Preliminary Test of a Real-Time, Interactive Silent Speech Interface Based on Electromagnetic Articulograph

Jun Wang
Dept. of Bioengineering
Callier Center for Communication Disorders
University of Texas at Dallas
wangjun@utdallas.edu

Ashok Samal
Dept. of Computer Science & Engineering
University of Nebraska-Lincoln
samal@cse.unl.edu

Jordan R. Green
Dept. of Communication Sciences & Disorders
MGH Institute of Health Professions
jgreen2@mghihp.edu

Abstract

A silent speech interface (SSI) maps articulatory movement data to speech output. Although still in experimental stages, silent speech interfaces hold significant potential for facilitating oral communication in persons after laryngectomy or with other severe voice impairments. Despite the recent efforts on silent speech recognition algorithm development using offline data analysis, online test of SSIs have rarely been conducted. In this paper, we present a preliminary, online test of a real-time, interactive SSI based on electromagnetic motion tracking. The SSI played back synthesized speech sounds in response to the user's tongue and lip movements. Three English talkers participated in this test, where they mouthed (silently articulated) phrases using the device to complete a phrase-reading task. Among the three participants, 96.67% to 100% of the mouthed phrases were correctly recognized and corresponding synthesized sounds were played after a short delay. Furthermore, one participant demonstrated the feasibility of using the SSI for a short conversation. The experimental results demonstrated the feasibility and potential of silent speech interfaces based on electromagnetic articulograph for future clinical applications.

1 Introduction

Daily communication is often a struggle for persons who have undergone a laryngectomy, a surgical removal of the larynx due to the treatment of cancer (Bailey et al., 2006). In 2013, about 12,260 new cases of laryngeal cancer were estimated in the United States (American Cancer Society, 2013). Currently, there are only limited treatment options for these individuals including (1) esophageal speech, which involves oscillation of the esophagus and is difficult to learn; (2) tracheo-esophageal speech, in which a voice prosthesis is placed in a tracheo-esophageal puncture; and (3) electrolarynx, an external device held on the neck during articulation, which produces a robotic voice quality (Liu and Ng, 2007). Perhaps the greatest disadvantage of these approaches is that they produce abnormal sounding speech with a fundamental frequency that is low and limited in range. The abnormal voice quality output severely affects the social life of people after laryngectomy (Liu and Ng, 2007). In addition, the tracheo-esophageal option requires an additional surgery, which is not suitable for every patient (Bailey et al., 2006). Although research is being conducted on improving the voice quality of esophageal or electrolarynx speech (Doi et al., 2010; Toda et al., 2012), new assistive technologies based on non-audio information (e.g., visual or articulatory information) may be a good alternative approach for providing natural sounding speech output for persons after laryngectomy.

Visual speech recognition (or automatic lip reading) typically uses an optical camera to obtain lip and/or facial features during speech (including lip contour, color, opening, movement, etc.) and then classify these features to speech units (Meier et al., 2000; Oviatt, 2003). However, due to the lack of information from tongue, the primary articulator, visual speech recognition (i.e., using visual information only, without tongue and audio information) may obtain a low accuracy (e.g., 30% - 40% for phoneme classification, Livescu et al., 2007). Furthermore, Wang and colleagues (2013b) have showed any single tongue sensor (from tongue tip to tongue body

Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), pages 38–45,
Baltimore, Maryland USA, August 26 2014.

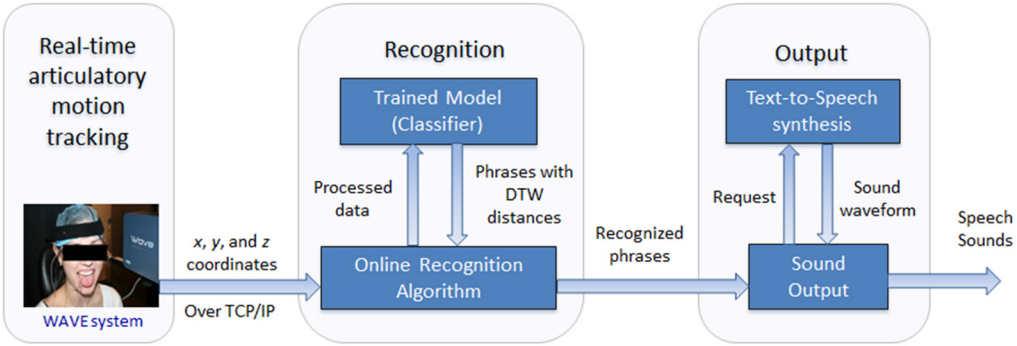

Figure 1. Design of the real-time silent speech interface.

back on the midsagittal line) encodes significantly more information in distinguishing phonemes than do lips. However, visual speech recognition is well suited for applications with small-vocabulary (e.g., a lip-reading based command-and-control system for home appliance) or using visual information as an additional source for acoustic speech recognition, referred to as audio-visual speech recognition (Potamianos et al., 2003), because such a system based on portable camera is convenient in practical use. In contrast, SSIs, with tongue information, have potential to obtain a high level of silent speech recognition accuracy (without audio information). Currently, two major obstacles for SSI development are lack of (a) fast and accurate recognition algorithms and (b) portable tongue motion tracking devices for daily use.

SSIs convert articulatory information into text that drives a text-to-speech synthesizer. Although still in developmental stages (e.g., speaker-dependent recognition, small-vocabulary), SSIs even have potential to provide speech output based on prerecorded samples of the patient's own voice (Denby et al., 2010; Green et al., 2011; Wang et al., 2009). Potential articulatory data acquisition methods for SSIs include ultrasound (Denby et al., 2011; Hueber et al., 2010), surface electromyography electrodes (Heaton et al., 2011; Jorgensen and Dusan, 2010), and electromagnetic articulograph (EMA) (Fagan et al., 2008; Wang et al., 2009, 2012a).

Despite the recent effort on silent speech interface research, online test of SSIs has rarely been studied. So far, most of the published work on SSIs has focused on development of silent speech recognition algorithm through offline analysis (i.e., using prerecorded data) (Fagan et al., 2008; Heaton et al., 2011; Hofe et al., 2013; Hueber et al., 2010; Jorgenson et al., 2010; Wang et al., 2009a, 2012a, 2012b, 2013c). Ultrasound-based SSIs have been tested online with multiple subjects and encouraging results were obtained in a phrase reading task where the subjects were asked to silently articulate sixty phrases (Denby et al., 2011). SSI based on electromagnetic sensing has been only tested using offline analysis (using pre-recorded data) collected from single subjects (Fagan et al., 2008; Hofe et al., 2013), although some work simulated online testing using prerecorded data (Wang et al., 2012a, 2012b, 2013c). Online tests of SSIs using electromagnetic articulograph with multiple subjects are needed to show the feasibility and potential of the SSIs for future clinical applications.

In this paper, we report a preliminary, online test of a newly-developed, real-time, and interactive SSI based on a commercial EMA. EMA tracks articulatory motion by placing small sensors on the surface of tongue and other articulators (e.g., lips and jaw). EMA is well suited for the early state of SSI development because it (1) is non-invasive, (2) has a high spatial resolution in motion tracking, (3) has a high sampling rate, and (4) is affordable. In this experiment, participants used the real-time SSI to complete an online phrase-reading task and one of them had a short conversation with another person. The results demonstrated the feasibility and potential of SSIs based on electromagnetic sensing for future clinical applications.

2 Design

2.1 Major design

Figure 1 illustrates the three-component design of the SSI: (a) real-time articulatory motion tracking using a commercial EMA, (b) online silent speech recognition (converting articulation information to text), and (c) text-to-speech synthesis for speech output.

The EMA system (Wave Speech Research

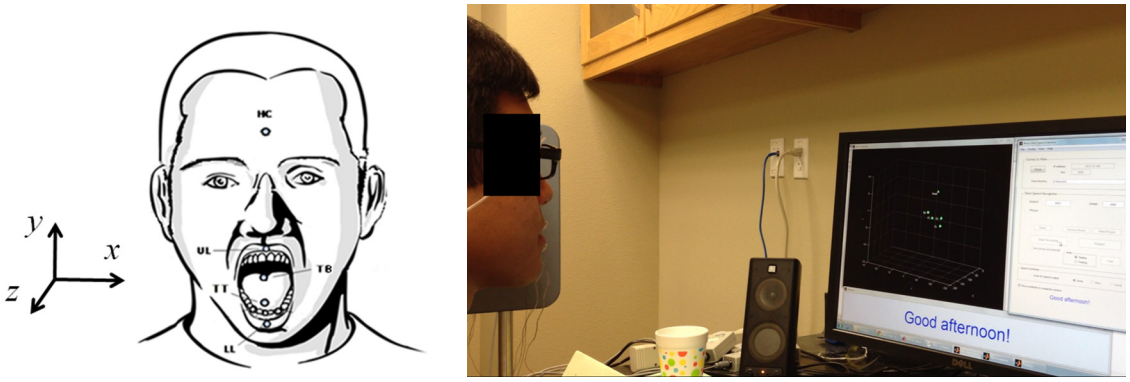

Figure 2. Demo of a participant using the silent speech interface. The left picture illustrates the coordinate system and sensor locations (sensor labels are described in text); in the right picture, a participant is using the silent speech interface to finish the online test.

system, Northern Digital Inc., Waterloo, Canada) was used to track the tongue and lip movement in real-time. The sampling rate of the Wave system was 100 Hz, which is adequate for this application (Wang et al., 2012a, 2012b, 2013c). The spatial accuracy of motion tracking using Wave is 0.5 mm (Berry, 2011).

The online recognition component recognized functional phrases from articulatory movements in real-time. The recognition component is modular such that alternative classifiers can easily replace and be integrated into the SSI. In this preliminary test, recognition was speaker-dependent, where training and testing data were from the same speakers.

The third component played back either pre-recorded or synthesized sounds using a text-to-speech synthesizer (Huang et al., 1997).

2.2 Other designs

A *graphical user interface* (GUI) is integrated into the silent speech interface for ease of operation. Using the GUI, users can instantly re-train the recognition engine (classifier) when new training samples are available. Users can also switch output voice (e.g., male or female).

Data transfer through TCP/IP. Data transfer from the Wave system to the recognition unit (software) is accomplished through TCP/IP, the standard data transfer protocols on Internet. Because data bandwidth requirement is low (multiple sensors, multiple spatial coordinates for each sensor, at 100 Hz sampling rate), any 3G or faster network connection will be sufficient for future use with wireless data transfer.

Extensible (closed) vocabulary. In the early stage of this development, closed-vocabulary silent speech recognition was used; however, the vocabulary is extensible. Users can add new phrases into the system through the GUI. Adding a new phrase in the vocabulary is done in two steps. The user (the patient) first enters the phrase using a keyboard (keyboard input can also be done by an assistant or speech pathologist), and then produces a few training samples for the phrase (a training sample is articulatory data labeled with a phrase). The system automatically re-trains the recognition model integrating the newly-added training samples. Users can delete invalid training samples using the GUI as well.

2.3 Real-time data processing

The tongue and lip movement positional data obtained from the Wave system were processed in real-time prior to being used for recognition. This included the calculation of head-independent positions of the tongue and lip sensors and low pass filtering for removing noise.

The movements of the 6 DOF head sensor were used to calculate the head-independent movements of other sensors. The Wave system represents object orientation or rotation (denoted by *yaw*, *pitch*, and *roll* in Euler angles) in quaternions, a four-dimensional vector. Quaternion has its advantages over Euler angles. For example, quaternion avoids the issue of gimbal lock (one degree of freedom may be lost in a series of rotation), and it is simpler to achieve smooth interpolation using quaternion than using Euler angles (Dam et al., 1998). Thus, quaternion has been widely used in computer graphics, computer vision, robotics, virtual reality, and flight dynamics (Kuipers, 1999). Given the unit quaternion

$$q = (a, b, c, d) \quad (1)$$

where $a^2 + b^2 + c^2 + d^2 = 1$, a 3×3 rotation matrix R can be derived using Equation (2):

$$R = \begin{bmatrix} a^2+b^2-c^2-d^2 & 2bc-2ad & 2bd+2ac \\ 2bc+2ad & a^2-b^2+c^2-d^2 & 2cd-2ab \\ 2bd-2ac & 2cd+2ab & a^2-b^2-c^2+d^2 \end{bmatrix} \quad (2)$$

For details of how the quaternion is used in Wave system, please refer to the Wave Real-Time API manual and sample application (Northern Digital Inc., Waterloo, Canada).

3 A Preliminary Online Test

3.1 Participants & Stimuli

Three American English talkers participated in this experiment (two males and one female with average age 25 and SD 3.5 years). No history of speech, language, hearing, or any cognitive problems were reported.

Sixty phrases that are frequently used in daily life by healthy people and AAC (augmentative and alternative communication) users were used in this experiment. Those phrases were selected from the lists in Wang et al., 2012a and Beukelman and Gutmann, 1999.

3.2 Procedure

Setup

The Wave system tracks the motion of sensors attached on the articulators by establishing an electromagnetic field by a textbook-sized generator. Participants were seated with their head within the calibrated magnetic field (Figure 2, the right picture), facing a computer monitor that displays the GUI of the SSI. The sensors were attached to the surface of each articulator using dental glue (PeriAcryl Oral Tissue Adhesive). Prior to the experiment, each subject participated in a three-minute training session (on how to use the SSI), which also helped them adapt to the oral sensors. Previous studies have shown those sensors do not significantly affect their speech output after a short practice (Katz et al., 2006; Weismer and Bunton, 1999).

Figure 2 (left) shows the positions of the five sensors attached to a participant's forehead, tongue, and lips (Green et al., 2003; 2013; Wang et al., 2013a). One 6 DOF (spatial and rotational) head sensor was attached to a nose bridge on a pair of glasses (rather than on forehead skin directly), to avoid the skin artifact (Green et al., 2007). Two 5 DOF sensors - TT (Tongue Tip) and TB (Tongue Body Back) - were attached on the midsagittal of the tongue. TT was located approximately 10 mm from the tongue apex (Wang et al., 2011, 2013a). TB was placed as far back as possible, depending on the participant's tongue length (Wang et al., 2013b). Lip movements were captured by attaching two 5 DOF sensors to the vermilion borders of the upper (UL) and lower (LL) lips at midline. The four sensors (i.e., TT, TB, UL, and LL) placements were selected based on literature showing that they are able to achieve as high recognition accuracy as that obtained using more tongue sensors for this application (Wang et al., 2013b).

As mentioned previously, real-time preprocessing of the positional time series was conducted, including subtraction of head movements from tongue and lip data and noise reduction using a 20 Hz low pass filter (Green et al., 2003; Wang et al., 2013a). Although the tongue and lip sensors are 5D, only the 3D spatial data (i.e., *x*, *y*, and *z*) were used in this experiment.

Training

The training step was conducted to obtain a few samples for each phrase. The participants were asked to silently articulate all sixty phrases twice at their comfortable speaking rate, while the tongue and lip motion was recorded. Thus, each phrase has at least two samples for training. Dynamic Time Warping (DTW) was used as the classifier in this preliminary test, because of its rapid execution (Denby et al., 2011), although Gaussian mixture models may perform well too when the number of training samples is small (Broekx et al., 2013). DTW is typically used to compare two single-dimensional time-series,

Training_Algorithm

Let $T_1 \ldots T_n$ be the sets of training samples for n phrases, where
$T_i = \{T_{i,1}, \ldots T_{i,j}, \ldots T_{i,mi}\}$ are m_i samples for phrase i.

```
for i = 1 to n     // n is the number of phrases
    L_i = sum(length(T_ij)) / m_i,  j = 1 to m_i;
    T = T_i1;      // first sample of phrase i
    for j = 2 to m_i
        (T', T'_ij) = MDTW(T, T_ij);
        T = (T' + T'_ij) / 2;//amplitude mean
        T = time_normalize(T, L_i);
    end
    R_i = T;   // representative sample for phrase i
end
Output(R);
```

Figure 3. Training algorithm using DTW. The function call MDTW() returns the average DTW distances between the corresponding sensors and dimensions of two data samples.

thus we calculated the average DTW distance across the corresponding sensors and dimensions of two data samples. DTW was adapted as follows for training.

The training algorithm generated a *representative* sample based on all available training samples for each phrase. Pseudo-code of the training algorithm is provided in Figure 3, for the convenience of description. For each phrase *i*, first, MDTW was applied to the first two training samples, $T_{i,1}$ and $T_{i,2}$. Sample T is the amplitude-mean of the warped samples $T'_{i,1}$ and $T'_{i,2}$ (time-series) (Line 5). Next, T was time-normalized (stretched) to the average length of all training samples for this phrase (L_i), which was to reduce the effects of duration change caused by DTW warping (Line 6). The procedure continued until the last training sample $T_{i, mi}$ (m_i is the number of training samples for phrase *i*). The final T was the representative sample for phrase *i*.

The training procedure can be initiated by pressing a button on the GUI anytime during the use of the SSI.

Testing

During testing, each participant silently articulated the same list of phrases while the SSI recognized each phrase and played corresponding synthesized sounds. DTW was used to compare the test sample with the representative training sample for each phrase (R_i, Figure 3). The phrase that had the shortest DTW distance to the test sample was recognized. The testing was triggered by pressing a button on the GUI. If the phrase was incorrectly predicted, the participant was allowed to add *at most* two additional training samples for that phrase.

Figure 2 (right) demonstrates a participant is using the SSI during the test. After the participant silently articulated "*Good afternoon*", the SSI displayed the phrase on the screen, and played the corresponding synthesized sound simultaneously.

Finally, one participant used the SSI for a bidirectional conversation with an investigator. Since this prototype SSI has a closed-vocabulary recognition component, the participant had to choose the phrases that have been trained. This task was intended to provide a demo of how the SSI is used for daily communication. The script of the conversation is as below:

Investigator: *Hi DJ, How are you?*
Subject: *I'm fine. How are you doing?*
Investigator: *I'm good. Thanks.*
Subject: *I use a silent speech interface to talk.*
Investigator: *That's cool.*
Subject: *Do you understand me?*
Investigator: *Oh, yes.*
Subject: *That's good.*

Subject	Accuracy (%)	Latency (s)	# of Training Samples
S01	100	3.086	2.0
S02	96.67	1.403	2.4
S03	96.67	1.524	3.1

Table 1. Phrase classification accuracy and latency for all three participants.

4 Results and Discussion

Table 1 lists the performance using the SSI for all three participants in the online, phrase-reading task. The three subjects obtained a phrase recognition accuracy from 96.67% to 100.00%, with a latency from 1.403 second to 3.086 seconds, respectively. The high accuracy and relatively short delay demonstrated the feasibility and potential of SSIs based on electromagnetic articulograph.

The order of the participants in the experiment was S01, S02, and then S03. After the experiment of S01, where all three dimensional data (*x*, *y*, and *z*) were used, we decided to use only *y* and *z* for S02 and S03 to reduce the latency. As listed in Table 1, the latencies of S02 and S03 did significantly reduce, because less data was used for online recognition.

Surprisingly, phrase recognition without using *x* dimension (left-right) data led to a decrease of accuracy and more training samples were required; prior research suggests that tongue movement in this dimension is not significant during speech in healthy talkers (Westbury, 1994). This observation is supported by participant S01, who had the highest accuracy and needed fewer training samples for each phrase (column 3 in Table 1). S02 and S03 used data of only *y* and *z* dimensions. Of course, data from more subjects are needed to confirm the potential significance of the *x* dimension movement of the tongue to silent speech recognition accuracy.

Data transfer between the Wave machine and the SSI recognition component was done through TCP/IP protocols and in real-time. In the future, this design feature will allow the recognition component to run on a smart phone or any wearable devices with an Internet connection (Cellu-

lar or Wi-Fi). In this preliminary test, the individual delays caused by TCP/IP data transfer, online data preprocessing, and classification were not measured and thus unknown. The delay information may be useful for our future development that the recognition component is deployed on a smart-phone. A further study is needed to obtain and analyze the delay information.

The bidirectional dialogue by one of the participants demonstrated how the SSI can be used in daily conversation. To our best knowledge, this is the first conversational demo using a SSI. An informal survey to a few colleagues provided positive feedback. The conversation was smooth, although it is noticeably slower than a conversation between two healthy talkers. Importantly, the voice output quality (determined by the text-to-speech synthesizer) was natural, which strongly supports the major motivation of SSI research: to produce speech with natural voice quality that current treatments cannot provide. A video demo is available online (Wang, 2014).

The participants in this experiment were young and healthy. It is, however, unknown if the recognition accuracy may decrease or not for users after laryngectomy, although a single patient study showed the accuracy may decrease 15-20% compared to healthy talkers using an ultrasound-based SSI (Denby et al., 2011). Theoretically, the tongue motion patterns in (silent) speech after the surgery should be no difference with that of healthy talkers. In practice, however, some patients after the surgery may be under treatment for swallowing using radioactive devices, which may affect their tongue motion patterns in articulation. Thus, the performance of SSIs may vary and depend on the condition of the patients after laryngectomy. A test of the SSI using multiple participants after laryngectomy is needed to understand the performance of SSIs for those patients under different conditions.

Although a demonstration of daily conversation using the SSI is provided, SSI based on the non-portable Wave system is currently not ready for practical use. Fortunately, more affordable and portable electromagnetic devices are being developed as are small handheld or wearable devices (Fagan et al., 2008). Researchers are also testing the efficacy of permanently implantable and wireless sensors (Chen et al., 2012; Park et al., 2012). In the future, those more portable, and wireless articulatory motion tracking devices, when they are ready, will be used to develop a portable SSI for practice use.

In this experiment, a simple DTW algorithm was used to compare the training and testing phrases, which is known to be slower than most machine learning classifiers. Thus, in the future, the latency can be significantly reduced by using faster classifiers such as support vector machines (Wang et al., 2013c) or hidden Markov models (Heracleous and Hagita, 2011; King et al., 2007; Rudzicz et al., 2012; Uraga and Hain, 2006).

Furthermore, in this proof-of-concept design, the vocabulary was limited to a small set of phrases, because our design required the whole experiment (including training and testing) to be done in about one hour. Additional work is needed to test the feasibility of open-vocabulary recognition, which will be much more usable for people after laryngectomy or with other severe voice impairments.

5 Conclusion and Future Work

A preliminary, online test of a SSI based on electromagnetic articulograph was conducted. The results were encouraging revealing high phrase recognition accuracy and short playback latencies among three participants in a phrase-reading task. In addition, a proof-of-concept demo of bidirectional conversation using the SSI was provided, which shows how the SSI can be used for daily communication.

Future work includes: (1) testing the SSI with patients after laryngectomy or with severe voice impairment, (2) integrating a phoneme- or word-level recognition (open-vocabulary) using faster machine learning classifiers (e.g., support vector machines or hidden Markov models), and (3) exploring speaker-independent silent speech recognition algorithms by normalizing the articulatory movement across speakers (e.g., due to the anatomical difference of their tongues).

Acknowledgements

This work was in part supported by the 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. Thomas F. Campbell, Dr. William F. Katz, Dr. Gregory S. Lee, Dr. Jennell C. Vick, Lindsey Macy, Marcus Jones, Kristin J. Teplansky, Vedad "Kelly" Fazel, Loren Montgomery, and Kameron Johnson for their support or assistance. We also thank the anonymous reviewers for their comments and suggestions for improving the quality of this paper.

References

American Cancer Society. 2013. *Cancer Facts and Figures 2013*. American Cancer Society, Atlanta, GA. Retrieved on February 18, 2014.

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

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

Beukelman, D. R., and Gutmann, M. 1999. Generic Message List for AAC users with ALS. http://aac.unl.edu/ALS_Message_List1.htm

Broekx, L., Dreesen, K., Gemmeke, J. F., and Van Hamme, H. 2013. Comparing and combining classifiers for self-taught vocal interfaces, *ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 21-28, 2013.

Chen, W.-H., Loke, W.-F., Thompson, G., and Jung, B. 2012. A 0.5V, 440uW frequency synthesizer for implantable medical devices, *IEEE Journal of Solid-State Circuits*, 47:1896-1907.

Dam, E. B., Koch, M., and Lillholm, M. 1998. *Quaternions, interpolation and animation*. Technical Report DIKU-TR-98/5, University of Copenhagen.

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

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

Doi, H., Nakamura, K., Toda, T., Saruwatari, H., Shikano, K. 2010. Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models, *IEICE Transactions on Information and Systems*, E93-D, 9:2472-2482.

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

Green, P. D., Khan, Z., Creer, S. M. and Cunningham, S. P. 2011. Reconstructing the voice of an individual following Laryngectomy, *Augmentative and Alternative Communication*, 27(1):61-66.

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

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

Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., and Rybchenko, S. I. 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing, *Speech Communication*, 55(1):22-32.

Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., and Rybchenko, S. I. 2011. Speech Synthesis Parameter Generation for the Assistive Silent Speech Interface MVOCA, *Proc. Interspeech*, 3009-3012.

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

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

Heaton, J. T., Robertson, M., and Griffin, C. 2011. Development of a wireless electromyographically controlled electrolarynx voice prosthesis, *Proc. of the 33rd Annual Intl. Conf. of the IEEE Engineering in Medicine & Biology Society*, Boston, MA, 5352-5355.

Heracleous, P., and Hagita, N. 2011. Automatic recognition of speech without any audio information, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2392-2395.

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

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

Kent, R. D., Adams, S. G., and Tuner, G. S. 1996. Models of speech production, in *Principles of Experimental Phonetics*, Ed., Lass, N. J., Mosby: St Louis, MO.

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

Kuipers, J. B. 1999. *Quaternions and rotation Sequences: a Primer with Applications to Orbits, Aerospace, and Virtual Reality*, Princeton University Press, Princeton, NJ.

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

Livescu, K., Çetin, O., Hasegawa-Johnson, Mark, King, S., Bartels, C., Borges, N., Kantor, A., et al. (2007). Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 621-624.

Meier, U., Stiefelhagen, R., Yang, J., and Waibel, A. (2000). Towards Unrestricted Lip Reading. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(5): 571-585.

Oviatt, S. L. 2003. Multimodal interfaces, in *Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, Eds. Julie A. Jacko and Andrew Sears (Mahwah, NJ:Erlbaum): 286–304.

Park, H., Kiani, M., Lee, H. M., Kim, J., Block, J., Gosselin, B., and Ghovanloo, M. 2012. A wireless magnetoresistive sensing system for an intraoral tongue-computer interface, *IEEE Transactions on Biomedical Circuits and Systems*, 6(6):571-585.

Potamianos, G., Neti, C., Cravier, G., Garg, A. and Senior, A. W. 2003. Recent advances in the automatic recognition of audio-visual speech, *Proc. of IEEE*, 91(9):1306-1326.

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(4): 1190-1207.

Toda, T., Nakagiri, M., Shikano, K. 2012. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement, *IEEE Transactions on Audio, Speech and Language Processing*, 20(9): 2505-2517.

Uraga, E. and Hain, T. 2006. Automatic speech recognition experiments with articulatory data, *Proc. Inerspeech*, 353-356.

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

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

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

Wang, J., Samal, A., Green, J. R., and Rudzicz, F. 2012b. Whole-word recognition from articulatory movements for silent speech interfaces, *Proc. Interspeech*, 1327-30.

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

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

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

Wang J. 2014. DJ and his friend: A demo of conversation using a real-time silent speech interface based on electromagnetic articulograph. [Video]. Available: http://www.utdallas.edu/~wangjun/ssi-demo.html

Weismer, G. and Bunton, K. (1999). Influences of pellet markers on speech production behavior: Acoustical and perceptual measures, *Journal of the Acoustical Society of America*, 105: 2882-2891.

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