TONGUE AND LIP MOTION PATTERNS IN VOICED, WHISPERED, AND SILENT VOWEL PRODUCTION

Kristin J. Teplansky[1,2], *Brian Y. Tsang*[1,3], *Jun Wang*[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
The University of Texas at Dallas, TX, USA
[3]Department of Electrical and Computer Engineering
The University of Texas at Austin, TX, USA

{kristin.teplansky@utdallas.edu, briantsang2020@utexas.edu, wangjun@utdallas.edu}

ABSTRACT

The speech production process during vocalized speech involves the integration of auditory and vocal motor sensory feedback. The aim of this study was to evaluate articulatory behaviour in different modes of speech during the production of eight English vowels to improve our understanding of the connection between tongue motion patterns in normal (vocalized), whispered and silent (mouthed) speech. EMA was used to record movements of sensors attached to the tongue and lips. Experimental results suggest that articulatory trajectories are longer in duration and slower in speed in silent speech compared to normal and whispered speech. A machine learning algorithm (support vector machine) showed higher accuracy in vowel classification for voiced speech than in the whispered and silent conditions. In addition, articulatory distinctiveness vowel spaces showed a smaller area in silent speech than in normal and whispered speech. The results may suggest less distinct tongue movement patterns in silent speech.

Keywords: Silent speech, tongue kinematics, support vector machine, EMA, articulatory vowel space

1. INTRODUCTION

The production of speech is a sensorimotor process that is highly dependent on proper coordination between vocal fold vibration and articulatory movements on a rapid time scale, where auditory and vocal motor sensory feedback are highly integrated [21]. In vocalized speech, the vibration of the vocal folds modulates the airflow, while the vocal tract acts as a resonator that modifies the sound [9]. Previous investigations of timing relations [10] and articulatory perturbations [24] have shown the complexity of the highly coordinated interaction between the oral articulators and the larynx.

A substantial amount of literature has studied the phonetic distinctness of vowels, duration and formant frequencies in voiced [12, 13, 25] and whispered speech [15, 17-19, 27, 31]. Previous investigations on whispered speech have reported increased first formant frequency [17, 27], converged adjacent vowels, and longer duration of vowels [27] and sentences [31]. However, research on silent tongue and lip motion patterns is relatively limited, due to lack of data.

Janke and colleagues [16] used EMG to investigate audible, whispered, and silently articulated consonants and vowels. They found stronger articulatory muscle activation (i.e., hyperarticulation) during the production of bilabial consonants and rounded vowels in silent speech. Differences in articulatory strategies have also been reported during the production of French VCV utterances [14] and words [4]. Specifically, articulatory strategies were less resistant to coarticulation [14], hypoarticulation of the lips and reduced word duration in silent speech [4]. More recently, Dromey and Black [7] investigated kinematic patterns of the tongue, lips, and jaw during sentence production in voiced, whispered, and mouthed speech, where they found an increased number of articulatory sub-movements, increased sentence duration, and reduced peak velocity [7]. Additional investigations are needed for better understanding of these speaking modes.

At this time, articulatory motion patterns in silent speech at the phoneme level are poorly understood. In addition to scientific knowledge, a better understanding of articulatory movement patterns in silent speech may contribute to improved algorithm designs for mapping articulation to speech in silent speech interface (SSI) [6, 22], an assistive technology that may help the oral communication of laryngectomees (individuals after a surgical removal of larynx due to the treatment of laryngeal cancer), who are no longer able to phonate, and as a result these individuals often experience a degraded quality of life [2, 23].

The aim of the current study was to investigate tongue and lip movements during the production of normal (vocalized), whispered, and silent (mouthed) isolated vowels. Besides kinematic analysis (e.g., velocity and duration), we also applied a machine learning classifier (i.e., support vector machine, SVM) to determine if there are distinct patterns in the tongue and lip motion of voiced, whispered, and silent speech.

2. METHOD

2.1. Subjects

This study analysed data collected from 12 (six females, six males) English speakers with no reported history of speech, language, or hearing disorders. They ranged from 21 to 31 years ($M_{age} = 23.83$, $SD_{age} = 3.24$). Informed consent was obtained from each participant prior to data collection.

2.1.1. Speech tasks

Each participant repeated a list of eight isolated vowels in consonant-vowel-consonant (CVC) structure embedded in bilabial /b/ (i.e., /bæb/, /bɑb/, /b^b/, /bɔb/, /bob/, /beb/, /bib/, /bub/). Each pseudo word was produced at the speakers habitual speaking rate in the following three conditions: normal (voiced), whispered (unvoiced), and silently articulated (mouthed) 10-12 times. The investigator perceptually verified that the participants were successful in producing whispered and silent speech.

2.2. Setup and procedure

The NDI Wave electromagnetic tracking device, an electromagnetic articulography (EMA, Wave System, Northern Digital Inc., Waterloo, Canada) was used to collect synchronized speech acoustics and kinematic data from the tongue and lips. Each sensor reported *x* (lateral), *y* (superior-inferior), and *z* (anterior-posterior) positions of the articulators with a spatial tracking accuracy of 0.5 mm [1]. An optimal four-sensor setup [28] were attached the tongue tip (TT, 5-10 mm from the apex), tongue back (TB, 20–30 mm from the TT), and vermillion borders of the upper lip (UL) and lower lip (LL). The sensors were attached using non-toxic dental glue (PeriAcryl 90, GluStitch). To help the participant adapt to the wired electrodes, they were asked to speak for 3-5 minutes prior to data collection. Previous studies suggest that the sensors do not significantly interfere with speech production [20]. Fig. 1 gives an illustration of the sensor locations. Although rare, occasional sensor defects and samples affected by mispronunciations occurred and were excluded from analysis.

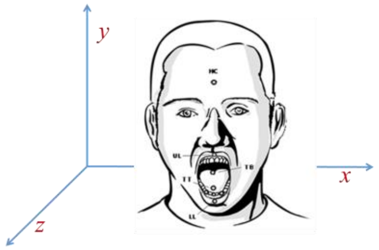


Figure 1. Sensor positions for data collection. Head centre (HC); upper lip (UL); lower lip (LL); tongue tip (TT); tongue back (TB). Sensor data were collected along the *x*-axis (lateral movements), *y*-axis (superior-inferior movements), and *z*-axis (anterior-posterior movements).

2.3. Data Preprocessing

Prior to analysis the translation and rotation components of head motion were subtracted from the tongue and lip movements to obtain head-independent data. To obtain synchronized acoustic and kinematic signals, a high-quality lapel microphone was placed approximately 15-20 cm from the participants' mouth during each recording. The data were recorded directly onto a computer hard drive at a sampling rate of 16 kHz. Simultaneous acoustic and kinematic data assisted in segmenting articulatory motion data during the voiced and whispered condition.

A customized MATLAB software program (SMASH, [11]) was used to identify the target stimuli. First, the CVC articulatory motion was aligned with the acoustic waveform. Then, visual inspection of the onset and offset of each CVC pseudo word was manually identified using SMASH [11]. The silent speech condition was segmented based on visual inspection of the motion data and prior repetitions of voiced and whispered data produced by the same participant.

3. METHODS

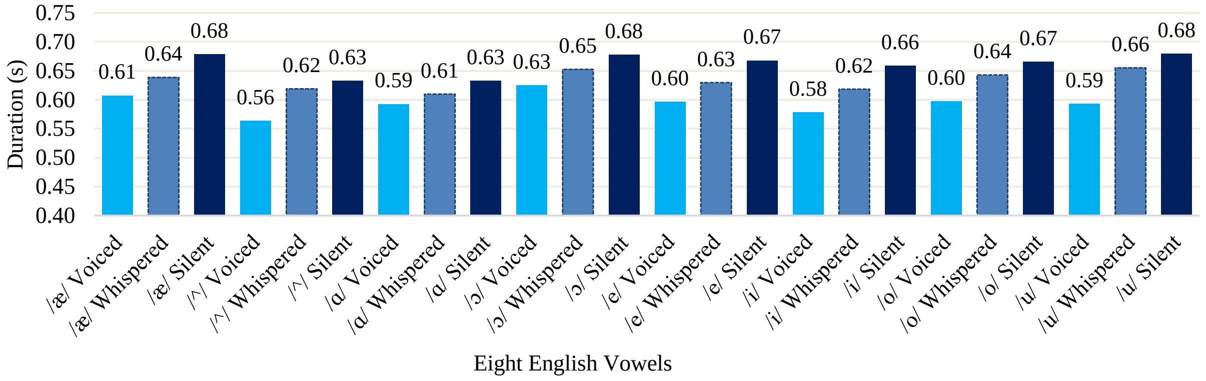
3.1. Kinematic measures

The following measures were collected along the *x*-dimension, *y*-dimension, and *z*-dimension:

1. Duration (sec.): Measured from vowel onset to offset.
2. Average speed (mm/s): The average of instant speed, calculated as the change in displacement over time.

To determine if duration and average 3D speed of tongue and lip movements during vowel production were different depending on the mode of speech

Figure 2. Average duration for all speakers during the production of four corner vowels across three speech conditions



(voiced, whispered, silent), two-factor repeated measures ANOVAs were conducted. Greenhouse-Geisser corrected values were reported with rounded degrees of freedom.

3.2. Articulatory Distinctiveness Space (ADS)

Wang and colleagues [29, 30] recently developed articulatory distinctiveness spaces (ADS), which resemble the traditional acoustic space. The ADS is based on the principal components of the entire time-series patterns of articulatory movements during the production of vowels and consonants. We applied the ADS for normal, whispered, and silent speech in this study. The following steps were used to calculate the articulatory distinctiveness space (ADS): First, a machine learning classifier (SVM) was used to classify the three speaking modes based on articulatory motion data [5], where a high accuracy indicates a distinct pattern. Second, Procrustes analysis [8], a bi-dimensional shape analysis, was used to calculate the pair-wise distance between vowels in three speaking conditions (voiced, whispered, and silent). Finally, multidimensional scaling [3] was used to create a space that preserves the distance relationship among all the vowel pairs.

4. RESULTS & DISCUSSION

4.1 Duration

Descriptive statistics are provided in Fig. 2. The results indicate a statistically significant main effect of speech mode (normal, whispered, silent), $F(1,15) = 7.962$, $p < .01$, $(\eta_p)^2 = .420$ and main effect of the vowel *(p* < .001). There was not a statistically significant interaction of speech mode and vowel ($p = .071$). The pairwise comparisons for the main effect of speech mode using a Bonferroni correction indicated that the duration of vowels was significantly longer in the silent speech ($p < .05$) condition ($M = 0.66$, $SD = 0.02$) and whispered condition ($M = 0.63$, $SD = 0.02$) than the voiced condition ($M = 0.59$, $SD = 0.02$). Although a longer duration was evident in the whispered condition ($M = 0.63$, $SD = 0.02$) when compared to the voiced condition, this finding was not statistically significant ($p = .022$). Experimental results show a significantly longer vowel duration in silent speech than voiced speech, which is consistent with the finding in [4, 7].

4.2 Average 3D speed of tongue and lip movement

Table 1. Main effect of speech mode (normal, whispered, silent).

Sensor	*df*	*F*	*p*	$(\eta_p)^2$
TT	1, 16	30.143	< .001	0.733
TB	1, 14	38.457	< .001	0.778
UL	1,18	26.401	< .001	0.706
LL	1, 20	21.296	< .001	0.659

Table 2. Interactions (vowel vs. speech mode) related to average 3D speed of the individual sensors.

Sensor	*df*	*F*	*p*	$(\eta_p)^2$
TT	3,42	4.663	<.01	0.298
TB	5,55	1.664	.158	0.131
UL	5,55	3.202	<.05	0.225
LL	3,42	9.651	<.001	0.467

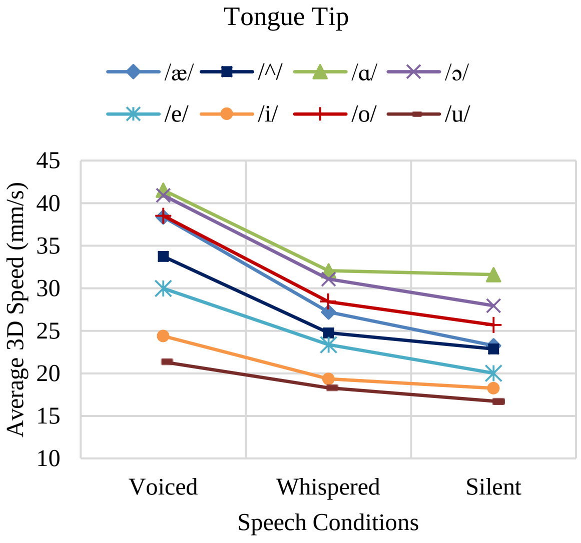
Note. For interaction analysis, two-way repeated measures ANOVA was used.

Table 3. Mean and standard deviation of tongue and lip speed for each speech mode, averaged across all vowels.

Sensor	Voiced	Whispered	Silent
TT	33.60 ± 7.76	25.56 ± 5.08	23.29 ± 5.01
TB	29.39 ± 3.42	21.81 ± 3.12	19.08 ± 2.65
UL	21.29 ± 1.86	16.40 ± 1.27	14.67 ± 1.00
LL	63.02 ± 13.46	51.15 ± 9.18	50.85 ± 6.55

The main effects of speech mode (normal, whispered, silent) are listed in Table 1. The results for the interaction term are provided in Table 2. The main effect of the vowel was statistically significant for the TT ($p < .001$), TB ($p < .01$), UL ($p < .05$), and LL ($p < .001$). Follow-up comparisons for the main effect of speech mode using Bonferroni adjustments reflect a significant difference in average speed of the tongue and lip movements among all three speaking conditions. Whispered speech was significantly slower than voiced speech for the TT ($p < .001$), TB ($p < .001$), UL ($p < .01$), and LL ($p < .001$). Silent speech was significantly slower than normal speech for the TT ($p < .001$), TB ($p < .001$), UL ($p < .001$) and LL ($p < .01$), which is also consistent with [7] using sentence stimuli. Table 3 provides the mean and standard deviation of average 3D speed. Fig. 3 provides averaged 3D speed of TT for each vowel.

Figure 3. Average 3D speed (mm/s) of the tongue tip for eight vowels in three speech conditions (voiced, whispered, and silent speech).



4.3 Articulatory Distinctiveness Space (ADS)

The overall classification accuracy of the vowels was highest for voiced vowels (88%), followed by silently produced vowels (81%) and whispered vowels (81%). Fig. 4 displays the three ADSs, where the articulatory space for voiced vowels displays a larger, expanded and distinct articulatory space ($Area_{voiced}$ = 0.0293). Whispered vowel space ($Area_{Whispered}$ = 0.0134) is smaller than the voiced space and the silent speech ($Area_{silent}$ = 0.0084) is the smallest. The results suggest that silently articulated vowels are less distinct than voiced and whispered vowels.

As illustrated in Fig. 4, low-back and centralized vowels (i.e., /ɑ/, /^/ and /ɔ/) are especially impacted when speech occurs without phonation (i.e., whispered or silent). Both machine learning (SVM) classification and Procrustes analysis (that generated the distances between vowels) indicate that distinct articulatory patterns occur in the three modes of speech. SVM classification accuracies are provided in Table 4. A visualized space (ADS) demonstrated the difference between these vowels.

Figure 4. Articulatory distinctiveness space for normal voiced (square), whispered (triangle) and silent (circle) vowels.

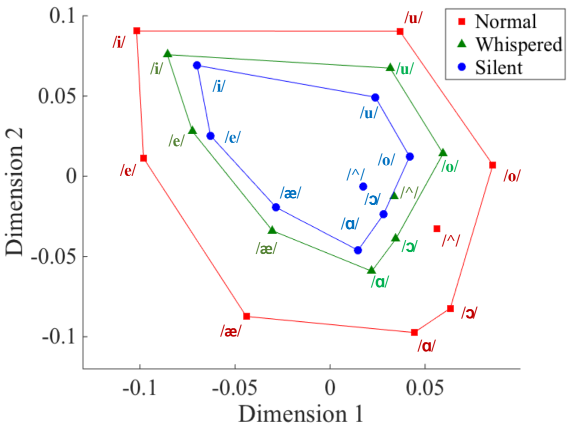


Table 4. SVM vowel classification accuracies for all vowels in normal (voiced), whispered, and silent speech conditions.

Truth / Predicted	*Voiced*	*Whispered*	*Silent*	*Total*
Voiced	**963**	94	34	1091
Whispered	94	**885**	112	1091
Silent	43	164	**884**	1091
Total	1100	1143	1030	**3273**

5. CONCLUSION AND FUTURE WORK

Our study showed distinct patterns of tongue and lip movements during the production of eight vowels in different speaking modes (voiced, whispered, and silent). Specifically, tongue and lip movements during silent speech are slower, longer in duration and less distinct. The results may be due to the fact that silent speech is not a frequently used mode of speech and does not provide auditory feedback. Additional studies on the target patient population (e.g., laryngectomees) may provide a better understanding of how articulatory patterns are affected by laryngeal surgery and improve current assistive speech technologies.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support provided by the National Institute of Health (NIH) through grant R03DC013990 and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We also thank Beiming Cao, Annette Glotfelty, Sonya Mehta, Alyssa Shrode, and volunteering participants.

7. REFERENCES

[1] Berry, J. 2011. Accuracy of the NDI wave speech research system. *J. of Speech, Language, and Hearing Research.* 54, 1295-1301.

[2] Cao, B., Kim, M., Mau, T., Wang, J. 2016. Recognizing Whispered Speech Produced by an Individual with Surgically Reconstructed Larynx Using Articulatory Movement Data. *Workshop on Speech and Language Processing for Assistive Technologies*. 80–86.

[3] Cox, R. F., Cox, M. A. A. 1994. *Multidimensional scaling.* London, UK: Chapman and Hall.

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

[5] Chang, C., Lin, C. 2018. LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*. 2(3), 1–27.

[6] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., Brumberg, J. S. 2010. Silent speech interfaces. *Speech Communication*. *52*(4), 270–287.

[7] Dromey, C., Black, K. M. 2017. Effects of Laryngeal Activity on Articulation. *IEEE/ACM Transactions on Audio Speech and Language Processing*. 25(12), 2272–2280.

[8] Dryden, I. L., Mardia, K. V. 1998. *Statistical shape analysis.* Hoboken, NJ: Wiley.

[9] Fant, G. (1960). *Acoustic theory of speech production.* The Hague, Netherlands: Mouton.

[10] Gracco, V. L., Loifqvist, A. 1994. Speech Motor Coordination and Control: Evidence from Lip, Jaw, and Laryngeal Movements. *J. of Neuroscience.* 74(11), 6565–6597.

[11] Green, J. R., Wang, J., Wilson, D. 2013. SMASH: A tool for articulatory data processing and analysis, *Interspeech.* 1331-35.

[12] Hillenbrand, J., Gayvert, R. 1993. Vowel classification based on fundamental frequency and formant frequencies. *J. of Speech and Hearing Research*, 36, 694-700.

[13] Hillenbrand, J., Gettly, A. L., Clark, J. M., and Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. of Acoust. Soc. of Am*. 97, 3099-3111.

[14] Hueber, T., Badin, P., Savariaux, C., Vilain, C., and Bailly, G. 2011. Differences in articulatory strategies between silent, whispered, and normal speech? A pilot study using electromagnetic articulography. *Proc. of ISSP, 0-1.*

[15] Itoh, T., Takeda, K, Itakura, F. 2005. Acoustic analysis and recognition of whispered speech. *IEEE Workshop on Automatic Speech Recognition and Understanding*. 45, 139-152.

[16] Janke, M., Wand, M., Schultz, T. 2010. Impact of lack of acoustic feedback in EMG-based silent speech recognition. Proc. *Interspeech*. 2686–2689.

[17] Jovičić, S. T. 1998. Formant feature differences between whispered and voiced sustained vowels. *Acta Acustica.* 84(4), 739-743.

[18] Jovičić, S. T., Šarić, Z. 2008. Acoustic Analysis of Consonants in Whispered Speech. *J. of Voice*. 22(3), 263-274.

[19] Kallail, K. J., Emanuel, F. W. 1984. Formant-frequency difference between isolated whispered and phonated vowel samples produced by adult female subject. *J. of Speech and Hearing Research*. 27, 245 – 251.

[20] Katz, W., Bharadwaj, S., Rush, M., Stettler, M. 2006. Influences of EMA receiver coils on speech production by normal and aphasic/apraxic talkers. *J. of Speech, Language, and Hearing Research*. 49, 645 – 659.

[21] Kent, D. R., Adams, G. S., Turner, S. 1996. Models of Speech Production. *Principles of Experimental Phonetics.* N.J. Lass, Ed., St. Louis, MO: Mosby.

[22] Kim, M., Cao, B., Mau, T., Wang, J. 2017. Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network. *IEEE/ACM Transactions on Audio, Speech and Language Processing*. 25(12), 2323–2336.

[23] Mertl, J., Žáčková, E., Řepová, B. 2018. Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis. *Disability and Rehabilitation. Assistive Technology*. (13)4, 342-352.

[24] Munhall, K. G., Löfqvist, A., Kelso, J. A. S., Lofqvist, A. 1994. Lip–larynx coordination in speech: Effects of mechanical perturbations to the lower lip Lip-larynx coordination in speech: Effects of mechanical perturbations to the lower lip. *J. of the Acoust. Soc. of Am.* (95), 3605.

[25] Neel, A. T. 2008. Vowel space characteristics and vowel identification accuracy. *J. of Speech, Language, and Hearing Research*. 51(3), 574-585.

[26] Schultz, T., Wand, M., Hueber, T., Krusienski, D., Herff, C., Brumberg, J. 2017. Biosignal-Based Spoken Communication: A Survey. *IEEE/ACM Transactions on Audio, Speech and Language Processing*. 25(12), 2257–2271.

[27] Sharifzadeh, H. R., McLoughlin, I. V., Russell, M. J. 2012. Toward a comprehensive vowel space for whispered speech. *International Symposium on Chinese Spoken Language Processing*. 26, 49-56.

[28] Wang, J., Samal, A., Rong, P., Green, J. 2016. An optimal set of flesh points on tongue and lips for speech-movement classification. *J. of Speech, Language, and Hearing Research*. *59*(1), 15-26.

[29] Wang, J., Green, J. R., Samal, A., Marx, D. B. 2011. Quantifying articulatory distinctiveness of vowels. Proceedings of the Annual Conference of the International Speech Communication Association, *Interspeech*. 277–280.

[30] Wang, J., Green, J. R., Samal, A., Yunusova, Y. 2013. Articulatory distinctiveness of vowels and consonants: A data-driven approach. *J. of Speech, Language, and Hearing Research.* 56(5), 1539.

[31] Zhang, C., Hansen, J. H. L. 2007. Analysis and classification of speech mode: Whispered through shouted. *Interspeech*. 4, 2396–2399.