatory distinctiveness space area [14]. These findings suggest that changes in laryngeal activity and in the absence of acoustic feedback impact articulatory behavior [16]. Unlike silent speech, alaryngeal speakers still receive auditory feedback; however, their speech is less intelligible. Moreover, alaryngeal speakers are likely to make articulatory adjustments to maximize speech intelligibility after laryngeal amputation.

A better understanding of articulatory movement patterns of alaryngeal speech may contribute to clinical practice and lead to improved algorithm designs for mapping articulation to speech, which can be used in the development of assistive technology that has the potential to provide laryngectomee's a more natural sounding voice (e.g., silent speech interfaces,

http://dx.doi.org/10.21437/Interspeech.2020-2854

SSI's) [16]–[19]. Ultrasound [18] and electromagnetic articulography [19] signals have shown lower silent speech recognition performance in alaryngeal speakers than their healthy counterparts.

This study sought to characterize tongue and lip motion patterns of alaryngeal speech. We directly examined tongue and lip movements derived from alaryngeal speakers as well as healthy controls. This study adds to the limited available literature on disordered articulation due to a laryngectomy.

2. Method

2.1. Participants and speech stimuli

This study included 9 participants (3 were female). Five of the participants were healthy speakers who produced voiced speech ($M_{age} = 24.80, SD_{age} = 3.19$) and four individuals had a laryngectomy and produced alaryngeal speech ($M_{age} = 53.25, SD_{age} = 21.76$). Of the alaryngeal speakers, two used a tracheoesophageal prosthesis (TEP), one used whisper [2], and one used an electrolarynx to produce speech. The alaryngeal speakers had their surgery four-five years prior to the data collection session. The healthy speakers had no reported history of speech, language, or cognitive issues. Each participant produced a list of 46 phrases at their habitual speaking rate and loudness. The phrases (e.g., '*call me back when you can.', 'I need some assistance.')* were selected because they are simple, functional, and commonly used in alternative and augmentative communication (AAC) devices [20]. All participants signed a consent form prior to participation in this study.

2.2. Tongue and lip motion tracking device

The NDI Wave system (Northern Digital Inc., Waterloo, Ontario, Canada), is a commercially available electromagnetic articulography that establishes an electromagnetic field to track tiny sensor coils in real-time. The spatial precision of Wave is ~0.5 mm [21]. The sampling rate is 100 Hz. A reference sensor was placed on the center of the head (HC) to derive and isolate head movements from the articulatory data.

An optimal four-sensor setup [22] was used to derive tongue and lip motion data. Two sensors were attached to tongue tip (TT, ~5 mm from tongue apex) and tongue back (TB, ~30 mm from TT) surface using non-toxic PeriAcryl 90 dental glue (GluStitch, Delta, British Columbia, Canada). Two sensors were adhered to the vermillion border of the upper lip (UL) and lower lip (LL) at midline using medical tape. The articulatory data included in the analysis was derived from the TT, TB, UL, and LL positional data. Please see Figure 1 an illustration of the Wave system and sensor setup. Each participant was provided approximately five minutes to talk prior to formal data collection to allow them to adapt to the wired sensors.

2.3. Data processing

The raw positional data was processed prior to data analysis. A 20 Hz low-pass filter was applied to remove noise. The head rotation and translations were subtracted from the articulatory movement data to obtain head-independent data. The orientation of the 3D Cartesian coordinate system is illustrated in Figure 1. The *x*-dimension captures lateral movements, the *y*-dimension captures superior-inferior movements, and the *z*-dimension captures anterior-posterior articulatory movements. The kinematic data were visually inspected for tracking errors prior to data analysis. Although rare, invalid data samples did occur and were excluded from data analysis.

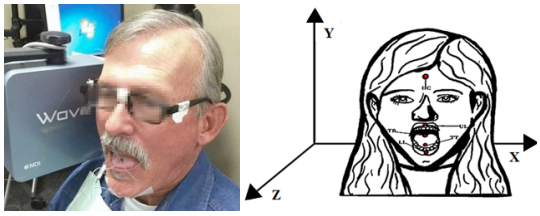

Figure 1. *The Wave system (left) and sensor placement (right) used for data collection.*

2.4. Kinematic measures

The following measures were derived from the positional data:

1. Duration (in seconds): Articulatory movement measured from sentence onset to offset.
2. Range (mm): The maximum position subtracted from the minimum position.
3. Cumulative Path Distance (mm): The length of the articulatory trajectory (path).
4. Speed (mm/s): The change in displacement over time.

The kinematic measurements were selected because they provide an overall view of tongue and lip motion behavior. All participants produced the same 46 phrases. The four metrics were derived from each phrase and then averaged for each participant. To determine if alaryngeal and healthy (voiced) speech were significantly different, independent samples t-tests were conducted. The second analysis included a machine learning classifier.

2.5. Support vector machine

A support vector machine (SVM) classifier [23] was trained by the tongue and lip movements derived from each of the speech phrases in our dataset. The model was then used to classify alaryngeal and healthy speech to determine if there is a difference between alaryngeal and healthy articulation patterns. The data were pre-processed prior to being fed into the SVM. Specifically, a linear interpolation was used to account for potential missing data points, the data were z-score normalized, and a 20Hz Butterworth low-pass filter was applied. Each kinematic sample was resampled to a fixed length of 100 data points. The 100-sample signals were concatenated to form a 1,200-dimensional feature vector (100 samples x 4 sensors x 3-dimensions). We trained a linear support vector machine classifier to learn differences between alaryngeal and healthy speakers' tongue and lip motions during speech.

Leave-one-participant-out cross-validation was used to evaluate the out-of-sample performance of the proposed classifier. That is, for each execution one participant was removed from the model and used for testing. The remaining data were used for model training.

3. Results

3.1 Movement Duration

An independent samples t-test was used to investigate tongue and lip movement duration in alaryngeal and voiced speakers.

Please see Figure 2 for descriptive statistics. The average duration of articulatory movements was longer for alaryngeal speakers ($M_{alaryngeal} = 1.95$) than healthy speakers ($M_{voiced} = 1.44$). The difference was statistically significant ($p < .05$).

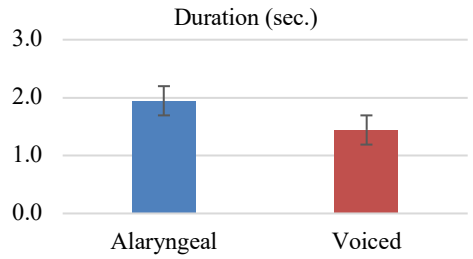

Figure 2. *Average duration of phrase production for alaryngeal and voiced speech.*

3.2 Range of Movement

The descriptive statistics for each sensor are provided in Table 1 and are plotted in Figure 3. There was a significant difference in the range of TT lateral movements ($p < .05$). Specifically, alaryngeal speakers use larger lateral movements of the TT. The TB, UL, and LL were not significantly different between alaryngeal and healthy speakers, possibly due to the small number of subjects.

Table 1. *Mean and standard deviation of movement range for each sensor.*

Sensor Location	Alaryngeal	Voiced
TT	6.10 ± 1.39	3.57 ± 0.76
TB	4.14 ± 1.32	3.18 ± 1.00
UL	2.49 ± 1.19	1.65 ± 0.22
LL	4.50 ± 2.53	2.85 ± 0.50

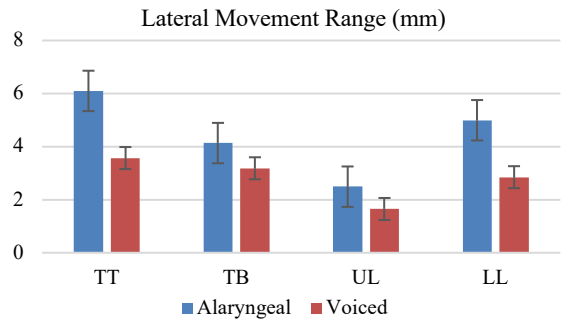

Figure 3. *Average range of lateral movement (x-dimension) with standard error bars.*

3.3 Tongue and Lip Speed

The descriptive statistics of tongue and lip movement 3D speed are provided in Table 2 and plotted in Figure 4. The results indicated that the tongue back speed was significantly slower in alaryngeal speakers than healthy voiced speakers ($p < .05$). There was no statistical difference between the two groups for the tongue tip or lip movement speed.

Table 2. *Mean and standard deviation for average 3D speed (mm/s) of each sensor.*

Sensor Location	Alaryngeal	Voiced
TT	43.35 ± 6.61	48.07 ± 9.14
TB	30.83 ± 7.81	46.23 ± 8.84
UL	10.73 ± 2.83	11.41 ± 2.44
LL	36.47 ± 15.44	34.10 ± 5.64

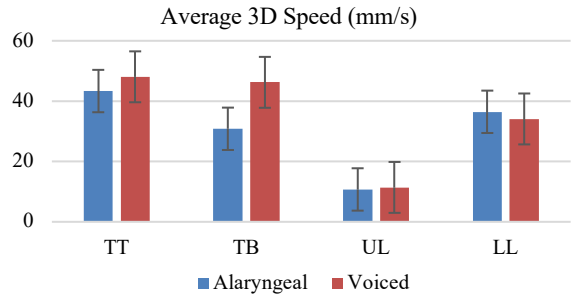

Figure 4. *Average 3D speed of the tongue and lips with standard error bars.*

3.4 Cumulative Path Distance

The descriptive statistics for each sensor are provided in Table 3 and plotted in Figure 5. The results showed a significant difference in lateral TT cumulative path distance ($p < .01$). Alaryngeal speakers use longer TT lateral movements than healthy speakers to produce speech. No significant difference was found for the TB, UL, or LL for the *y*-dimension (superior-inferior) or *z*-dimension (anterior-posterior).

Table 3. *Mean and standard deviation of cumulative path distance (mm) of lateral movements (x-dimension) for each sensor.*

Sensor Location	Alaryngeal	Voiced
TT	22.63 ± 4.71	11.10 ± 2.93
TB	16.18 ± 6.63	10.51 ± 3.07
UL	10.10 ± 5.73	5.21 ± 0.70
LL	21.13 ± 13.29	9.48 ± 0.79

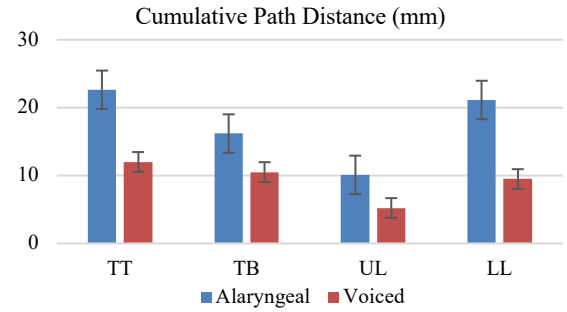

Figure 5. *Average cumulative path distance of lateral movement (x-dimension) with standard error bars.*

3.5 SVM Classification

The results of the SVM classification model are displayed in the form of a confusion matrix in Table 4. Sensitivity, specificity, and balanced accuracy were used as measures of

performance. Sensitivity provides the proportion of true positives correctly identified by the model, whereas specificity provides the proportion of true negatives correctly identified by the model. Balanced accuracy was used to determine how well the model classifies both classes (alaryngeal and voiced) from tongue and lip positional data. The classifier correctly identified most of the phrases when produced by typical speakers. However, false negatives occurred, where the model predicted alaryngeal speech to be healthy speech. Overall, this model is obtained a sensitivity (41.5%), a specificity (70.9%), and a balanced accuracy (58.0%).

Table 4. *Classification matrix.*

		Predicted	
		Alaryngeal	Voiced
Actual	Alaryngeal	**83**	101
	Voiced	67	**163**

Despite the high number of errors seen in individual predictions, if these errors are distributed in a relatively uniform manner across participants, then it may be possible to significantly improve the classification performance. This can be accomplished by aggregating the predictions across multiple phrases of speech data [24]. To do this, we randomly select groups of five and 40 phrases from our initial set of 46 phrases. Then we averaged the posterior probability predictions to generate the aggregated prediction. The process was then repeated via a 1000-iteration Monte-Carlo simulation in which different groups of phrases are selected randomly at each iteration. The results of this process along with the receiver operating characteristic (ROC) curve of the initial model are in Figure 6.

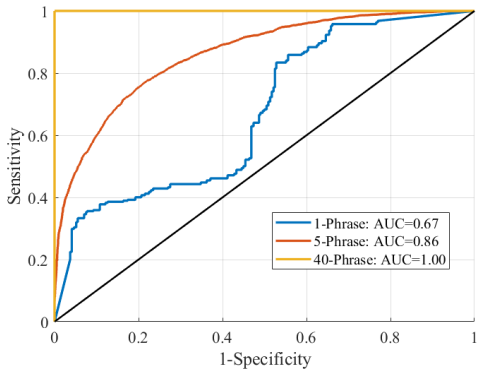

Figure 6. *SVM classification model results.*

ROC curves illustrate the trade-off between sensitivity and specificity for classification models. Better classification models will have curves that arc closer to the top left corner of the plot. Measuring the area under the curve (AUC) provides an indicator of the overall ability for the classifier to accurately discriminate between the two groups. The results show a significant improvement in performance when the predictions are aggregated across multiple phrases. Moreover, when 40 phrases are used, the model can perfectly distinguish between the two speaker groups. Thus, while differences in the articulatory motion patterns across these two groups are difficult to detect using short samples of speech, they may be reliably detected from larger speech samples.

4. Discussion

The results of our study provide preliminary evidence of distinct tongue movement patterns in alaryngeal speech during speech production. Specifically, phrases produced by alaryngeal speakers are longer in duration. Alaryngeal speakers also move the back of their tongue at a slower speed than healthy controls. A possible interpretation of this finding is that alaryngeal speakers are over articulating or exaggerating their movements [25] in attempt to produce more intelligible speech [26]. Moreover, alaryngeal speakers require more time and effort to generate an adequate airflow to produce speech, which may also contribute to longer duration [17], [27]. Another interesting finding was that alaryngeal speakers use larger lateral tongue movements than healthy speakers. Prior research has suggested that lateral tongue movements are insignificant in speech measurements of healthy speakers [28]. Our results suggest this is not necessarily true for disordered speech. A possible interpretation of this finding is that alaryngeal speakers are compensating for mechanical restrictions that may occur from the incision along the upper border of the hyoid bone during a total laryngectomy, where the hypoglossal nerve may be injured [29], [30]. The goal of the SVM classification model was to determine if there are distinct articulation pattern differences between alaryngeal and silent speech. Although the model exhibited inconsistent performance in single-phrase classification decisions, this performance improved rapidly when decisions were aggregated across a larger set of stimuli, reaching perfect performance when aggregated across 40 phrases. This suggests that while differences in the articulation patterns were not pervasive across individual data samples, they were consistent across all of the participants included in the study. Our results suggest that additional consideration is needed when developing SSI models based on healthy speakers who produce silent speech. In short, our results suggest that there are subtle differences in articulatory strategies between healthy and alaryngeal speakers.

One limitation of this project that deserves mention is that the alaryngeal speakers used different modes of speech (i.e., TEP and electrolarynx). This may cause within-group differences, which will be investigated in future work. Additionally, studies with a larger number of participants are needed to verify these findings.

5. Conclusion and Future Work

This study investigated tongue and lip movements derived from alaryngeal and healthy speakers. The results suggest that individuals who have had a laryngectomy use longer duration, slower tongue back speed, and larger lateral tongue movements during speech production than healthy speakers. The results of our SVM model suggests that alaryngeal speakers have distinct articulatory movement patterns from healthy speakers. Future studies will include additional speech stimuli and a larger number of subjects to verify these findings.

6. Acknowledgements

This work was supported by the National Institutes of Health under award numbers R03DC013990 and R01DC016621, and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We thank Dr. Anusha Thomas, Elizabeth M. Finch, Ryan Larson, Megan M. Welsh, and the volunteering participants.

7. References

[1] O. Ceachir, R. Hainarosie, and V. Zainea, "Total laryngectomy - past, present, future.," *Maedica (Buchar).*, vol. 9, no. 2, pp. 210–6, 2014.

[2] K. J. Lorenz, "Rehabilitation after total laryngectomy-A tribute to the pioneers of voice restoration in the last two centuries," *Front. Med.*, vol. 4, no. 81, pp. 1–12, 2017.

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

[4] M. Mignano, G. Acquaviva, F. Martini, C. Dominici, and R. Dellavalle, "Phonetic rehabilitation after laryngectomy," *Biomed. Pharmacother.*, vol. 47, no. 2–3, pp. 53–59, 1993.

[5] K. F. Nagle, "Elements of clinical training with the electrolarynx," in *Clinical Care and Rehabilitation in Head and Neck Cancer*, 1st ed., P. C. Doyle, Ed. Springer, Cham, 2019, pp. 129–143.

[6] J. Mertl, E. Žáčková, and B. Řepová, "Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis," *Disabil. Rehabil. Assist. Technol.*, vol. 13, no. 4, pp. 342–352, 2018.

[7] J. Robbins, H. B. Fisher, E. C. Blom, and M. I. Singer, "A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production," *J. Speech Hear. Disord.*, vol. 49, pp. 202–210, 1984.

[8] R. A. Kazi *et al.*, "Assessment of the formant frequencies in normal and laryngectomized individuals using linear predictive coding," *J. Voice*, vol. 21, no. 6, pp. 661–668, 2007.

[9] J. Robbins, "Acoustic differentiation of laryngeal, esophageal, and tracheoesophageal speech," *J. Speech Hear. Res.*, vol. 27, pp. 577–585, 1984.

[10] F. Debruyne, P. Delaere, J. Wouters, and P. Uwents, "Acoustic analysis of tracheo-oesophageal versus oesophageal speech," *J. Laryngol. Otol.*, vol. 108, no. 4, pp. 325–328, 1994.

[11] J.-S. Liao, "An acoustic study of vowels produced by alaryngeal speakers in taiwan," *Am. J. Speech-Language Pathol.*, vol. 25, pp. 481–492, 2016.

[12] N. L. Sisty and B. Weinberg, "Formant frequency characteristics of esophageal speech.," *J. Speech Hear. Res.*, vol. 15, no. 2, pp. 439–448, 1972.

[13] C. Dromey and K. M. Black, "Effects of laryngeal activity on articulation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2272–2280, 2017.

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

[15] L. Crevier-buchman *et al.*, "Articulatory strategies for lip and tongue movements in silent versus vocalized Speech," in *International Congress of Phonetic Science*, 2011, pp. 1–4.

[16] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, 2010.

[17] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2257–2271, 2017.

[18] B. Denby *et al.*, "Tests of an interactive, phrasebook-style, post-laryngectomy voice-replacement system," *17th Int. Congr. Phonetic Sci. (ICPhS XVII)*, pp. 572–575, 2011.

[19] 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 Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2323–2336, 2017.

[20] J. Wang, A. Samal, and J. Green, "Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph," pp. 38–45, 2015.

[21] J. J. Berry, "Accuracy of the NDI wave speech research system," *J. Speech, Lang. Hear. Res.*, vol. 54, pp. 1295–1301, 2011.

[22] J. Wang, A. Samal, P. Rong, and J. R. . Green, "An optimal set of flesh points on tongue and lips for speech-movement classification," *J. Speech, Lang. Hear. Res.*, vol. 59, pp. 15–26, 2016.

[23] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–39, 2011.

[24] K. H. An *et al.*, "Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks," *Interspeech*, pp. 1913–1917, 2018.

[25] J. P. Searl and M. A. Carpenter, "Acoustic cues to the voicing feature in tracheoesophageal speech," *J. Speech, Lang. Hear. Res.*, vol. 45, no. 2, pp. 282–294, 2006.

[26] J. P. Searl, "Magnitude and variability of oral pressure in tracheoesophageal speech," *Folia Phoniatr. Logop.*, vol. 54, no. 6, pp. 312–328, 2002.

[27] C. E. Stepp, J. T. Heaton, and R. E. Hillman, "Post-laryngectomy speech respiration patterns," *Ann. Otol. Rhinol. Laryngol.*, vol. 117, no. 8, pp. 557–563, 2008.

[28] J. Wang, W. F. Katz, and T. F. Campbell, "Contribution of tongue lateral to consonant production," *Proc. Annu. Conf. Int. Speech Commun. Assoc. Interspeech*, pp. 174–178, 2014.

[29] A. W. Schwartz, W. H. Hollinshead, and K. D. Devine, "Laryngectomy: Anatomy and technique," *Surg. Clin. North Am.*, vol. 43, no. 4, pp. 1063–1079, 1963.

[30] K. M. Hiiemae and J. B. Palmer, "Tongue movements in feeding and speech," *Crit. Rev. Oral Biol. Med.*, vol. 14, no. 6, pp. 413–429, 2003.