INTERSPEECH 2014

Across-speaker Articulatory Normalization for Speaker-independent Silent Speech Recognition

Jun Wang [1, 2], *Ashok Samal* [3], *Jordan R. Green* [4]

[1] Department of Bioengineering
[2] Callier Center for Communication Disorders
University of Texas at Dallas, Dallas, Texas, USA
[3] Department of Computer Science & Engineering
University of Nebraska-Lincoln, Lincoln, Nebraska, USA
[4] Department of Communication Sciences & Disorders
MGH Institute of Health Professions, Boston, Massachusetts, USA
wangjun@utdallas.edu, samal@cse.unl.edu, jgreen2@mghihp.edu

Abstract

Silent speech interfaces (SSIs), which recognize speech from articulatory information (i.e., without using audio information), have the potential to enable persons with laryngectomy or a neurological disease to produce synthesized speech with a natural sounding voice using their tongue and lips. Current approaches to SSIs have largely relied on speaker-dependent recognition models to minimize the negative effects of talker variation on recognition accuracy. Speaker-independent approaches are needed to reduce the large amount of training data required from each user; only limited articulatory samples are often available for persons with moderate to severe speech impairments, due to the logistic difficulty of data collection. This paper reported an across-speaker articulatory normalization approach based on Procrustes matching, a bidimensional regression technique for removing translational, scaling, and rotational effects of spatial data. A dataset of short functional sentences was collected from seven English talkers. A support vector machine was then trained to classify sentences based on normalized tongue and lip movements. Speaker-independent classification accuracy (tested using leave-one-subject-out cross validation) improved significantly, from 68.63% to 95.90%, following normalization. These results support the feasibility of a speaker-independent SSI using Procrustes matching as the basis for articulatory normalization across speakers.

Index Terms: silent speech recognition, speech kinematics, Procrustes analysis, support vector machine

1. Introduction

The options for augmenting oral communication in persons with moderate to severe speech and voice disorders are currently very limited [1]. Persons with neurological speech disorders eventually abandon oral communication to rely on assistive devices that are much slower than normal speech. Although persons incapable of talking due to a laryngectomy (complete removal of the larynx) currently have several options to restore speech (i.e., esophageal speech, tracheo-esophageal speech, and electrolarynx), these approaches frequently produce abnormal sounding voice with a pitch that is abnormally low and limited in range [2].

In the future, silent speech interfaces (SSIs) may overcome these limitations by allowing people to generate natural sounding speech from the movements of their tongue and lips [3, 4]. SSIs have three basic components: a speech movement recorder, a speech movement recognizer [5], and a speech playback device [6, 7]. SSIs require that speech movements be recorded in real-time, and then rapidly mapped on to the intended units of speech including phonemes, words, or sentences. A variety of techniques have been used to record speech movements including ultrasound [8, 9], surface electromyography [10, 11], and electromagnetic articulograph (EMA) [12, 13]. The current project used EMA, which registers the 3D motion of sensors adhered to the tongue and lips. To date, most of the research on the recognition component has focused on developing speaker-dependent (within-speaker) approaches, where training data and testing data are from the same speaker. Speaker-dependent approaches have been used to mitigate the negative effects of inter-talker variation on the recognition of speech movements. This variation has multiple sources including gender, dialect, individual vocal tract anatomy, and different co-articulation patterns [14, 15] - all of which challenge attempts to develop accurate speaker-independent recognition models.

To minimize such inter-talker effects, researchers have normalized the articulatory movements of vowels [16, 17, 18, 21], consonants [19, 21], and pseudo-words [20] by aligning the tongue position to a reference (e.g., palate [16, 17], or a general tongue shape [19]). One promising approach is Procrustes matching, a bidimensional regression technique [22] that has been used to minimize the effect of translational, scaling, and rotational differences of articulatory data across speakers [20, 21].

This study examined the potential benefits of Procrustes-based normalization on speaker-independent silent speech recognition. Tongue and lip movements were recorded from multiple speakers while producing short functional sentences (e.g., How are you doing?). The time-varying positions of the tongue and lips were classified into sentences using a support vector machine. Speaker-independent classification models (using leave-one-subject-out cross validation) were used to compare the recognition accuracy of non-normalized and normalized speech movement data.

2. Method

2.1. Design of our silent speech interface

Figure 1 illustrates the three-component design of the SSI: (a) real-time articulatory data acquisition, (b) online silent speech recognition (converting articulation information to text), and (c) text-to-speech synthesis for speech output. The first

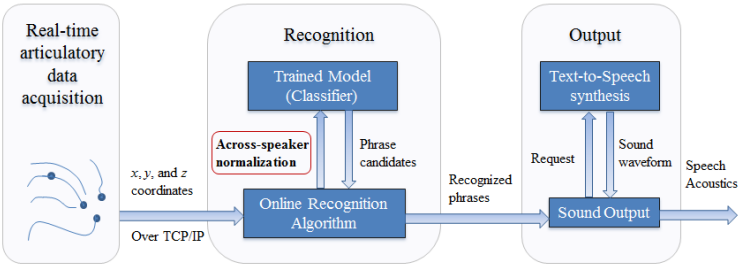

Figure 1. *Conceptual design of the speaker-independent silent speech interface, where an across-speaker articulatory normalization component is embedded in data preprocessing stage.*

component tracks articulatory motion in real-time using EMA. The second component recognizes a set of phrases from articulatory data (i.e., without using audio data). The third component generates synthesized sounds for the recognized phrases. For details of the SSI design, please refer [23, 24].

This study was to determine if across-speaker articulaory normalization, based on Procrustes-matching, is effective for enhancing the speaker-independent recognition accuracy. If effective, the normalization will be embedded into the online recognition component of the future SSI development.

2.2. Across-speaker articulatory normalization using Procrustes matching

2.2.1. Procrustes matching

Procrustes matching (or Procrustes analysis, Procrustes transformation [22]), is a robust bi-dimensional shape analysis. In Procrustes analysis, a shape is represented by a set of ordered landmarks on the surface of an object. Procrustes distance is calculated as the summed Euclidean distances between the corresponding landmarks of two shapes after the locational, rotational, and scaling effects are removed from the two shapes [25, 26].

Figure 2 shows an example of articulatory shape corresponding to "how are you doing?". The shape contains 40 landmarks that are discretized from the continuous motion paths of four sensors attached on tongue and lips, named as TT (Tongue Tip), TB (Tongue Body), UL (Upper Lip), and LL (Lower Lip). Section 3 will give details of the sensor setup.

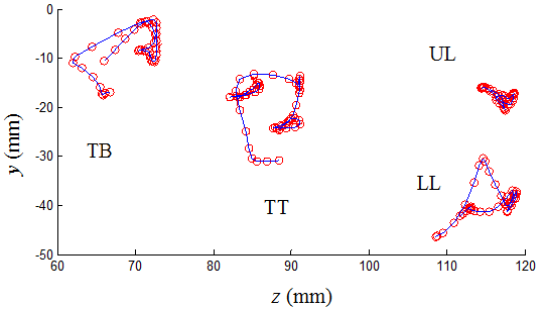

Figure 2. *Example of a shape (sampled motion path of four articulators) for producing "How are you doing?" Each curve is down-sampled to 40 points (indicated by red circles). In this coordinate system, y is vertical and z is anterior-posterior.*

Each shape is a 40 × 2 (y and z coordinates) array. A prior study has shown 40 data points are sufficient to capture the motion patterns in short phrases [27]. A step-by-step of Procrustes matching between two shapes includes (1) aligning the centroids of the two shapes, (2) scaling the shapes to a unit size, and (3) rotating one shape to match the other [25, 26].

Let S a set of landmarks as shown below.

$$S = \{(y_i, z_i)\}, i = 1...n \quad (1)$$

where (y_i, z_i) represents the i'th data point (spatial coordinates) of a sensor, and n is the total number of data points (n = 40 in Figure 2). The Procrustes matching can be described using parameters $\{(c_y, c_z), (\beta_y, \beta_z), \theta\}$. S is transformed to:

$$\begin{bmatrix} \overline{y_i} \\ \overline{z_i} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \beta_y \\ \beta_z \end{bmatrix} \begin{bmatrix} y_i - c_y \\ z_i - c_z \end{bmatrix} \quad (2)$$

where (c_y, c_z) are the translation factors; Scaling factor β is done for each dimension separately, which is the square root of the sum of the squares of all data points along the dimension; θ is the angle to rotate [22].

2.2.2. Across-speaker articulatory normalization

In addition to the removal of translational, scaling, and rotational effects in typical Procrustes matching [20, 21], our normalization approach transformed each participant's articulatory shape into an "normalized shape", which had a centroid at the origin (0, 0), a unit size, and aligned to the vertical line formed by the average positions (centroids) of the upper and lower lips.

The normalization procedure was done in three steps. First, all articulatory data (e.g., a shape in Figure 2) of each speaker were translated to the centroid of that shape (average position of all data points in the shape). This step removed the locational effects between speakers (see Figure 3 left panel). Second, the articulatory data from each speaker were scaled to unit size. This step reduced the effect of difference in vocal tract and tongue sizes of talkers. Third, shapes of speakers were rotated to make sure the sagittal plane was oriented such that the centroid of lower and upper lip movements defined the vertical axis. *This step reduced the variation of rotational effects due to the facial anatomy difference between speakers.* Thus, in Equation 2, (c_y, c_z) are the centroid of shape S; Scaling factor (β_y, β_z) is the square root of the sum of all data points along each dimension of S; θ is angle of the S to the reference shape in which upper lip and lower lip form a vertical line (right panel in Figure 3).

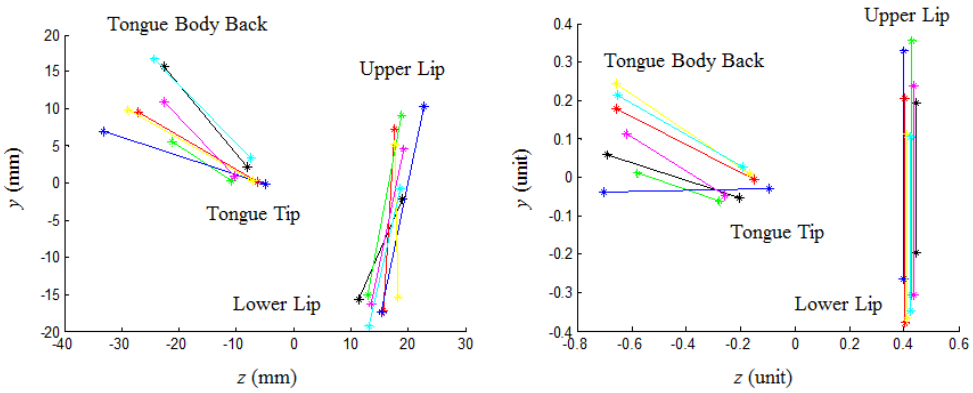

Figure 3. *Centroids of each articulator (TT, TB, UL, and LL) of the seven talkers producing a short phrase "how are you doing?" Left panel is translated data (centroids of the whole shapes are at zero point); Right panel is translated, scaled (to unit size), and rotated (UL and LL are in a vertical line) data.*

Figure 3 illustrates the normalization approach using the centroids of the motion paths of each sensor of all speakers. In the left panel, the shapes composed by the discretized motion paths of four sensors (i.e., TT, TB, UL, and LL) were translated (moved to the zero point). In the right panel, the shapes were translated, scaled (to unit size), and rotated to an angle such that upper lip and lower lip form a vertical line.

2.3. Evaluation: Classification using SVM

Support vector machines (SVMs) are widely used soft margin classifiers that find separating hyperplanes with maximal margins between classes in a high dimensional space [28]. A radial basis function (RBF) was used as the kernel function in this experiment, where λ is an empirical parameter (Equation 3). A kernel function is used for describing the distance between two data points (i.e., *u* and *v* in Equation 3).

$$K_{RBF}(u,v) = \exp(1-\lambda \parallel u-v \parallel) \qquad (3)$$

LIBSVM [29], a widely used implementation of SVM, was used in this experiment. SVMs have been successfully used in silent speech recognition by classifying phonemes [4, 30], words [23, 24], and phrases [27] from articulatory movement data. In this experiment, a SVM was used to classify the phrases collected from multiple speakers under three different configurations: speaker-dependent, speaker-independent without normalization, and speaker-independent with normalization. After the normalization, *y* and *z* coordinates of each individual sensor were concatenated as a one-dimensional vector that was fed into the SVM [27]. Cross-validation was used in all three configurations. Leave-one-sample-out cross validation was used in the speaker-dependent setting. Leave-one-subject-out cross validation was used in the speaker-independent settings, where all samples from one speaker was used for testing and the rest data from other speakers were used for training in each execution. The classification accuracies were used to evaluate the effectiveness of the across-speaker normalization approach.

3. Data collection

3.1. Participants and stimuli

Seven English talkers participated in the data collection. No history of speech, language, hearing, or cognitive problems were reported. Each speaker participated in one session in which he/she repeated a sequence of twelve short functional phrases. The phrases were selected from [27]. The participants were asked to pronounce the list of phrases and repeated the list multiple times at their habitual speaking rate and loudness.

3.2. Tongue motion tracking device

The electromagnetic articulograph (EMA) AG500 (Carstens Medizinelektronik GmbH, Bovenden, Germany) was used to collect the 3-D movement time-series data of the head, tongue, and lips for all participants. EMA records tongue movements by establishing a calibrated electromagnetic field in a cube that induces electric current into tiny sensor coils that are attached to the surface of the articulators. A similar data collection procedure has been used in [25, 27, 31]. The spatial precision of motion tracking using EMA (AG500) is approximately 0.5 mm [32]. The sampling rate for recording is 200 Hz [32].

3.3. Procedure

Participants were seated with their head within a calibrated magnetic field. Seven sensors were attached to the surface of each articulator using dental glue (PeriAcryl 90, GluStitch) or tape. Before the beginning of actually data recording, a three-minute training session helped the participants to adapt to the wired sensors. Previous studies have shown these sensors do not significantly affect their speech output [33].

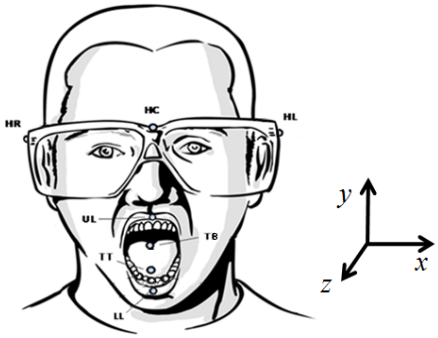

Figure 4. *Positions of sensors. Sensor labels are described in text.*

Figure 4 shows the positions of the seven sensors attached to a participant's head, tongue, and lips. HC (Head Center) was on the bridge of the glasses; HL (Head Left) and HR (Head Right) were on the left and right outside edge of each lens, respectively. The movements of HC, HL, and HR were used to calculate the movements of other articulators independent of the head. TT (Tongue Tip), TB (Tongue Body Back) were attached at the midline of the tongue [34, 35]. TT was about approximately 10 mm from the actual tongue tip. TB was as far back as possible and about 30 to 40mm from the tongue apex [25]. Lip sensors were attached to the vermilion borders of the upper (UL) and lower (LL) lips at midline. The data of TT, TB, UL, and LL were used for analysis, which was found optimal for this application [36].

3.4. Data preprocessing

The raw sensor position data were preprocessed prior to analysis. First, the head translations and rotations were subtracted from the tongue and lip locations to derive head-independent measurements of the analysis variables. The orientation of the derived 3-D Cartesian coordinate system is displayed in Figure 4, in which x is left-right, y is vertical, z is front-back. Second, a low pass filter (i.e., 20 Hz) [25, 31] was applied for removing noise.

Only y and z coordinates (Figure 4) of the tongue and lip sensors were used for analysis because the movement along the x axis is not significant in normal speech production [25, 35]. The center of the magnetic field is the origin (zero point) of the EMA coordinate system.

Samples affected by mispronunciations or sensor defects were rare, but were excluded from the experiment. In all, 2,076 samples (12 phrases × 7 speakers × 24.7 samples per phrase per speaker) were obtained and used in this study.

4. Results & Discussion

Classification accuracy. Figure 5 gives the classification accuracies of the three configurations: speaker-dependent, speaker-independent without normalization, and speaker-independent with normalization. Paired t-test was used to test the significance of the differences in the accuracies between the approaches. The recognition accuracy for the speaker-dependent approach (94.31%) was significantly higher ($p < 0.01$) than for the speaker-independent, unnormalized approach (68.63%). These findings provided evidence of the large variation in the speech movements across speakers. The recognition accuracy for the speaker-independent with normalization approach (95.90%) was also significantly higher ($p < 0.01$) than for the speaker-independent without normalization approach. These results suggest that our Procrustes matching-based approach was effective for across-speaker articulatory normalization.

Adaptability for online speaker-independent recognition. Although the experiment was conducted offline (data were collected before the analysis), the normalization approach can be easily integrated into online speaker-independent silent speech recognition (Figure 1), because this approach involves only three preprocessing steps: translation (move the data sample to zero point), scaling (to unit size), and rotation (so that upper lip and lower lip centroids can form a vertical line). The approach is also advantageous for other online applications because it does not require pre-recorded training data. Of course, EMA system is currently too cumbersome for clinical applications. However, our articulatory normalization approach is not dependent on a particular data acquisition device and can be easily applied to any portable devices when they are ready in the future.

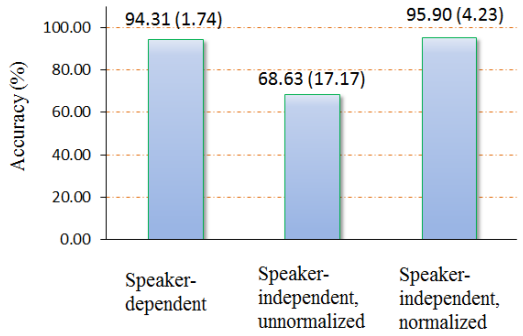

Figure 5: *Average phrase classification accuracy of different configurations. Standard deviations are in parenthesis.*

Other implications. This articulatory normalization approach, although only applied to improving speech movement recognition in this study, may have implications for other related applications. For example, the normalization may be useful for scientific studies of tongue kinematics during speech [25], recognition of speech with articulatory data for healthy [37] and disordered populations [38], EMA-based speech therapy [39], and tongue motion as visual feedback for secondary language pronunciation learning [40]).

Limitations. Although the experimental results were encouraging, the data set used in the experiment contained only a small number of unique phrases. Further studies with a larger vocabulary (including phonemes and words) from a larger number of subjects with different genders and dialects are necessary to explore the limits of the current approach.

5. Conclusions & Future Work

This paper investigated the across-speaker articulatory normalization based on Procrustes matching for speaker-independent silent speech recognition. Seven native English speakers participated in this study, in which short phrases were produced. Experimental results showed the effectiveness of the normalization approach for improving the accuracy of speaker-independent silent speech recognition.

Future work includes articulatory normalization for each individual articulator [20], testing the normalization approach when used in a real-time silent speech interface [13], and evaluating the efficacy of the approach for subjects with speech impairment (e.g., after laryngectomy).

6. Acknowledgements

This work was in part supported by Callier Excellence in Education Fund, University of Texas at Dallas, and grants awarded by the National Institutes of Health (R01 DC009890 and R01 DC013547). We would like to thank Dr. Tom Carrell, Dr. Lori Synhorst, Dr. Mili Kuruvilla, Cynthia Didion, Cara Ullman, Rebecca Hoesing, Kate Lippincott, Kayanne Hamling, Kelly Veys, and Toni Hoffer for their contribution to subject recruitment, data collection, management, and processing.

7. References

[1] Bailey, B. J., Johnson, J. T., and Newlands, S. D., *Head and Neck Surgery – Otolaryngology*, Lippincot, Williams & Wilkins, Philadelphia, PA, USA, 4th Ed., 1779-1780, 2006.

[2] Liu, H., & Ng, M. L. "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, 34(3):327-332, 2007.

[3] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., and Brumberg, J. S. "Silent speech interfaces", *Speech Communication*, 52:270-287, 2010.

[4] Wang, J., Samal, A., Green, J. R., & Carrell, T. D. "Vowel recognition from articulatory position time-series data," *Proc. IEEE Intl. Conf. on Signal Processing and Communication Systems*, Omaha, NE, 1-6, 2009.

[5] Wang, J. *Silent speech recognition from articulatory motion*, Ph.D. Dissertation, Dept. of Comp. Sci., Univ. of Nebraska, Lincoln, NE, 2011.

[6] Huang, X. D., Acero, A., Hon, H.-W., Ju, Y.-C., Liu, J., Meredith, S., and Plumpe, M., "Recent Improvements on Microsoft's Trainable Text-to-Speech System: Whistler," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 959-962, 1997.

[7] Khan, Z., A., Green, P., Creer, S., and Cunningham, S. "Reconstructing the voice of an individual following laryngectomy," *Augmentative and Alternative Communication*, 27(1):61-66, 2011.

[8] Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips", *Speech Communication*, 52:288–300, 2010.

[9] Denby, B., Cai, J., Roussel, P., Dreyfus, G., Crevier-Buchman, L., Pillot-Loiseau, C., Hueber, and T., Chollet, G., "Tests of an interactive, phrasebook-style post-laryngectomy voice-replacement system", *the 17th International Congress on Phonetic Sciences*, Hong Kong, China, 572-575, 2011.

[10] Jorgensen, C. and Dusan, S., "Speech interfaces based upon surface electromyography", *Speech Communication*, 52:354–366, 2010.

[11] Heaton, J. T., Robertson, M., and Griffin, C., "Development of a wireless electromyographically controlled electrolarynx voice prosthesis", *Proc. IEEE Intl. Conf. on Engineering in Medicine & Biology Society*, Boston, MA, 5352-5355, 2011.

[12] Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E., and Chapman, P. M., "Development of a (silent) speech recognition system for patients following laryngectomy", *Medical Engineering & Physics*, 30(4):419-425, 2008.

[13] Wang, J., Samal, A., and Green, J. R., "Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph," *ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2014 (In press).

[14] Kent, R. D., Adams, S. G., and Tuner, G. S. *Models of speech production*. Lass, N. J. : Principles of experimental Phonetics. Mosby, 1996.

[15] Kent, R. D., and Minifie, F. D., "Coarticulation in recent speech production models", *Journal of Phonetics*, 5(2):115–133, 1977.

[16] Hashi, M., Westbury, J. R., and Honda, K., "Vowel posture normalization," *Journal of Acoustical Society of America*, 104(4):2426-2437, 1982.

[17] Johnson, K., Ladefoged, P., and Lindau, M., "Individual differences in vowel production," *Journal of Acoustical Society of America*, 94(2):701-714, 1993.

[18] Simpson, A. P., *Gender-specific differences in the articulatory and acoustics realization of interword vowel sequences in American English*, In Philip Hoole, Masaaki Honda, & Christine Mooshammer (eds.), 5th Seminar on Speech Production: Models and Data. Kloster Seeon, 209-212, 2000.

[19] Westbury, J. H., Hashi, M., and Lindstrom, M. J., "Differences among speakers in lingual articulation for American English /r/," Speech *Communication*, 26(3):203-226, 1998.

[20] Felps, D. and Gutierrez-Osuna, R., "Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis," Technical Report, Texas A&M University, 2010-5-3.

[21] Li, S. and Wang, L., "Cross linguistic comparison of Mandarin and English EMA articulatory data," *Interspeech*, Portland, OR, 903-906, 2012.

[22] Dryden I. L. and Mardia, K.V., *Statistical shape analysis*, Hoboken: John Wiley, 1998.

[23] Wang, J., Samal, A., Green, J. R., and Rudzicz, F., "Whole-word recognition from articulatory movements for silent speech interfaces", *Interspeech*, Portland, OR, 1327-30, 2012.

[24] Wang, J., Balasubramanian, A., Mojica de La Vega, L., Green, J. R., Samal, A., and Prabhakaran, B., "Word recognition from continuous articulatory movement time-series data using symbolic representations," *ACL/ISCA Interspeech Workshop on Speech and Language Processing for Assistive Technologies*, Grenoble, France, 119-127, 2013.

[25] Wang, J., Green, J. R., Samal, A. and Yunusova, Y. "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, 56:1539-1551, 2013.

[26] Wang, J., Green, J. R., Samal, A., and Marx, D. B. "Quantifying articulatory distinctiveness of vowels", *Interspeech*, Florence, Italy, 277-280, 2011.

[27] Wang, J., Samal, A., Green, J. R., and Rudzicz, F., "Sentence recognition from articulatory movements for silent speech interfaces", *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 4985-4988, 2012.

[28] Boser, B., Guyon, I., Vapnik, V., "A training algorithm for optimal margin classifiers", *Conf. on Learning Theory (COLT)*, 144-152, 1992.

[29] Chang, C. -C., and Lin. C. -J., "LIBSVM: a library for support vector machines", *ACM Trans. on Intelligent Systems and Technology*, 2(27):1-27, 2011.

[30] Wang, J., Green, J. R., Samal, A., and Carrell, T. D., "Vowel recognition from continuous articulatory movements for speaker-dependent applications," *Proc. IEEE Intl. Conf. on Signal Processing and Communication Systems*, 1-7, 2010.

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

[32] Yunusova, Y., Green, J. R., and Mefferd, A., "Accuracy assessment for AG500 electromagnetic articulograph", *Journal of Speech, Language, and Hearing Research*, 52(2):547-555, 2009.

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

[34] Green, J. R. and Wang, Y., "Tongue-surface movement patterns during speech and swallowing", *Journal of Acoustical Society of America*, 113:2820-2833, 2003.

[35] Westbury, J. *X-ray microbeam speech production database user's handbook*. University of Wisconsin, 1994.

[36] Wang, J., Green, J. R., & Samal, A., "Individual articulator's contribution to phoneme production", *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 7795-89, 2013.

[37] King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M., "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, 121(2):723-742, 2007.

[38] Rudzicz, F., "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):947-960, 2011.

[39] Katz, W. F., Campbell, T. F., Wang, J., Farrar, E., Eubanks, J. C., Balasubramanian, A., Prabhakaran, B., and Rennaker, R., "Opti-speech: A real-time, 3D visual feedback system for speech training", *Interspeech*, Singapore, 2014 (In press).

[40] Ouni, S. "Tongue control and its implication in pronunciation training," Computer Assisted Language Learning, 16:1-15, 2013.