JSLHR

Research Article

An Optimal Set of Flesh Points on Tongue and Lips for Speech-Movement Classification

Jun Wang,[a,b,c] Ashok Samal,[d] Panying Rong,[e] and Jordan R. Green[e]

Purpose: The authors sought to determine an optimal set of flesh points on the tongue and lips for classifying speech movements.
Method: The authors used electromagnetic articulographs (Carstens AG500 and NDI Wave) to record tongue and lip movements from 13 healthy talkers who articulated 8 vowels, 11 consonants, a phonetically balanced set of words, and a set of short phrases during the recording. We used a machine-learning classifier (support-vector machine) to classify the speech stimuli on the basis of articulatory movements. We then compared classification accuracies of the flesh-point combinations to determine an optimal set of sensors.
Results: When data from the 4 sensors (T1: the vicinity between the tongue tip and tongue blade; T4: the tongue-body back; UL: the upper lip; and LL: the lower lip) were combined, phoneme and word classifications were most accurate and were comparable with the full set (including T2: the tongue-body front; and T3: the tongue-body front).
Conclusion: We identified a 4-sensor set—that is, T1, T4, UL, LL—that yielded a classification accuracy (91%–95%) equivalent to that using all 6 sensors. These findings provide an empirical basis for selecting sensors and their locations for scientific and emerging clinical applications that incorporate articulatory movements.

Speech sounds are the result of complex coordinated movements of a variety of vocal-tract structures. The underlying coordination of individual articulators required to produce fluent speech involves dozens of muscles spanning the diaphragm to the lips. Technologies used to register and display speech movements are developing rapidly and removing prior barriers to developing the next generation of technology-based solutions for assisting speech communication in persons with speech and voice impairments (Green, 2015). Emerging speech-movement-based technologies include silent-speech interfaces designed to assist individuals who have had laryngectomies (surgical removal of the larynx due to treatment of cancer) or individuals with severely impaired voice and speech (Denby et al., 2010; Fagan, Ell, Gilbert, Sarrazin, & Chapman, 2008; Wang & Ham, 2015; Wang, Samal, & Green, 2014; Wang, Samal, Green, & Carrell, 2009; Wang, Samal, Green, & Rudzicz, 2012a, 2012b); speech recognition using articulatory information (Hahm, Heitzman, & Wang, 2015; King et al., 2007; Rudzicz, Hirst, & van Lieshout, 2012); treatments that provide real-time visual feedback of speech movements (Katz et al., 2014; Katz & McNeil, 2010; Levitt & Katz, 2010); and computer-aided pronunciation training for second-language learners (Levitt & Katz, 2010; Ouni, 2014; Suemitsu, Dang, Ito, & Tiede, 2015).

Although recent advances in the development of electromagnetic tracking technology have made the recording of tongue movements feasible (Green, Wang, & Wilson, 2013; Wang, Samal, & Green, 2014), use of the technology in clinical populations continues to be challenging. Some patients do not tolerate fixing multiple sensors to the tongue, and the procedure for attaching sensors to the tongue is time intensive. The use of more sensors (particularly tongue sensors) than necessary comes at a cost of unnatural articulation, a larger likelihood that sensors will fall off, and possible interference between sensors (Wang, Green, & Samal, 2013). The goal of this study was to determine if there is an optimal set of sensors that can be used for encoding articulatory information from speech units of varying sizes: phonemes, words, and phrases. Here, we operationally define the *optimal set* as the set that has a minimum number of sensors but still encodes enough

[a]Speech Disorders & Technology Lab, The University of Texas at Dallas
[b]Callier Center for Communication Disorders, The University of Texas at Dallas
[c]University of Texas Southwestern Medical Center, Dallas
[d]University of Nebraska–Lincoln
[e]MGH Institute of Health Professions, Boston, MA
Correspondence to Jun Wang: wangjun@utdallas.edu

Editor: Jody Kreiman
Associate Editor: Kate Bunton
Received April 24, 2014
Revision received November 10, 2014
Accepted August 7, 2015
DOI: 10.1044/2015_JSLHR-S-14-0112

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

information that is statistically equivalent to that encoded by a full set of sensors. The optimal set may not be unique. A *full set* comprises four sensors on the tongue and two on the lips. In prior work, researchers have chosen the number of tongue sensors and their locations on the basis of long-standing assumptions about tongue-movement patterns, depending on the specific purpose of the study. Studies on speech articulation have most often used three or four tongue sensors (Green & Wang, 2003; Guenther et al., 1999; Perkell et al., 2004; Rudzicz et al., 2012; Wang, Green, Samal, & Yunusova, 2013; Westbury, 1994; Wrench, 2000; Yunusova, Weismer, Westbury, & Lindstrom, 2008; Yunusova, Weismer, & Lindstrom, 2011).

Determining how much redundancy there is among sensors will also be useful for making inferences about the degrees of freedom of control that talkers access during speech. In the past, researchers have used a variety of statistical techniques (e.g., principal-component analysis and factor-analysis models) to estimate the functional degrees of freedom of tongue control during speech (e.g., Beaudoin & McGowan, 2000; Beautemps, Badin, & Bailly, 2001; Harshman, Ladefoged, & Goldstein, 1977; Maeda, 1978; Slud, Stone, Smith, & Goldstein, 2002; Zerling, 1979). Using factor analysis on the vocal-tract contours of vowels extracted from X-ray images, Maeda (1990) derived seven articulatory parameters that correspond to jaw position, tongue-body position, tongue-body shape, tongue-tip position, lip height, lip protrusion, and larynx height. To account for the bulk of variance in articulatory movements during vowel production, Badin, Baricchi, and Vilain (1997) reconstructed the midsagittal tongue shape on the basis of three flesh points on the tongue (roughly corresponding to the positions of the tongue tip, tongue blade, and tongue dorsum) and the position of the jaw. In a similar manner, Qin, Carreira-Perpiñán, Richmond, Wrench, and Renals (2008) showed that three or four electromagnetic articulograph sensors are sufficient to predict the tongue contour with an error of 0.2–0.3 mm. Story's computational airway model (TubeTalker) uses only four major parameters (glottal/respiratory, vocal-tract area, length, and nasalization) to drive speech synthesis (Bunton & Story, 2012; Story, 2011). On the basis of these findings, we expected that we could achieve accurate speech-movement classification using fewer flesh points on the tongue and lips.

In this study, we tested speech-movement classification accuracies using data collected from individual flesh points and their combinations at three levels of speech units: phonemes, words, and phrases. A higher accuracy indicates that the flesh point or combination of flesh points encodes more information that distinguishes the speech sounds than other points (or combinations) that yield lower accuracy. We used a support-vector machine (SVM) to classify vowels, consonants, words, and phrase samples on the basis of the movement of individual sensors and groups of sensors. SVMs are widely used machine-learning classifiers (Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995) that can successfully classify phonemes, words, and phrases on the basis of articulatory movements (e.g., Wang et al., 2012a, 2012b; Wang, Balasubramanian, et al., 2013; Wang, Green, Samal, & Marx, 2011; Wang, Green, Samal, & Yunusova, 2013). We then used the resulting classification accuracies to answer the following experimental questions: (a) Which single flesh point on the tongue and lips encodes the most information for vowels, consonants, words, or phrases? (b) Is there an optimal set of flesh points on the tongue and lips (i.e., a minimum set of sensors) that can match the classification accuracy achieved using the full set of (six) flesh points?

Method

Participants

Thirteen monolingual, native speakers of English participated in the study. The average age of the participants was 26.7 years (SD = 14.1, range = 19–65). No participant reported any hearing or speech problem or had a prior history of hearing or speech impairment. All participants were from the Midwest region of the United States. Each talker participated in one data-collection session. Ten participated in a session assessing vowels, consonants, words, and phrases; one participated in a session in which only data on vowels and consonants were collected; and the other two participants attended a session for either vowels or words only.

Stimuli

Our method used as vowel stimuli eight major English vowels in symmetrical consonant–vowel–consonant (CVC) syllables: /bɑb/, /bib/, /beb/, /bæb/, /bʌb/, /bɔb/, /bob/, and /bub/. We selected these eight vowels because they sufficiently circumscribe the boundaries of feature-based descriptions of English articulatory vowel space (see Wang, Green, Samal, & Yunusova, 2013, Figure 3A, or Ladefoged & Johnson, 2011, Figure 2.2). For example, /ɑ/ is a low and back vowel; /i/ is high and front; /u/ is high and back; /æ/ is low and middle-front; and other vowels are produced in between. Therefore, these vowels provide a good representation of the variety of tongue-and-lip movement patterns necessary to enunciate English vowels. We held the consonant context constant across stimuli to minimize the influence of consonant coarticulation effects on vowel identity. We selected the context /b/, a bilabial, because it has a minimal coarticulation effect on the vowels compared with other consonants such as /k/ and /t/ (Lindblom & Sussman, 2012).

We selected as consonant stimuli 11 consonants in symmetrical vowel–consonant–vowel (VCV) syllables: /ɑbɑ/, /ɑgɑ/, /ɑwɑ/, /ɑvɑ/, /ɑdɑ/, /ɑzɑ/, /ɑlɑ/, /ɑrɑ/, /ɑʒɑ/, /ɑdʒɑ/, and /ɑjɑ/. These consonants represent the primary places and manners of articulation of English consonants. Consonants were embedded into the /ɑ/ context because this vowel is known to induce larger tongue movements than other vowels (Yunusova et al., 2008).

To assess word stimuli, we used a set of phonetically balanced words (Wang et al., 2012b). The 25-word sets are an alternative to the traditional 50-word phonetically

balanced sets, which are widely used in speech-perception research and hearing testing (Burke, Shutts, & King, 1965; Shutts, Burke, & Creston, 1964). We picked one of four lists as word stimuli for this study (for all four lists, see Shutts et al., 1964). The words are *job, need, charge, hit, blush, snuff, log, nut, frog, gloss, start, moose, trash, awe, pick, bud, mute, them, fate, tang, corpse, rap, vast, dab,* and *ways*.

Last, we used a set of short phrases that are frequently embedded in augmentative and alternative communication devices (Wang et al., 2012a). The phrases are "*How are you doing?,*" "*I am fine,*" "*I need help,*" "*That is perfect,*" "*Do you understand me?,*" "*Hello!,*" "*Why not?,*" "*Please repeat that,*" "*Good-bye,*" "*I don't know,*" and "*What happened?*"

Tongue-Motion Tracking Devices

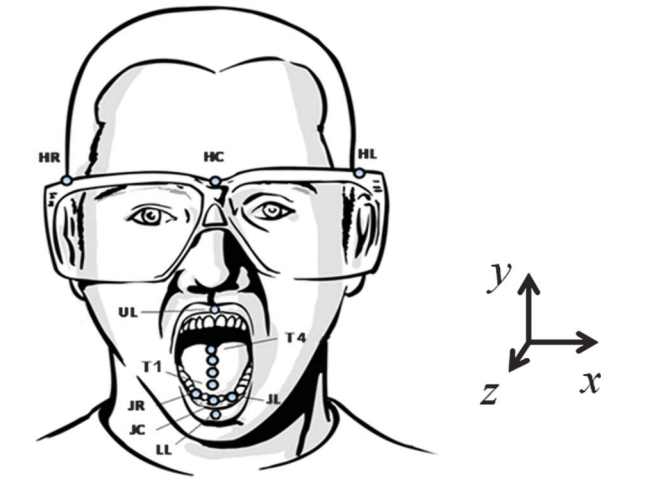
We collected 3-D movement time-series data from sensors placed on the tongue, lips, and jaw using two electromagnetic devices. An AG500 (Carstens Medizinelektronik GmbH, Bovenden, Germany) was used with 11 of the 13 participants, and a Wave system (Northern Digital Inc., Waterloo, Canada) was used for the other two participants. The two devices rely on the same electromagnetic tracking technology (Hoole & Zierdt, 2010; Perkell et al., 1992) and are safe for use with human participants (Hasegawa-Johnson, 1998). Both devices record articulatory movements by establishing a calibrated electromagnetic field that induces electric current into tiny sensor coils attached to the surface of the articulators; data-collection procedures are similar for both devices (Green et al., 2013). Thus, we describe here only the data-collection procedure for the Carstens AG500. For both devices, the spatial precision of motion tracking is approximately 0.5 mm (Berry, 2011; Yunusova, Green, & Mefferd, 2009). The sampling rates were 200 Hz for the AG500 and 100 Hz for the Wave.

Procedure

Participants were seated with their heads inside the calibrated magnetic field. Sensors were attached to the surface of each articulator using dental glue (PeriAcryl Oral Tissue Adhesive, GluStitch Inc., Delta, BC, Canada). Participants were then asked to produce the vowel, consonant, word, and phrase sequences at their habitually comfortable speaking rate and vocal intensity.

Figure 1 shows the placement of the 12 sensors attached to a participant's head, face, and tongue. Three of the sensors were attached to a pair of glasses: HC (head center) was on the bridge of the glasses, and HL (head left) and HR (head right) were on the left and right outside edges, respectively. We used the movements of the HC, HL, and HR sensors as references to calculate the movements of other articulators independent of the head (Green, Wilson, Wang, & Moore, 2007). Lip movements were captured by attaching two sensors to the vermilion borders of the upper (UL) and lower (LL) lips at midline. Four sensors—T1 (vicinity between the tongue apex and tongue blade), T2 (tongue-blade back), T3 (tongue-body front), and T4 (tongue-body back)—were attached at the midline of the tongue (Wang, Green, Samal,

Figure 1. Sensor positions. Sensor labels are described in the text. Adapted from Wang, Green, Samal, & Yunusova (2013).



& Marx, 2011; Wang, Green, Samal, & Yunusova, 2013; Westbury, 1994). Intervals between neighboring sensors were about 10 mm. T1 was 5–10 mm from the tongue apex. The movements of three jaw sensors—JL (jaw left), JR (jaw right), and JC (jaw center)—were recorded, but we did not analyze these data in the current study, because components of jaw movements were represented in both the tongue and lower-lip movement data.

To help participants pronounce the stimuli correctly, all stimuli were presented on a large computer screen in front of them, and prerecorded sounds were played when needed. For CVC and VCV stimuli, participants were asked to repeat what they heard and to put stress on the second syllable (rather than the first one). Participants were asked to rest (about 0.5 s) before each production to minimize between-stimuli coarticulation effects. This rest interval also facilitated segmenting the stimuli prior to analysis. The stimuli were presented in a fixed order (as listed earlier in Stimuli) across participants. Mispronunciations were rare but were identified and excluded from data analyses.

Each stimulus sequence was repeated multiple times by each participant. As mentioned previously, not all 13 participants participated in all sessions for vowels, consonants, words, and phrases. On average, 22 valid vowel samples were collected from each of the 11 participants, with the number of samples for each vowel varying from 16 to 29 per participant. In total, we analyzed 1,936 vowel samples, with 242 samples per vowel. The average number of valid consonant samples collected from each participant was 20, varying from 12 to 24 per participant. In total, 2,387 consonant samples (with 217 samples per consonant) were collected from the 10 participants. On average, 236 valid word samples per word were collected from the 11 participants, with the number of samples for each word varying from 16 to 26 per participant. In total, 5,900 word samples were collected and analyzed. In total, 3,012 phrase samples (251 samples per phrase) were collected, with the number of

samples varying from 19 to 27 per phrase per participant. In all, 13,235 samples (tongue and lip movements producing vowels, consonants, words, and phrases) were analyzed in this experiment.

Data Preprocessing

Figure 1 shows the orientation of the coordinate system. The origin of the coordinate system is the center of the magnetic field. Prior to analysis, we subtracted the translation and rotation components of head movement from the tongue and lip movements. The resulting head-independent tongue and lower-lip sensor positions included two components: tongue/lip movement and jaw movement. To achieve a balance of noise reduction and retaining the maximum amount of information, we used a low-pass filter on the articulatory-movement data; the cutoff frequency depended on the linguistic unit: 10 Hz for phonemes, 20 Hz for words, and 40 Hz for phrases.

We recorded acoustic and kinematic signals simultaneously, directly onto a hard drive of a computer at the sampling rate of 16 kHz with 16-bit resolution. A high-quality lapel microphone (Crown CM311, AKG Acoustics GmbH, Vienna, Austria) was mounted on a microphone stand approximately 15 cm from the participant's mouth, which was outside the magnetic field. Acoustic recordings were used only for segmenting articulatory-movement data. First, sequences of movements were aligned with acoustic waveforms. Then the onset and offset of the utterances were identified visually on the basis of acoustic-waveform data using SMASH, a MATLAB-based program developed by our group (Green et al., 2013). All manual segmentation results were checked and verified by the data analysts. On occasion, erroneous samples were collected due to a sensor falling off during recording or sounds not being produced correctly. We excluded these erroneous samples from analysis.

Only *y*- (vertical) and *z*- (anterior–posterior) coordinates of the sensors (i.e., UL, LL, T1, T2, T3, and T4) were used for analysis, because the movement along the *x*- (lateral) axis is not significant during speech of healthy talkers (Westbury, 1994). Beautemps et al. (2001) found that the midsagittal tongue contour explained 96% of tongue-data variance in a study using cineradio- and labio-film data of French phonemes (without /ɹ/ or /l/).

Analysis

We used an SVM to classify the production samples for each level of speech unit. The SVM was selected over other classifiers because of its high accuracy in classifying vowels, consonants, words, and phrases on the basis of tongue- and lip-motion data in our prior work (Wang, Balasubramanian, et al., 2013; Wang, Green, & Samal, 2013; Wang, Green, Samal, & Yunusova, 2013). The SVM also outperformed other approaches such as neural networks and decision trees for this application (Wang et al., 2009).

In machine learning, a classifier (computational model) predicts classes (groups, categories) of new data samples on the basis of a training data set, in which the classes are known. In this classification method, a data sample is defined by an array of values (attributes). A classifier makes predictions regarding data classes by analyzing these attributes. The accuracy of the prediction is based on pattern consistency in the data and on the classifier's power. An SVM tries to maximize distances between boundaries of different classes in order to obtain the best generalization of patterns between training data and testing data. SVM classifiers project training data into a higher dimensional space and then separate classes using a linear separator (Boser et al., 1992; Cortes & Vapnik, 1995). The linear separator maximizes the margin between groups of training data through an optimization procedure (Chang & Lin, 2011). A kernel function is used to describe the distance between two samples (u and v in Equation 1). The following radial basis function was used as the kernel function K_{RBF} in this study, where λ is an empirical parameter (Wang et al., 2012a, 2012b):

$$\mathrm{K}_{\mathrm{RBF}}(u, v) = \exp(1 - \lambda||u - v||). \quad (1)$$

For more details, please refer to Chang and Lin (2011), who describe the implementation of SVM used in this study.

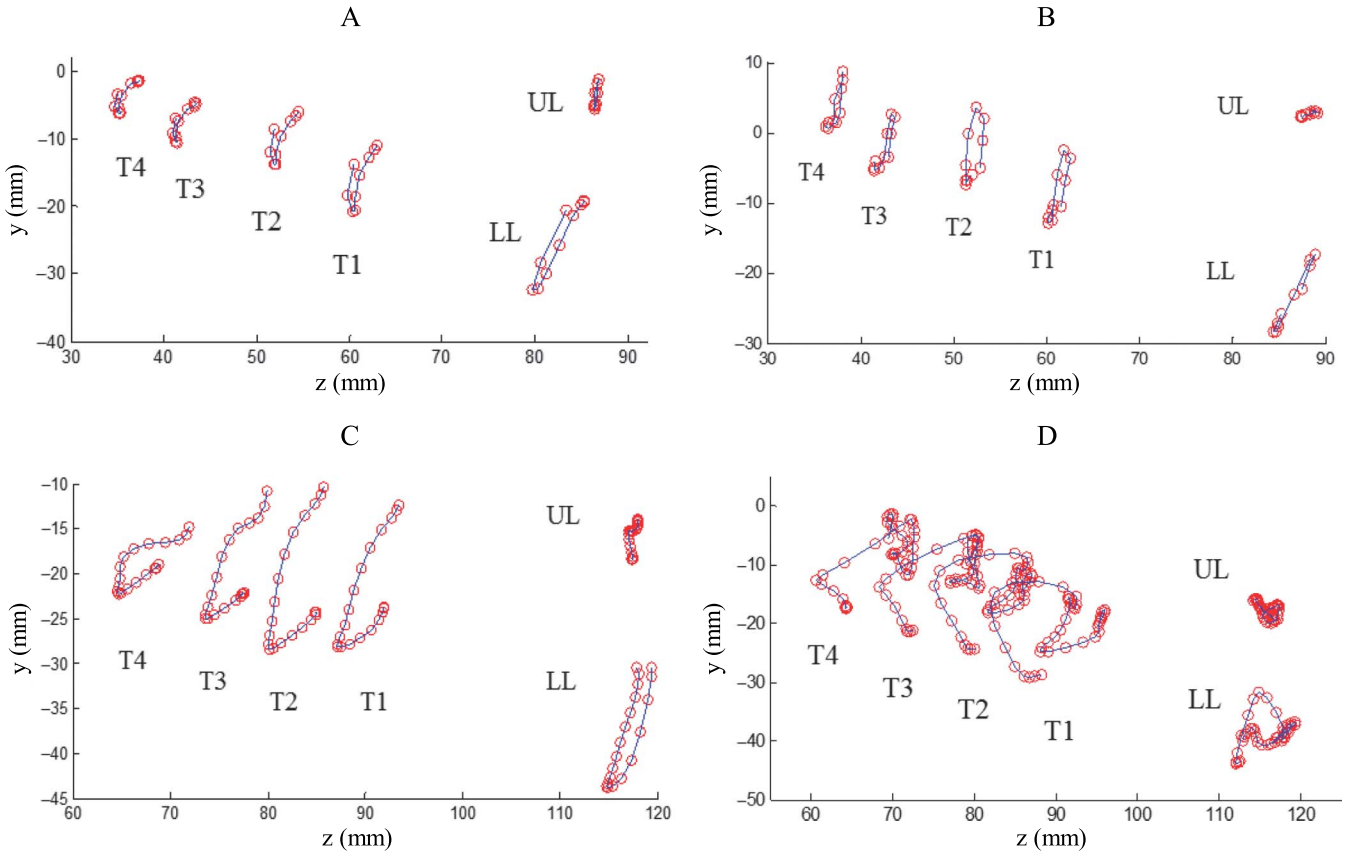
In this study, a sample (u or v in Equation 1) was a concatenation of time-sampled motion paths of sensors as data attributes. First, for each dimension y or z separately, movement data generated by each individual sensor for each stimulus (vowel or consonant) were time-normalized and sampled to a fixed length (i.e., 10 frames for CVCs or VCVs; 20 for words; 40 for phrases). The length was fixed because the SVM requires input samples to be in a fixed-width array. The predominant frequency of tongue and lip movements is about 2–3 Hz for simple CVC or VCV utterances (Green et al., 2007); thus we think 10, 20, and 40 samples adequately preserve the motion patterns of our selected phonemes, words, and short phrases. Figure 2 illustrates the down-sampled motion paths of multiple sensors for producing vowels (in VCV form), consonants (in CVC form), words, and short phrases.

Then the arrays of *y*- or *z*-coordinates for those sensors were mean-normalized (the mean values were subtracted for each dimension separately) and concatenated into one sample (vector) representing a vowel, consonant, word, or phrase. That is, the concatenated sample is a single dimension vector that contains the *y*- and *z*-coordinates of sensors. Overall, each sample contained q (q = 10 for a phoneme, 20 for a word, or 40 for a phrase × 2 dimensions × p sensors) attributes for p sensors ($1 \le p \le 6$). The following is an example of a vowel sample (a single dimension vector) with six sensors (with length 120 = 10 attributes × 2 dimensions × 6 sensors [T1, T2, T3, T4, UL, LL]):

$$\left[\mathrm{T1}_{y1}, \mathrm{T1}_{y2}, \ldots, \mathrm{T1}_{y10}, \mathrm{T1}_{z1}, \mathrm{T1}_{z2}, \ldots \mathrm{T1}_{z10}, \mathrm{T2}_{y1}, \mathrm{T2}_{y2}, \ldots, \mathrm{LL}_{z10}\right].$$

An integer (e.g., in a word-classification experiment, 1 for *job* and 2 for *need*) was added to the concatenated sample as a label for this training sample. There is no such label for testing samples. Thus, a testing-sample vector has one fewer element than a training-sample vector.

Figure 2. Examples of the sampled motion path of the six sensors for producing a vowel (CVC), consonant (VCV), word, or short phrase. The blue curves are the original motion paths; the red circles are the sampled data points. Sensor labels are described in the text. (a) Example of sampled articulatory-movement path of /bɑbɑ/ of all six sensors. Each curve is down-sampled to 10 points (indicated by red circles). Adapted from Wang, Green, Samal, & Yunusova (2013). (b) Example of sampled motion path of six sensors for producing /ɑrɑ/. Each curve is down-sampled to 10 points (indicated by red circles). (c) Example of sampled motion path of six sensors for producing the word "job." Each curve is down-sampled to 20 points (indicated by red circles). (d) Example of a sampled motion path of six sensors for producing "How are you doing?" Each curve is down-sampled to 40 points (indicated by red circles). UL = upper lip; LL = lower lip; T1 = vicinity between the tongue apex and tongue blade; T2 = tongue-blade back; T3 = tongue-body front; T4 = tongue-body back.



Cross-validation, a standard procedure to test classification algorithms in machine learning, was used to evaluate the accuracy of articulatory-movement classification using the SVM. Training data and testing data are unique in cross validation. In this experiment, we used leave-one-out cross-validation. In each execution, one sample for each stimulus (a vowel, consonant, word, or phrase) in the data set was selected for testing and the rest were used for training. There were a total of m executions, where m is the number of samples per stimulus. The average classification accuracy of all m executions was considered as the overall classification accuracy (Wang, 2011). Classifications for vowels, consonants, words, and phrases were conducted separately.

Results

Vowel Classification Using Individual Sensors

Table 1 (Column 1) lists the average vowel-classification accuracy across participants using individual flesh points, with standard deviations in parentheses. Our data show that the T4 (tongue-body back) sensor yielded the highest classification accuracy. Paired-samples t tests showed that the accuracy obtained from any individual tongue sensor (i.e., T1, T2, T3, or T4) was significantly higher than those from UL and LL; the accuracy obtained from LL was significantly higher than that from UL ($p < .01$). The accuracy of T1 was significantly different from those of T2 ($p < .05$) and T4 ($p < .05$); there was no significant difference in accuracy among T2, T3, and T4. Figure 3 plots the accuracies for individual articulators.

Consonant Classification Using Individual Sensors

Table 1 (Column 2) shows the average consonant-classification accuracy across participants using single sensors. Figure 4 plots the accuracies and shows that T1 (tongue tip) obtained the highest classification accuracy. Similar to the results for vowel classification, paired t tests showed that the accuracy obtained from any single tongue

Table 1. Average (*SD*) vowel-, consonant-, word-, and phrase-classification accuracies (%) across participants, using single sensors.

Sensor	Vowel classification	Consonant classification	Word classification	Phrase classification
T1	81.62 (8.82)	**80.82** (8.10)	**89.98** (3.47)	91.82 (3.66)
T2	84.57 (7.83)	74.79 (13.48)	88.10 (6.39)	90.83 (5.36)
T3	83.56 (7.35)	75.10 (13.35)	81.99 (6.66)	89.82 (6.49)
T4	**85.59** (6.00)	80.16 (11.16)	89.61 (5.57)	93.17 (3.92)
UL	62.38 (9.95)	53.96 (12.65)	74.34 (8.07)	95.62 (2.63)
LL	73.29 (8.71)	66.55 (12.39)	84.51 (6.71)	**95.90** (2.23)

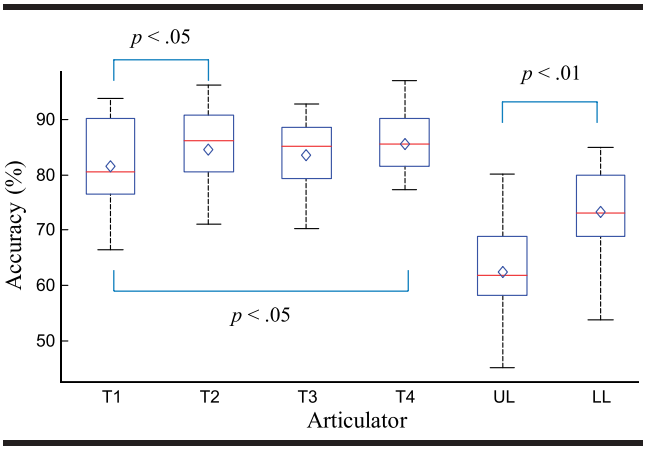
Note. UL = upper lip; LL = lower lip; T1 = vicinity between the tongue apex and tongue blade; T2 = tongue-blade back; T3 = tongue-body front; T4 = tongue-body back. The highest accuracy in each category (column) is shown in bold.

sensor was significantly higher than those from LL and UL; the accuracy from LL was significantly higher than that from UL ($p < .01$). It is interesting to note that unlike with the vowel-classification results, there was no significant difference among any of the tongue sensors (T1, T2, T3, or T4).

Word Classification Using Individual Sensors

Table 1 (Column 3) lists the average word-classification accuracy across participants using single sensors. The highest accuracy was obtained from T1. Similar to the results for vowel and consonant classification, the accuracy obtained from any single tongue sensor was significantly higher than that from any lip sensor (LL or UL)—except for T3, which had no significant differences compared to UL and LL; the accuracy from LL was not significantly different than that from UL ($p < .01$). Unlike the vowel- and consonant-classification results, word-classification accuracy obtained from either T1 or T2 was significantly higher than that from T3 ($p < .01$); however, there was no significant difference compared to the accuracy obtained from T4. There were no significant differences observed among T1, T2, or T4. Figure 5 plots the accuracies for each individual sensor.

Figure 3. Average vowel-classification accuracies across participants for individual sensors, shown as box-whisker plots (diamond is the mean value; red line is the median; edges of the boxes are 25th and 75th percentiles). There is a significant difference between any tongue sensor and any lip sensor (not displayed).



Phrase Classification Using Individual Sensors

Table 1 (Column 4) shows the average phrase-classification accuracies across participants using single sensors. It is surprising that, unlike with vowel, consonant, and word classification, a lip sensor (i.e., LL) achieved the highest accuracy; the accuracy from LL was significantly greater than that from any single tongue sensor, but not significantly greater than that from UL. The accuracy from UL was significantly higher than those from T2 and T3 but not from T1 and T4. There were no significant differences among any individual tongue sensors. Figure 6 plots the accuracies for each individual sensor.

Classification Using Sensor Combinations

To determine an optimal set (a minimum set of sensors that can match the classification accuracy of data generated by all six sensors), we compared the classification accuracies of all possible combinations of sensors. We named this hypothesized optimal combination/set $\boldsymbol{A}$:

$$A = \{\mathrm{T1}, \mathrm{T4}, \mathrm{UL}, \mathrm{LL}\}. \quad (2)$$

Figure 4. Average consonant-classification accuracies across participants for individual sensors, shown as box-whisker plots (diamond is the mean value; red line is the median; edges of the boxes are 25th and 75th percentiles). There is a significant difference between any tongue sensor and any lip sensor (not displayed). The red cross is an outlier.

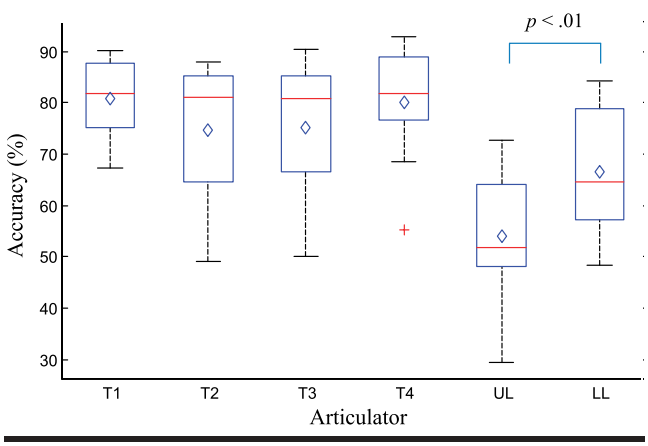
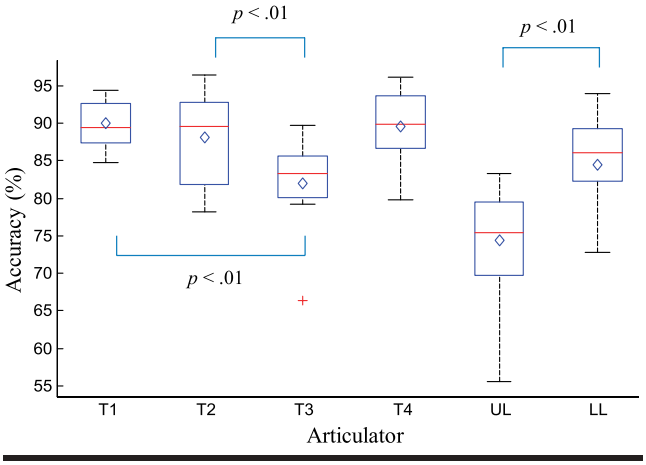
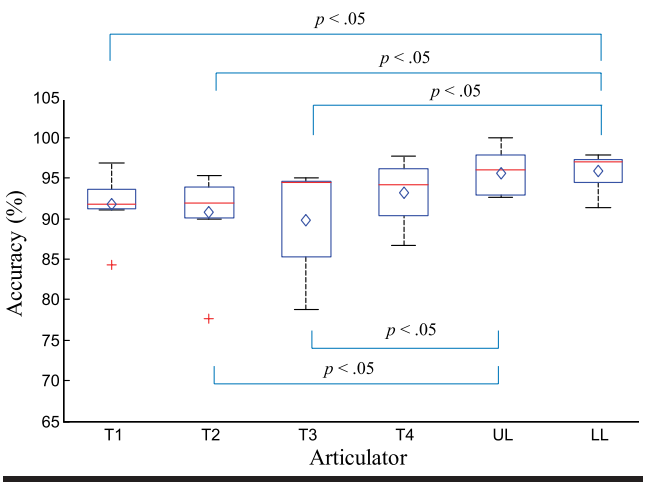


Figure 5. Average word-classification accuracies across participants for individual sensors, shown as box-whisker plots (diamond is the mean value; red line is the median; edges of the boxes are 25th and 75th percentiles). There is a significant difference between any tongue sensor and any lip sensor (not displayed), except between T3 and UL.



On the basis of our prior findings—specifically that T1 obtained the highest accuracy in consonant and word classification, whereas T4 obtained the highest accuracy in vowel classification (Wang, Samal, & Green, 2013)—we hypothesized that the boundaries defined by the movements of the tongue and lips during speech can be captured on the basis of data from only the following four sensors: T1, T4, UL, and LL. Although UL on its own is not a primary articulator during speech, the vertical and anterior distance between UL and LL determines lip aperture, which is an important articulatory gesture in both vowel (e.g., /u/) and consonant (e.g., /p/, /b/) productions (Ladefoged & Johnson, 2011). Tongue sensors are not able to capture these movements. Therefore, we included both UL and LL in our hypothesis to represent lip movement during speech.

Figure 6. Average phrase-classification accuracies across participants for individual sensors, shown as box-whisker plots (diamond is the mean value; red line is the median; edges of the boxes are 25th and 75th percentiles).



First, we compared the accuracy obtained from $\boldsymbol{A}$ (Eq. 2) to the accuracies obtained from different combinations using fewer sensors—that is, {T1, T4}, {T1, T4, UL}, and {T1, T4, LL}—and single sensors {T1}, {T4}, {UL}, and {LL}. These comparisons were made to verify that no combination with fewer sensors than $\boldsymbol{A}$ has a similar or higher accuracy compared to that of $\boldsymbol{A}$. Second, $\boldsymbol{A}$ was compared to those combinations without lip sensors but with more tongue sensors—i.e., {T1, T4, T2}, {T1, T4, T3}, and {T1, T4, T2, T3}—to verify that lip sensors are needed to maintain accuracy. Last, $\boldsymbol{A}$ was compared to those combinations with extra sensors—that is, $\boldsymbol{A} \cup$ {T2}, $\boldsymbol{A} \cup$ {T3}, and $\boldsymbol{A} \cup$ {T2, T3}—to verify that extra (tongue) sensors do not improve the classification accuracy.

Table 2 lists the accuracies obtained from $\boldsymbol{A}$ and from all other relevant combinations, as well as any statistically significant differences between them. As anticipated, the accuracy obtained from $\boldsymbol{A}$ was significantly higher than the accuracy obtained from any combination with fewer sensors or any combination with extra sensors in vowel-, consonant-, and word-classification experiments.

Discussion

We used a machine-learning classifier (an SVM) to classify vowels, consonants, words, and phrases on the basis of movement data from individual tongue and lip sensors or their combinations. The accuracies of data from each sensor indicated the level of information that is encoded in these individual sensors. We identified an optimal set of four sensors (i.e., tongue tip [T1], tongue-body back [T4], upper lip [UL], and lower lip [LL]) that could be used to accurately classify phonemes, words, and phrases. These findings will inform future work designed to improve the assessment and treatment of communication impairments using speech-movement-based technologies.

Guidance on Sensor Selection

To the best of our knowledge, this study is the first to empirically determine an optimal number of sensors and sensor locations that can be used for classification studies of speech movements. Our results suggest that (a) the tongue-body back (T4) sensor conveys the most information for classifying vowels, (b) the tongue tip (T1) conveys the most information for distinguishing consonants or words, and (c) a lip sensor (LL) achieved the highest accuracy for classifying a small set of phrases. For vowels, consonants, and words, classification accuracy was greatest when it was based on data from the four sensors T1, T4, UL, and LL. In contrast, the movement of either the upper or lower lip was sufficient for recognizing a small set of phrases. These findings may help guide the selection of the number of sensors and their locations for use in speech kinematic studies or in similar clinical applications using electromagnetic articulographs.

Table 2. Average vowel-, consonant-, word-, and phrase-classification accuracies (%) across participants on selected sensors and sensor combinations.

Sensor or sensor combination	Vowel classification	Consonant classification	Word classification	Phrase classification
{T1}	81.62***	80.82***	89.98***	91.82**
{T4}	85.59***	80.16***	89.61**	93.17
{UL}	62.38***	53.96***	74.34***	95.62
{LL}	73.29***	66.55***	84.51**	**95.90***
{T1, T4}	88.20***	88.02**	93.55**	92.97*
{T1, T4, UL}	90.66*	90.33	96.42	94.83
{T1, T4, LL}	90.78*	90.55**	95.59**	93.51**
{T1, T4, T2}	87.17***	87.54**	93.38**	93.73**
{T1, T4, T3}	87.29**	87.55**	91.49***	93.26*
{T1, T4, T2, T3}	86.57**	87.45*	92.10**	93.20*
{**T1, T4, UL, LL**}	**91.67**	**91.8**	**96.88**	95.08
{**T1, T4, UL, LL**} ∪ {T2}	91.37	91.26	96.21*	95.14
{**T1, T4, UL, LL**} ∪ {T3}	91.31	91.45	95.22*	95.55
{**T1, T4, UL, LL**} ∪ {T2, T3}	90.71	91.37	94.76*	95.12

Note. The highest accuracy in each category (column) is shown in bold. Indicated *p* values refer to significant differences between that sensor or sensor combination and ***A*** ({T1, T4, UL, LL}).

*$p < .05$. **$p < .01$. ***$p < .001$.

Vowel Classification Using Individual Sensors

In general, our findings for vowel classification suggest that tongue sensors contribute more to vowel classification than do lip sensors. This finding is consistent with the long-standing descriptive data from classical phonetics studies, which assert that vowels are distinguished by tongue height and front–back position (Ladefoged & Johnson, 2011). The finding that the accuracy obtained from LL is higher than that from UL was not surprising, because the movement of LL included the movements of the jaw, which is also a major articulator (Kent, Adams, & Turner, 1996).

The finding that T2 and T4 obtained significantly higher accuracies than T1 may be explained by the fact that they are farther away from parts of the tongue responsible for the main vocalic constriction—such as palatal for /i/-like vowels, velar for /u/-like vowels, and pharyngeal for /ɑ/-like vowels—than other sensors are.

The position and height of the tongue tip may be more variable than those of the tongue body and tongue back during vowel production. Our findings suggest that if only one sensor can be used in a vowel-production study, it should be the tongue-body back (T4) rather than the tongue tip. It is important to note that our data were collected from English speakers, which procures a bias toward generating accurate information from tongue sensors because English does not have any front rounded vowels (or back unrounded vowels). Generalization of these findings to other languages will require future investigations.

Consonant Classification Using Individual Sensors

For consonant classification, our findings suggest that no single tongue flesh point alone conveys more information than any other, which may be explained by the tight biomechanical coupling between adjacent tongue regions (Green & Wang, 2003). These findings are consistent with the suggestion that consonant production involves more independent motions of different parts of the tongue than vowel production does. Each tongue sensor encoded a similar amount of information for classification purposes. However, T1 and T4 achieved about 5% higher accuracy than either T2 or T3; the differences, however, were not statistically significant. With more data, these differences might be statistically significant. As mentioned previously, this conclusion is based on a speech-movement classification experiment, and the findings may not generalize to other applications. For example, it is inappropriate to use T1 or T4 only in a study that involves dorsal consonants.

Word Classification Using Individual Sensors

The finding that the tongue tip (T1) and tongue blade (T2) sensors encode significantly more information than the tongue body front (T3) sensor may reflect the quasi-independent movement of these regions during consonant production. This finding suggests that sensors placed on the front of the tongue (including T1 and T2) encode critical information for distinguishing phonetically balanced words. Word findings furthermore confirmed that T1 appears to be the best sensor to use if only one tongue sensor can be used in a word-level speech kinematic study.

Phrase Classification Using Individual Sensors

Our phrase-classification results that lip sensors had significantly higher accuracies than any single (or combined) tongue sensor were surprising. This finding suggests that lip movements encode information that can adequately distinguish a small set of short phrases and that the movement of tongue sensors may have larger variation than lip sensors during connected speech. Additional experiments and analysis are required to determine if these findings can be generalized to a larger set of phrases.

Classification Using Sensor Combinations

In this experiment, we first hypothesized that the sensor set {T1, T4, UL, LL} would be optimal for classification of movement data into speech units, on the basis of the literature and our preliminary study (Wang, Green, & Samal, 2013). To test this hypothesis, we compared this set with other sensor combinations with fewer or more sensors (see Table 2). We observed that classification accuracy decreased when classification was based on fewer sensors than these four and that the inclusion of additional sensors (i.e., T2 and T3) did not increase accuracy. These findings were consistent with those observed for vowel, consonant, and word classification. The results suggest that classification accuracy will decrease significantly if any sensors in ***A*** are not included. Moreover, the addition of extra sensors did not significantly increase classification accuracy for vowels, consonants, and phrases, and even decreased accuracy for word classification. Our results suggest that {T1, T4, UL, LL} is an optimal set to encode the articulatory distinctiveness among different phonemes, words, and phrases. We note that the set ***A*** may not be unique. An additional study is required to examine, for example, if either T1 or T4 is replaceable (by T2 or T3).

Degrees of Freedom of Tongue Movement in Speech Production

The major finding of this study—that two flesh points on the midsagittal region of the tongue can represent the principal movements of the tongue during speech—is consistent with previous literature demonstrating that most of the variance in speech acoustics can be accounted for by a small number of parameters that represent the functional degrees of freedom of tongue movement during speech. Prior work has relied primarily on dimensionality-reduction techniques (Badin, Bailly, Raybaudi, & Segebarth 1998; Engwall, 2000; Harshman et al., 1977; Hoole, 1999; Yehia & Tiede, 1997) to extract the shared variance across articulators. In contrast, our approach tested all relevant combinations of these flesh points to identify an optimal set. Both approaches are successful because many regions of the tongue are biomechanically coupled, which limits their movement independence and the available coordinative options. Green and Wang (2003) found that, across consonants, adjacent tongue regions are highly coupled during speech, whereas the tongue tip and tongue-body back (named T1 and T4, respectively) move quasi-independently. Convergence across methodologies further supports the conclusion that a small number of functional degrees (highly constrained) of freedom are used during speech.

Limitations of This Study

Classification in this study was based only on the vertical and anterior–posterior aspects of speech movements. Previous studies have demonstrated, however, that two dimensions are adequate for representing three-dimensional vocal-tract shape changes produced during speech. For example, some studies extending 2-D midsagittal vocal-tract models to 3-D models have shown that five midsagittal articulatory parameters (i.e., jaw height, tongue body, tongue dorsum, tongue tip, and tongue advancement) accounted for most of the variance in 3-D tongue movement (Badin et al., 1998; Engwall, 2000). Jaw height, tongue body, and tongue dorsum were the main predictors for both sagittal and lateral dimensions. Yehia and Tiede (1997) showed that 3-D vocal-tract shapes can be approximated by linear combinations of four 3-D basis functions estimated from the midsagittal configurations of the vocal tract. All of these findings suggest that the shape of the vocal tract (in both 2- and 3-D) can be represented by a limited number of parameters that correspond to how talkers constrain the large number of degrees of freedom of articulatory movement. Accounting for lateral movements may, however, be particularly important when analyzing atypical speech movements of persons with motor-speech disorders.

All the tongue sensors in this study were placed at midsagittal locations on the tongue, lips, and jaw. Additional work is required to determine if sensors placed on the lateral margins of the tongue encode unique features that distinguish, for example, /l/ from /r/ or /s/ from /ʃ/ (Ji, Berry, & Johnson, 2014; Wang, Katz, & Campbell, 2014).

In the current study, the flanking sounds were held constant during vowel (i.e., CVC) and consonant (i.e., VCV) classification. Of course, phoneme classification will become significantly more difficult when phonemes are embedded in a wider variety of contexts, such as different words. Additional work is required to explore the limits of phoneme classification solely on the basis of speech-movement data.

Last of all, we selected stimuli in this study for the general purpose of studying selected applications such as small-vocabulary classification (Hofe et al., 2013; Wang et al., 2012a, 2012b) and visual feedback for speech therapy (Katz et al., 2014). The stimuli set (eight vowels and 11 consonants) did not account for some potential important contrasts that are needed to improve continuous silent-speech recognition (Hahm & Wang, 2015; Wang & Hahm, 2015) and speech recognition from combined acoustic and articulatory data (Hahm, Heitzman, & Wang, 2015; King et al., 2007; Rudzicz et al., 2012). However, a recent separate study confirmed this four-sensor sensor set is also optimal for continuous silent speech recognition (Wang, Hahm, & Mau, 2015).

Conclusion

This study investigated classification of major English vowels, consonants, a phonetically balanced set of words, and a small set of short phrases on the basis of articulatory-movement time-series data from six sensors placed on the tongue and lips. Experimental results were based on 13,235 speech samples obtained from 13 healthy, native English speakers. A machine-learning classifier (SVM) was used to compare accuracies among individual sensors and

combinations of sensors. Classification accuracy of vowels, consonants, words, and phrases was found to be as high (approximately 91%–95%) for a set of four-sensor set (tongue tip, tongue-body back, upper lip, and lower lip) as it was when data were included from all six sensors. Thus, we conclude that {T1, T4, UL, LL} is an optimal set for speech-movement classification or relevant applications. Of course, the number of sensors and their locations may vary depending on the purpose of the study and its application. But in general, using two tongue sensors is more practical than using more sensors (e.g., three or four) for investigating disordered speech articulation (e.g., Rong, Yunusova, Wang, & Green, 2015; Yunusova, Green, Wang, Pattee, & Zinman, 2011), and two are sufficient for relevant assistive technology (e.g., Wang, Samal, & Green, 2014).

Acknowledgments

This work was supported in part by the Excellence in Education Fund, The University of Texas at Dallas; the Barkley Trust, Barkley Memorial Center, University of Nebraska–Lincoln; National Institutes of Health Grants R01 DC009890 (principal investigator: Jordan R. Green), R01 DC013547 (principal investigator: Jordan R. Green), and R03 DC013990 (principal investigator: Jun Wang); and the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant (principal investigator: Jun Wang). About a quarter of the data and analysis was presented at the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (Vancouver, Canada) and published in the conference's annual proceedings (Wang, Green, & Samal, 2013). We would like to thank Tom D. Carrell, Mili Kuruvilla, Lori Synhorst, Cynthia Didion, Rebecca Hoesing, Kayanne Hamling, Katie Lippincott, Kelly Veys, Tony Boney, and Lindsey Macy for their contribution to participant recruitment, data management, data collection, data processing, and other support.

References

Badin, P., Bailly, G., Raybaudi, M., & Segebarth, C. (1998). A three-dimensional linear articulatory model based on MRI data. In *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis* (pp. 249–254). Jenolan Caves, Australia: ECSA.

Badin, P., Baricchi, E., & Vilain, A. (1997). Determining tongue articulation: From discrete fleshpoints to continuous shadow. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (Eds.), *Proceedings of the Eurospeech '97—5th European Conference on Speech Communication and Technology, Vol. 1* (pp. 47–50). Grenoble, France: European Speech Communication Association.

Beaudoin, R. E., & McGowan, R. S. (2000). Principal components analysis of X-ray microbeam data for articulatory recovery. In *Proceedings of the Fifth Seminar on Speech Production* (pp. 225–228). Kloster Seeon, Bavaria, Germany: International Seminar on Speech Production.

Beautemps, D., Badin, P., & Bailly, G. (2001). Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *The Journal of the Acoustical Society of America, 109,* 2165–2180.

Berry, J. J. (2011). Accuracy of the NDI Wave speech research system, *Journal of Speech, Language, and Hearing Research, 54,* 1295–1301.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceeding of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). New York, NY: Association for Computing Machinery.

Bunton, K., & Story, B. H. (2012). The relation of nasality and nasalance to nasal port area based on a computational model. *The Cleft Palate–Craniofacial Journal, 49,* 741–749.

Burke, K. S., Shutts, R. E., & King, W. P. (1965). Range of difficulty of four Harvard phonetically balanced word lists. *The Laryngoscope, 75,* 289–296.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2,* 27.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20,* 273–297.

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

Engwall, O. (2000). A 3D tongue model based on MRI data. In *Proceedings of the Sixth International Conference on Spoken Language Processing, Vol. 3* (pp. 901–904). Beijing, China: International Speech Communication Association.

Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E., & Chapman, P. M. (2008). Development of a (silent) speech recognition system for patients following laryngectomy. *Medical Engineering & Physics, 30,* 419–425.

Green, J. R. (2015). Mouth matters: Scientific and clinical applications of speech movement analysis. *Perspectives on Speech Science and Orofacial Disorders, 25,* 6–16.

Green, J. R., & Wang, Y.-T. (2003). Tongue-surface movement patterns during speech and swallowing. *The Journal of the Acoustical Society of America, 113,* 2820–2833.

Green, J. R., Wang, J., & Wilson, D. L. (2013). SMASH: A tool for articulatory data processing and analysis. In F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, & P. Perrier (Eds.), *Interspeech 2013—14th Annual Conference of the International Speech Communication Association* (pp. 1331–1335). Lyon, France: International Speech Communication Association.

Green, J. R., Wilson, E. M., Wang, Y.-T., & Moore, C. A. (2007). Estimating mandibular motion based on chin surface targets during speech. *Journal of Speech, Language, and Hearing Research, 50,* 928–939.

Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., & Perkell, J. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *The Journal of Acoustical Society of America, 105,* 2854–2865.

Hahm, S., Heitzman, D., & Wang, J. (2015). Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization. In *6th Workshop on Speech and Language Processing for Assistive Technologies* (pp. 47–54). Dresden, Germany: Association for Computational Linguistics.

Hahm, S., & Wang, J. (2015). Silent speech recognition from articulatory movements using deep neural network. In Scottish Consortium for ICPhS 2015 (Eds.), *Proceedings of the 18th International Congress of Phonetic Sciences* (no. 524, pp. 1–5). Glasgow, Scotland: University of Glasgow.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America, 62,* 693–707.

Hasegawa-Johnson, M. (1998). Electromagnetic exposure safety of the Carstens Articulograph AG100. *The Journal of the Acoustical Society of America, 104,* 2529–2532.

Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., & Rybchenko, S. I. (2013). Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication, 55,* 22–32.

Hoole, P. (1999). On the lingual organization of the German vowel system. *The Journal of the Acoustical Society of America, 106,* 1020–1032.

Hoole, P., & Zierdt, A. (2010). Five-dimensional articulography. In B. Maassen & P. van Lieshout (Eds.), *Speech motor control: New developments in basic and applied research* (pp. 331–349). Oxford, United Kingdom: Oxford University Press.

Ji, A., Berry, J., & Johnson, M. T. (2014). The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7719–7723). Florence, Italy: Institute of Electrical and Electronics Engineers.

Katz, W., Campbell, T., Wang, J., Farrar, E., Eubanks, J. C., Balasubramanian, A., . . . Rennaker, R. (2014). Opti-Speech: A real-time, 3D visual feedback system for speech training. In H. Li, H. Meng, B. Ma, E. S. Chng, & L. Xie (Eds.), *Interspeech 2014—15th Annual Conference of the International Speech Communication Association* (pp. 1174–1178). Singapore: International Speech Communication Association.

Katz, W. F., & McNeil, M. R. (2010). Studies of articulatory feedback treatment for apraxia of speech based on electromagnetic articulography. *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders, 20*(3), 73–79.

Kent, R. D., Adams, S. G., & Turner, G. S. (1996). Models of speech production. In N. J. Lass (Ed.), *Principles of experimental phonetics* (pp. 3–45). St. Louis, MO: Mosby-Yearbook.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007). Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America, 121,* 723–742.

Ladefoged, P., & Johnson, K. (2011). *A course in phonetics* (6th ed.). Boston, MA: Wadsworth.

Levitt, J. S., & Katz, W. F. (2010). The effects of EMA-based augmented visual feedback on the English speakers' acquisition of the Japanese flap: A perceptual study. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Interspeech 2010—11th Annual Conference of the International Speech Communication Association* (pp. 1862–1865). Makuhari, Japan: International Speech Communication Association.

Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics, 40,* 1–19.

Maeda, S. (1978). Une analyse statistique sur les positions de la langue: Étude préliminaire sur les voyelles françaises [A statistical analysis of tongue positions: Preliminary study of French vowels]. In *Actes des 9èmes Journées d'Études sur la Parole* (pp. 191–199). Lannion, France: Groupement d'Acousticiens de la Langue Française.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 131–149). Dordrecht, the Netherlands: Kluwer Academic.

Ouni, S. (2014). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning, 27,* 439–453.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., & Jackson, M. T. T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America, 92,* 3078–3096.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of Acoustical Society of America, 116,* 2338–2344.

Qin, C., Carreira-Perpiñán, M. Á., Richmond, K., Wrench, A., & Renals, S. (2008). Predicting tongue shapes from a few landmark locations. In *Interspeech 2008—9th Annual Conference of the International Speech Communication Association* (pp. 2306–2309). Brisbane, Australia: International Speech Communication Association.

Rong, P., Yunusova, Y., Wang, J., & Green, J. R. (2015). Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach. *Behavioral Neurology, 2015,* 183027.

Rudzicz, F., Hirst, G., & van Lieshout, P. (2012) Vocal tract representation in the recognition of cerebral palsied speech. *Journal of Speech, Language, and Hearing Research, 55,* 1190–1207.

Shutts, R. E., Burke, K. S., & Creston, J. E. (1964). Derivation of twenty-five-word PB lists. *Journal of Speech and Hearing Disorders, 29,* 442–447.

Slud, E., Stone, M., Smith, P. J., & Goldstein, M., Jr. (2002). Principal components representation of the two-dimensional coronal tongue surface. *Phonetica, 59,* 108–133.

Story, B. H. (2011). TubeTalker: An airway modulation model of human sound production. In S. Fels & N. d'Alessandro (Eds.), *Proceedings of the First International Workshop on Performative Speech and Singing Synthesis* (pp. 1–8). Vancouver, Canada: P3S 2011.

Suemitsu, A., Dang, J., Ito, T., & Tiede, M. (2015). A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning. *The Journal of Acoustical Society of America, 138*(4), EL382–EL387.

Wang, J. (2011). *Silent speech recognition from articulatory motion* (Unpublished doctoral dissertation). University of Nebraska–Lincoln, Lincoln, NE.

Wang, J., Balasubramanian, A., Mojica de La Vega, L., Green, J. R., Samal, A., & Prabhakaran, B. (2013). Word recognition from continuous articulatory movement time-series data using symbolic representations. In *4th Workshop on Speech and Language Processing for Assistive Technologies* (pp. 119–127). Grenoble, France: Association for Computational Linguistics.

Wang, J., Green, J. R., & Samal, A. (2013). Individual articulator's contribution to phoneme production. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 7785–7789). Vancouver, Canada: Institute of Electrical and Electronics Engineers.

Wang, J., Green, J. R., Samal, A., & Marx, D. B. (2011). Quantifying articulatory distinctiveness of vowels. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 277–280). Florence, Italy: INTERSPEECH.

Wang, J., Green, J. R., Samal, A., & Yunusova, Y. (2013). Articulatory distinctiveness of vowels and consonants: A data-driven approach. *Journal of Speech, Language, and Hearing Research, 56,* 1539–1551.

Wang, J., & Hahm, S. (2015). Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training. In S. Möller, H. Ney, B. Möbius, E. Nöth, & S. Steidl (Eds.), *Interspeech 2015—16th Annual Conference of the International Speech Communication Association*

(pp. 2415–2419). Dresden, Germany: International Speech Communication Association.

Wang, J., Hahm, S., & Mau, T. (2015). Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition. In *6th Workshop on Speech and Language Processing for Assistive Technologies* (pp. 79–85). Dresden, Germany: Association for Computational Linguistics.

Wang, J., Katz, W. F., & Campbell, T. F. (2014). Contribution of tongue lateral to consonant production. In H. Li, H. Meng, B. Ma, E. S. Chng, & L. Xie (Eds.), *Interspeech 2014—15th Annual Conference of the International Speech Communication Association* (pp. 174–178). Singapore: International Speech Communication Association.

Wang, J., Samal, A., & Green, J. R. (2014). Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph. In *4th Workshop on Speech and Language Processing for Assistive Technologies* (pp. 38–45). Baltimore, MD: Association for Computational Linguistics.

Wang, J., Samal, A., Green, J. R., & Carrell, T. D. (2009). Vowel recognition from articulatory position time-series data. In B. J. Wysocki & T. A. Wysocki (Eds.), *Proceedings of the 3rd International Conference on Signal Processing and Communication Systems* (p. 1–6). Omaha, NE: Institute of Electrical and Electronics Engineers.

Wang, J., Samal, A., Green, J. R., & Rudzicz, F. (2012a). Sentence recognition from articulatory movements for silent speech interfaces. In *2012 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 4985–4988). Kyoto, Japan: Institute of Electrical and Electronics Engineers.

Wang, J., Samal, A., Green, J. R., & Rudzicz, F. (2012b). Whole-word recognition from articulatory movements for silent speech interfaces. In *Interspeech 2012—13th Annual Conference of the International Speech Communication Association* (pp. 1327–1330). Portland, OR: International Speech Communication Association.

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook*. Madison, WI: University of Wisconsin.

Wrench, A. A. (2000). A multi-channel/multi-speaker articulatory database for continuous speech recognition research. *Phonus, 5,* 1–13.

Yehia, H., & Tiede, M. (1997). A parametric three-dimensional model of the vocal-tract based on MRI data. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3* (pp. 1619–1625). Munich, Germany: IEEE Computer Society Press.

Yunusova, Y., Green, J. R., & Mefferd, A. (2009). Accuracy assessment for AG500, electromagnetic articulograph. *Journal of Speech, Language, and Hearing Research, 52,* 547–555.

Yunusova, Y., Green, J. R., Wang, J., Pattee, G., & Zinman, L. (2011). A protocol for comprehensive assessment of bulbar dysfunction in amyotrophic lateral sclerosis (ALS). *Journal of Visualized Experiments, 48,* 2422.

Yunusova, Y., Weismer, G. G., & Lindstrom, M. J. (2011). Classification of vocalic segments from articulatory kinematics: Healthy controls and speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 54,* 1302–1311.

Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research, 51,* 596–611.

Zerling, J. P. (1979). *Articulation et coarticulation dans les groups occlusive-voyelle en français* [Articulation and coarticulation in occlusive-vowel groups in French] (Unpublished doctoral dissertation). Université de Nancy, Nancy, France.