Word Recognition from Continuous Articulatory Movement Time-Series Data using Symbolic Representations

Jun Wang [1], *Arvind Balasubramanian* [2], *Luis Mojica de la Vega* [2], *Jordan R. Green* [3]
Ashok Samal [4], *Balakrishnan Prabhakaran* [2]

[1] Callier Center for Communication Disorders
[2] Department of Computer Science
University of Texas at Dallas, Dallas, Texas, United States
[3] MGH Institute of Health Professions, Boston, Massachusetts, United States
[4] Department of Computer Science & Engineering
University of Nebraska-Lincoln, Lincoln, Nebraska, United States
`{wangjun, arvind, luis.mojica, prabha}@utdallas.edu`
`jgreen2@mghihp.edu, samal@cse.unl.edu`

Abstract

Although still in experimental stage, articulation-based silent speech interfaces may have significant potential for facilitating oral communication in persons with voice and speech problems. An articulation-based silent speech interface converts articulatory movement information to audible words. The complexity of speech production mechanism (e.g., co-articulation) makes the conversion a formidable problem. In this paper, we reported a novel, real-time algorithm for recognizing words from continuous articulatory movements. This approach differed from prior work in that (1) it focused on word-level, rather than phoneme-level; (2) online segmentation and recognition were conducted at the same time; and (3) a symbolic representation (SAX) was used for data reduction in the original articulatory movement time-series. A data set of 5,900 isolated word samples of tongue and lip movements was collected using electromagnetic articulograph from eleven English speakers. The average speaker-dependent recognition accuracy was up to 80.00%, with an average latency of 302 miliseconds for each word prediction. The results demonstrated the effectiveness of our approach and its potential for building a real-time articulation-based silent speech interface for clinical applications. The across-speaker variation of the recognition accuracy was discussed.

Index Terms: silent speech recognition, laryngectomy, support vector machine, SAX, time-series

1. Introduction

Persons who lose their voice after laryngectomy (a surgical removal of the larynx due to the treatment of cancer) or who have speech impairment struggle with daily communication [1]. In 2012, more than 52,000 new cases of head and neck cancers (including larynx, pharynx, etc.) were estimated in the United States [2]. Currently, there are only limited treatment options for these individuals, which include (1) "esophageal speech", which involves oscillation of the esophagus and can be difficult to learn; (2) electrolarynx, which is a mechanical device resulting in a robotic-like voice; and (3) augmented and alternative communication (AAC) devices (e.g., text-to-speech synthesizers operated with keyboards), which are limited by slow manual text input [1]. New assistive technologies are needed to provide a more efficient oral communication mode with natural voice for those individuals.

Silent speech interfaces (SSIs), although still in early development stages [3] (e.g., speaker-dependent recognition, small-vocabulary, devices are not ready for clinical use), may provide an alternative interaction modality for persons with voice and speech problems. The common purpose of SSIs is to convert non-audio articulatory data to text that drives a text-to-speech (TTS) synthesizer (e.g., [4]) (see Figure 1 for a schematic of our SSI design). Potential articulatory data transduction methods for SSIs include ultrasound [5, 6], surface electromyography electrodes [7, 8], and electromagnetic articulograph (EMA) [9, 10, 11]. The current project used EMA, which registers the 3D motion of sensors adhered to the tongue and lips.

One major challenge for building effective SSIs is developing accurate and fast algorithms that recognize words or sentences based on articulatory data (i.e., without audio information). Articulatory data have been successfully used to improve the accuracy of voiced speech recognition from both healthy talkers [12, 13] and neurologically impaired individuals [14]. This typically involves the use of *articulatory features* (AFs), which include lip rounding, tongue tip position, and manner of production, for example. Phoneme-level AF-based approaches have typically obtained word recognition accuracies less than 50% [13] because articulation can vary significantly within those categorical features depending on the surrounding sounds and the speaking context [15].

These challenges in phoneme-level recognition motivate a higher unit level of articulatory recognition, for example, word-level or sentence-level. Although sentence-level recognition accuracy is high [9], it lacks the scalability of phoneme- and word-level recognition because all sentences are required to be known prior to prediction. Word-level recognition may have better scalability than sentence-level recognition and the potential for higher accuracy than phoneme-level recognition. Word-level recognition from acoustic data has outperformed monophone recognition by approximately 25% [16, 17]. However, whole-word recognition has rarely been investigated in articulatory data probably due to logistic difficulty of collecting articulatory data [10, 11].

Online word recognition from continuous articulatory movements can be extremely challenging because word

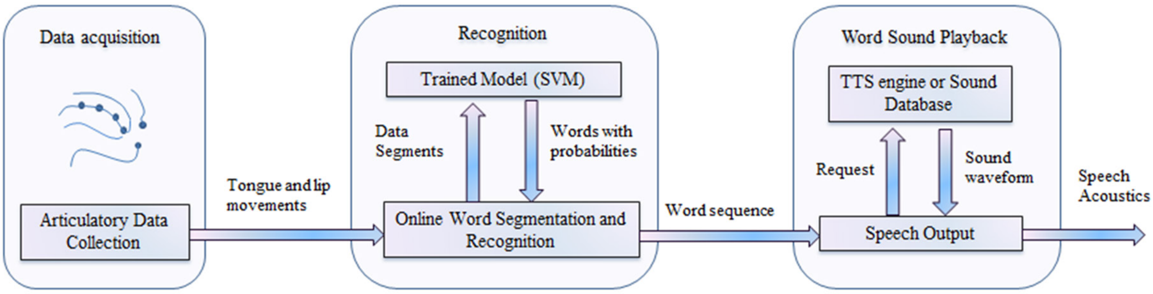

Figure 1. *Three-component design of the articulatory movement-based silent speech interface.*

boundaries (onset and offset) are difficult to identify. Recent works have shown offline word classification (word boundaries are known) accuracy can be greater than 90% for a small vocabulary [10, 11]. However, because of word segmentation issues, online recognition accuracy can be significantly lower than offline classification accuracy. Online word segmentation based on articulatory movements has rarely been attempted [18]. A threshold (e.g., 2 SD) of the articulatory movements has been successfully used for isolated word datasets [19, 20]. Such amplitude-based segmentation may not be well suited for words produced in a continuous sequence because of co-articulation (illustrated in Figure 1) or for words within sentences (connected speech). Co-articulation is an effect characterized by a sound is affected by its adjacent sounds [21, 22].

Figure 2 illustrates the articulatory movements for a word sequence with co-articulation produced by one of the participants. The top panel shows the continuous motion of sensors (*y* and *z* coordinates, where *y* is vertical and *z* is front-back) attached on the tongue and lips. T1, T2, T3, and T4 are four sensors attached on the midsaggital line of the tongue, from tip to back; UL is upper lip; LL is lower lip. Details of the coordinate system and the labels of the sensors are provided in Section 4. The bottom panel shows the synchronously recorded audio.

The goal of this project was to investigate word recognition from continuous articulatory movements. A novel, real-time algorithm for word recognition from continuous stream of articulatory movements has been recently proposed [10]. The algorithm was designed to solve the online segmentation and recognition problems simultaneously. The algorithm is characterized by the following: recognition is at the word level rather than the phoneme- or sentence-level; recognition employs a dynamic thresholding technique based on patterns in the probability change returned by a classifier; and the algorithm is extensible (i.e., it can be embedded with a variety of classifiers). The algorithm has been tested on the minimally processed articulatory movements [10]. Although the results were promising (missing only 1.93 words on a sequence with twenty-five words), false positives caused a relatively low overall accuracy.

The current project implemented the following three strategies for improving word recognition accuracy: (1) using symbolic aggregation approximation (SAX) representation to reduce the local variation in the original articulatory movement time-series data, (2) adding a look-back strategy to handle a situation in which two words are so close that the onset of the second word may not be accurately identified, and (3) using speaker-dependent thresholds to determine the word candidates during online recognition. A phonetically-balanced and isolated word dataset of tongue and lip movements was collected using electromagnetic articulograph and used to evaluate the effectiveness and efficiency of the improved algorithm.

2. Design & Method

The design of our articulation-based silent speech interface is illustrated in Figure 1, which contains three major components [9, 10]: (a) data acquisition, (b) online (word) recognition, and (c) sound playback or synthesis. Data acquisition is performed using an electromagnetic articulograph that tracks the motion of sensors attached on a speaker's tongue and lips.

The focus of this paper is the second component, online word recognition, whose goal is to recognize a set of isolated words from continuous articulatory data (without using audio data). The core recognition problems are to (1) convert a time-series of spatial configurations of multiple articulators to time-delimited words, and (2) identify the onset of those recognized words. Here, a spatial configuration is an ordered set of 3D locations of the sensors. In this whole-word recognition algorithm, segmentation and identification are conducted together in a variable-size sliding window. The algorithm is based on the premise that a word has its highest matching probability given an observation window with an appropriate starting point and width. A trained machine learning classifier that derives these matching probabilities is embedded into the algorithm, as described in the rest of this section. In the future, this algorithm will serve as the recognition component of our articulation-based SSI.

2.1. Symbolic representation of articulatory time-series data

SAX is a symbolic representation technique [23] that has been widely used in time-series data pattern analysis (e.g., [24, 25, 26]). The main idea of SAX is to represent the original time-series amplitude using discrete symbols that can still capture the patterns. The potential benefits of SAX are (i) efficient dimensionality reduction while retaining essential features; and (ii) lower bounding of the distance measure on the original series. To our best knowledge, however, SAX has not been used for articulatory movement time-series data analysis.

The underlying contention behind representing the tongue motion data in the form of symbols is to capture the motion pattern for a particular word and to reduce the local variation. If the motion trajectory can be captured in terms of symbols that represent different regions in the motion distribution space, then the symbolic representation should reduce the amount of data required while overcoming local variations and scaling effects, thus may enable efficient comparison of the

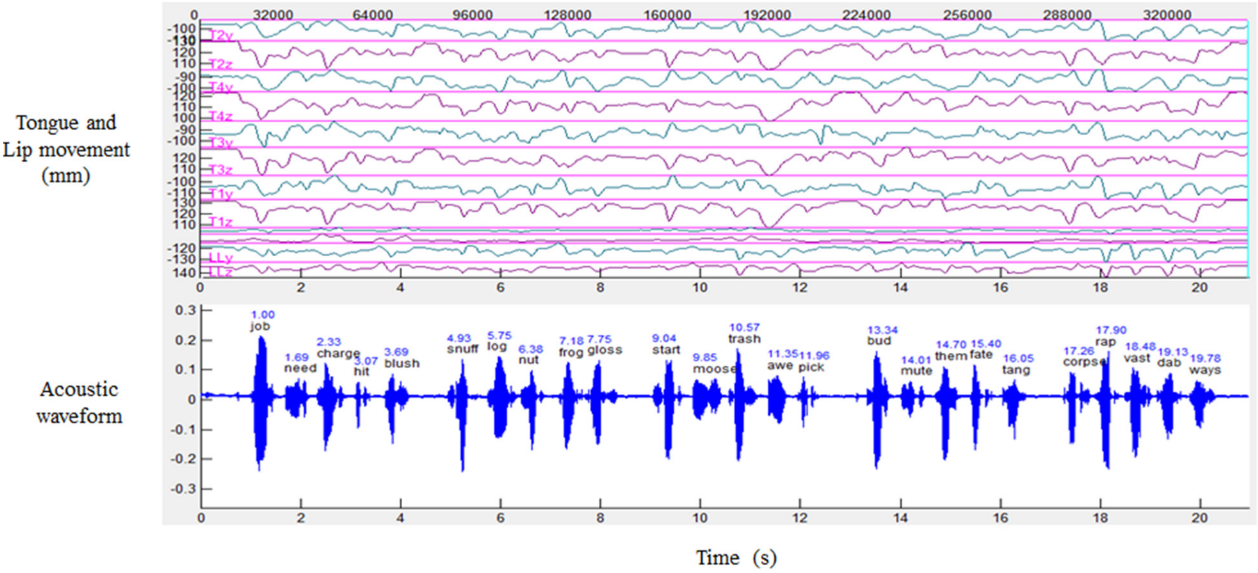

Figure 2. *Example of a sequence of tongue and lip movements (top panel) of twenty five words and synchronously recorded sounds (bottom panel). Labels of the tongue and lip sensors are described in text. The articulatory movement data was low-pass filtered (20 Hz). In the acoustic waveform panel, the numbers in blue above words are the actually occurrence time of that word.*

motion data of different words with a higher accuracy.

In this study, SAX symbolic representation was used to discretize the tongue and lip motion time series data. In SAX, each time sequence is z-normalized (mean = 0 and SD = 1), and split into *w* equal segments. For each segment, the mean is calculated and a symbol is assigned based on a set of breakpoints that divide the distribution space into α equiprobable regions, where α is the alphabet size. When α is given, the breakpoints (that separate the space to α regions) are definite. For the definition of breakpoints, please refer to [23]. Thus, each time subsequence is converted into a string of length *w*, formed by symbols from an alphabet of size α. Both the length *w* and the alphabet size α are pre-specified. Theoretically, an optimal combination of the two parameters – *w* and α – should be able to efficiently represent the variation in the sequences of any given time series data. Figure 3 illustrates how a time-series is converted to string of symbols (using *w* = 5, and α = 6).

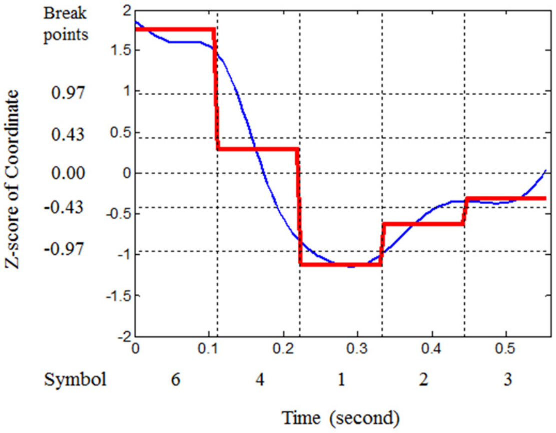

Figure 3. *Example of a symbolic representation of articulatory movement time-series data using SAX; the blue curve is the z-scored vertical coordinate of tongue tip producing a word "job"; the red segments are the discretized results. The original articulatory time-series data are finally converted into a string of symbols "64123".*

In this project, however, a word sample contains multiple time sequences, multi-dimensional coordinates (*y* and *z*) from multiple sensors. The following procedure was used to convert a data sample of original articulatory movement data to a string of symbols. The original data captured from all sensors was first time-normalized and amplitude shifted to have a mean of zero. These data arrays were then combined into a single-dimension data vector (with sequences of multi-dimension data from multiple sensors). The data vector was then converted into a single SAX vector. The reason for using concatenation of all sensor data (rather than converting on each sensor separately) to generate a single SAX vector is to preserve the relative variation in amplitude across sensors. Conversion to SAX reduced the data by a constant factor (number of data points for each sensor / *w*). The SAX vectors were served as input to the training and testing phases of the recognition module.

The optimal SAX parameters (*w* and α) needed to be determined before word recognition experiment could be conducted. Most of the words in our dataset were of the phonetic structure CVC (consonant-vowel-consonant) or CCVCC, thus, *w* = 5 was chosen as the length of symbol string for capturing the motion characteristics. A preliminary experiment was conducted to determine the best α value. Figure 4 gives the average word off-line classification accuracy across speakers for different α values (from 3 to 15), and *w* = 5. α = 6 resulted in the highest classification accuracy, and was thus used in the online recognition experiment, which will be described in the next two sub-sections.

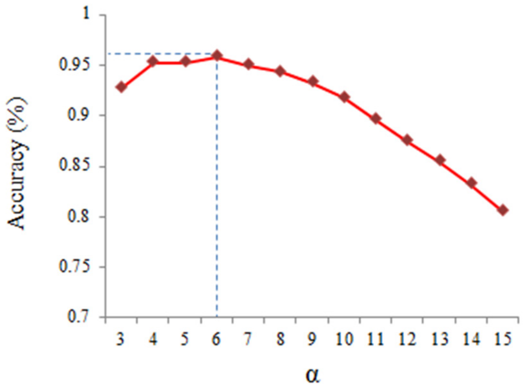

Figure 4. *Average offline classification accuracy across speakers using different* α *values.*

2.2. Model training

Support vector machine (SVM) [27], a widely used machine learning classifier, was used to recognize words in this project. SVMs are soft margin classifiers that find separating hyperplanes with maximal margins between classes in high dimensional space [28]. Model training was conducted by training a SVM using pre-segmented multi-dimension articulatory movement data from multiple sensors associated with known words. A kernel function is used to describe the distance between two data points (i.e., *u* and *v* in Equation 1). A radial basis function (RBF) was used as the kernel function in this experiment, where λ is an empirical parameter:

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

Details of the implementation of SVM used in this experiment were described in [28].

The training component was developed off-line before the SSI was deployed in a real-time application. Therefore, the time required to build the model is not a relevant problem. Rather, the time taken for a trained model to predict words is an important measure for evaluating real-time applications. To obtain a high speed in prediction, input data was minimally processed and converted to SAX symbols before being fed into the SVM. The sampled motion paths of all articulator were time-normalized to a fixed-width (SVMs require samples to have a fixed number of values) and concatenated as one vector of attributes. The vector was then converted to SAX symbols, which formed a word sample. To understand the improvement of using SAX itself, we compared the offline classification accuracy using SAX and using the minimally processed original time-series data (used in [10]).

2.3. Online recognition

A prediction window with variable boundaries was used to traverse the sequence of tongue and lip movement data to recognize words and their locations (onset) within the window based on the probabilities returned by LIBSVM, which extends the generic SVM by providing probability estimates transformed from SVM decision values [28]. The SVM was trained offline using pre-segmented articulatory movement data. Pseudo-code of the original whole-unit recognition algorithm is provided in [9].

The major steps of the *improved* word recognition algorithm are described as below. Steps 1 to 3 are for finding word candidates; Steps 4 to 6 are to verify those candidates; Step 7 is sound playback of recognized words.

In Step 1 to 3 (Figure 5), word candidates are identified within the prediction window based on the probabilities returned from the trained SVM. At each time point *t*, all possible word lengths (within the length range of training words with a step size Δ*t*) are considered and the maximum probability is returned as the probability for time point *t*. The word length in our list ranges from 370 to 885 ms. The offset of the probability function varied considerably across words, which made it difficult to identify a sensitive candidate threshold. Therefore, the probability associated with each word was baseline-corrected by subtracting the average probability derived from the first 600 ms of the test sequence. Candidates are identified in a prediction window (represented by its left and right boundaries, w_l and w_r) when probability values exceed a candidate threshold ($thres_c$). The candidate threshold was obtained empirically from training data. In the current experiment, a single constant threshold was used for all words (but varies for different subjects). In the future, each word will have its own threshold for each subject. In this speaking-dependent recognition experiment, the threshold varied slightly for different subjects (ranged from 0.30 to 0.40).

If no candidates are found in the current prediction window, w_r moves forward (to get more data), and the process goes back to step 1, until $w_r \leq w_l + l_{max}$, where l_{max} is the maximum word length in this data set.

In Step 4, a candidate is verified based on probability change trend. If the probabilities for that word are decreasing in a time span of half of the minimum word length, implying ongoing decreases, the candidate is confirmed; otherwise, the decision-making is delayed. This strategy is to confirm a word right after the peak probability of the word happens, while the peak probability is unknown in online recognition.

Look-back strategy. When the currently recognized word is very close to the next word, the location of w_l may be erroneously located after the actual beginning of the next

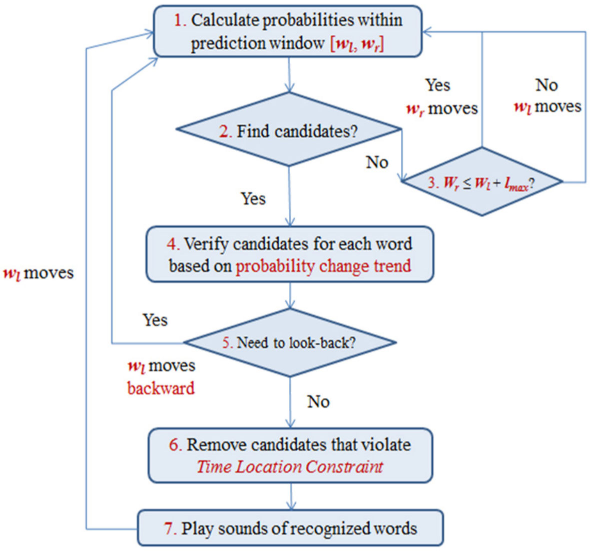
1. Calculate probabilities within prediction window [w_l, w_r]
2. Find candidates?
Yes
w_r moves
No
w_l moves
No
3. $W_r \leq W_l + l_{max}$?
Yes
4. Verify candidates for each word based on probability change trend
w_l moves
Yes
5. Need to look-back?
w_l moves backward
No
6. Remove candidates that violate *Time Location Constraint*
7. Play sounds of recognized words

Figure 5. *Schematic of the improved word recognition algorithm from continuous articulatory movement data.*

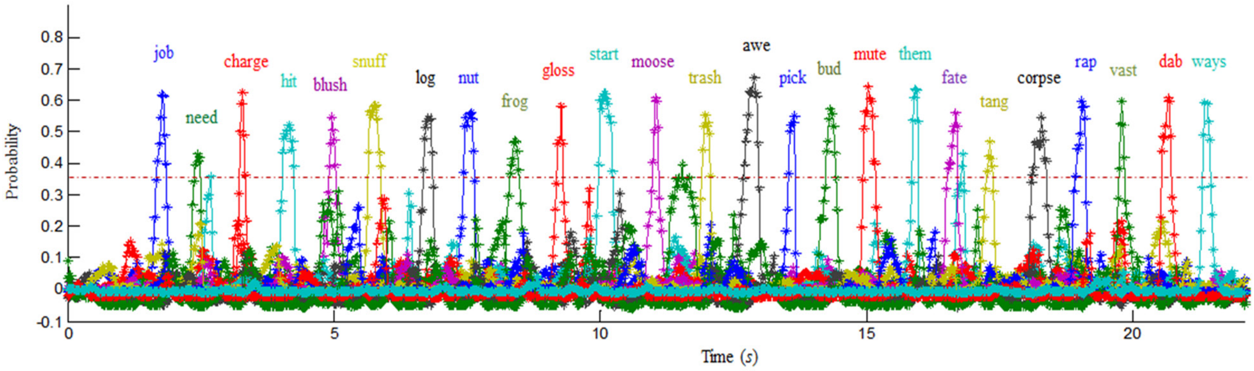

Figure 6. *Example of probabilities (baseline removed) of twenty-five words on a test sequence. The dashed horizontal line is the probability threshold for word candidates.*

word. This situation may cause error predictions, which was not considered in [10]. A look-back strategy was introduced to address this problem in this experiment (Step 5). A threshold $thres_{look\text{-}back}$ (> candidate threshold $thres_c$) is defined first. When a word candidate was found at time t_c with probability p_c, if $p_c \leq thres_{look\text{-}back}$, the window location before t_c was saved as the candidate predicted time location, which means $w_l = w_l - \Delta t$. In the current setting, w_l takes at most one step back because two or more step size back is unlikely to happen in real articulatory movement data (otherwise the two words may have overlap). Also to avoid dead loop of the execution, this procedure executes at most once in the implementation of the algorithm.

Time Location Constraint allows only one word to occur within each time span (Step 6). A time span must not be less than the minimum word length in the training data (i.e., 370 ms). If more than one word candidate is found within a time span, only the one with the highest probability is retained in the recognized word list.

In Step 7, after playing prerecorded audio samples of recognized words, the left boundary of the prediction window (w_l) moves to w_r. The whole procedure (Step 1 to 7) is repeated until the rightmost boundary of the prediction window (w_r) reaches the end of the input sequence.

2.4. Evaluation

Recognition accuracy and processing time were used to evaluate the performance of the word recognition algorithm.

A word prediction is correct if the expected word is identified within half a second of its actual occurrence time. That is, both missing values and wrongly predicted occurrence times are considered as errors. A false positive is a word that is recognized at a time point where there is actually no word. Figure 6 illustrates the word probability distribution on a selected sequence. In this example, all twenty-five words were correctly recognized.

Two measures were used to evaluate the efficiency of this algorithm: *prediction location offset* (machine-independent) and *prediction processing time*, or *latency* (machine-dependent). Prediction location offset was defined as the difference in location on a sequence between where a word is actually spoken and where it is recognized [29]. The prediction location offset provides an estimate of how much information is needed for predicting a word. Latency is the actual CPU time needed for predicting a word.

3. Data Collection

3.1. Participants and stimuli

Eleven healthy native English speakers participated in data collection. Each speaker participated in one session in which he/she repeated a sequence of twenty-five words (i.e., one of the four phonetically-balanced word lists in [30]) multiple times.

Subjects, who were blinded to the specific purpose of the research, were asked to pronounce the target words in their habitual speaking rate and loudness. Thus, the production contained co-articulation between adjacent words, although the co-articulation might not be similar to that in connected speech.

3.2. Tongue motion tracking devices

The electromagnetic articulograph (EMA) AG500 (Carstens Medizinelektronik GmbH, Bovenden, Germany) was used to collect the 3-D movement time-series data of the tongue, lips, and jaw for ten of the eleven participants. Wave Speech Research System (Northern Digital Inc., Waterloo, Canada) was used for the other participant. The two devices are based on the same electromagnetic tracking technologies [31, 32]. Both devices record 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, and they have similar data collection procedure [33]. Thus, only the data collection procedure using EMA will be described in this paper (in Section 3.3). The spatial precision of motion tracking using EMA (AG500) and Wave are both approximately 0.5 mm [34, 35]. The sampling rate of the original data is 200 Hz for EMA AG500 and 100 Hz for Wave, respectively.

3.3. Procedure

Participants were seated with their head within the calibrated magnetic field. Then sensors (pellets) were attached to the surface of each articulator using dental glue (PeriAcryl Oral Tissue Adhesive). The participants were then asked to produce the word sequences at their habitually comfortable speaking rate and loudness. Before the beginning of actually data recording, a two-minute training and practice helped the participants to adapt to the wired sensors. Previous studies have shown these sensors do not significantly affect their

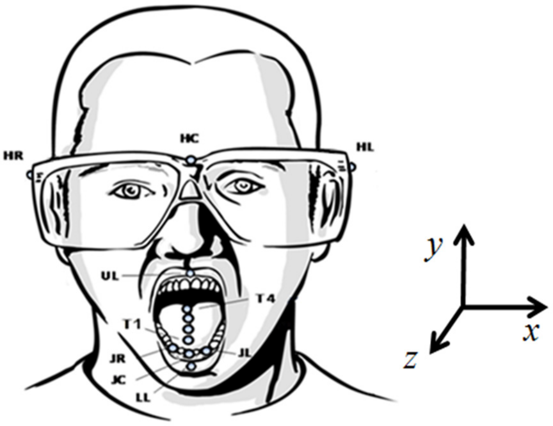

Figure 7. *Positions of sensors attached on the subject's head, tongue, lips, and jaw in data collection.*

speech output [36].

Figure 7 (picture adapted from [37]) shows the positions of 12 sensors attached to a participant's head, face, and tongue [38, 39]. Three of the sensors were attached to a pair of glasses. 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 sensors were used to calculate the movements of other articulators independent of the head. Four sensors - T1 (Tongue Tip), T2 (Tongue Blade), T3 (Tongue Body Front) and T4 (Tongue Body Back) - were attached approximately 10 *mm* from each other at the midline of the tongue [38, 39, 40]. Lip movements were captured by attaching two sensors to the vermilion borders of the upper (UL) and lower (LL) lips at midline.

Data from the four tongue sensors and the two lip sensors were used for this word recognition experiment. The movements of three jaw sensors, JL (Jaw Left), JR (Jaw Right), and JC (Jaw Center), were recorded for future use.

3.4. Data preprocessing

The time-series data of sensor locations recorded using EMA went through a sequence of preprocessing steps prior to analysis. First, the head movements and orientations were subtracted from the tongue and lip locations to give head-independent measurements of the analysis variables. The orientation of the derived 3-D Cartesian coordinate system is displayed in Figure 7. Second, a zero phase lag low pass filter (i.e., 20 Hz) [10, 40] was applied for removing noise. Third, all sequences were manually segmented based on synchronously recorded audio data and annotated with words using a Matlab-based software called SMASH [33].

Only y (vertical) and z (front-back) coordinates (see Figure 7) of the six tongue and lip sensors (i.e., T1, T2, T3, T4, UL, LL) were used for this word recognition experiment because the movement along the x axis (left-right) is not significant in normal speech production [38, 41]. In the future, however, x dimension will be used for predicting speech articulated by individuals with laryngectomy or other speech disorders. The center of the magnetic field is the origin (zero point) of the EMA coordinate system.

Error samples (e.g., mispronunciation or sensor falling off during the production) were rare and were excluded from the experiment. In all, 5,900 word samples (in 236 sequences) were obtained and used in this experiment.

4. Results & Discussion

Cross validation is a standard procedure to evaluate the performance of classification algorithms, where training data and test data are separate. Leave-one-out cross validation was conducted on the dataset from each subject in both training and online recognition, where one sequence (with twenty-five words) was used for testing and the rest of the sequences were used for training.

4.1. Training accuracy

The average training (offline classification) accuracy was 94.01% using minimally processed articulatory data (used in [10]) and 96.90% using SAX transformed data in the current experiment. A paired *t*-test showed that the 2.89% improvement in accuracy was statistically significant ($p < 0.001$).

The experimental results demonstrated that SAX is effective in retaining the articulatory movement patterns while reducing the local variation. SAX may have potential for a greater improvement in classification accuracy for a larger vocabulary.

4.2. Online recognition accuracy and processing time

The average online recognition accuracy across all subjects was 80.00% (SD = 10.95%). More specifically, our algorithm failed to recognize 1.96 words (SD = 0.88) and generated 3.04 (SD = 1.95) false positives in a sequence of twenty-five words. The average difference of correctly predicted word locations and their actual locations was 48 ms (SD = 9). The online word accuracy was improved up to 20%, compared with the performance of the original algorithm [10].

The average prediction location offset and latency were 150 ms (SD = 68) and 302 ms (SD = 11) for a word prediction, respectively. Latency was measured on a PC with 2.6 GHz dual-core CPU and 4GB memory.

Table 1 summarizes the performance findings of the original and current algorithm [10]. During offline classification, the only difference between the original

Table 1. *Summary of the performances of current and the original algorithm.*

Measure	The Original Algorithm	The Current Algorithm	Statistical Significance
Offline Classification Accuracy	94.01%	96.90%	$p < 0.001$
Online Missing Words	1.93	1.96	
Online False Positives	8.08	3.04	$p < 0.001$
Online Recognition Accuracy	60.00%	80.00%	$p < 0.001$

algorithm and the current algorithm was the use of SAX and only a modest improvement in recognition was achieved. For online recognition, the current algorithm implemented not only SAX, but also a look-back strategy, and speaker-dependent thresholds. This implementation improved overall accuracy by primarily reducing the number of false positives. Additional work, however, is needed to determine the individual benefit of each newly-added component (i.e., SAX, look-back strategy, and speaker-dependent thresholds).

The high accuracy showed the effectiveness of our proposed algorithm to address the challenge in word recognition caused by co-articulation. The low prediction location offset and latency demonstrated the potential of our approach for real-time applications. The low standard deviations of the accuracy and other measures across subjects indicate that our approach can be applied generally with multiple subjects.

4.3. Across-talker accuracy variation

Although speech articulation is thought to vary across talkers [21], reports on this variability have been limited because most silent speech recognition or relevant studies have involved less than five participants.

As reported previously, the standard deviation of the online word recognition accuracy across eleven subjects was 10.95%, which is not surprising. To examine across taker differences in our study, the eleven subjects were grouped into four groups according to their word recognition accuracy, < 70%, 70-80%, 80-90%, and ≥ 90%. Figure 8 shows the distribution of the subjects with regard to the word recognition accuracy. 18.18% of the subjects obtained an accuracy equivalent or greater than 90%; 36.36% obtained an accuracy greater than 80% but less than 90%; 27.27% obtained an accuracy between 70% and 80%; 18.18% obtained an accuracy less than 70%. In other words, 81.82% of the subjects obtained accuracy greater than 70%. It is notable that two of the participants had significantly lower recognition accuracies than the other nine participants, while the two participants had similarly high offline classification accuracies. Future work is required to determine the factors that account for across participant differences in recognition accuracy.

4.4. Adaptability for real online recognition

Our word recognition algorithm was designed for online recognition. In this experiment, the algorithm was tested using pre-recorded sequences of continuous articulatory movement data. That is, the algorithm was not tested in a real online recognition experimental setup. However, our experiment, to some extent, simulated online recognition. During the recognition, at time *t*, only data before ($t + l_{max}$) can be reached (l_{max} = 885 ms), which can be considered as an approximation of a real online recognition setting. Therefore, the word recognition algorithm used in this study should be well suited for real-time applications. Testing the algorithm in a real online recognition experimental setting is a next step.

4.5. Limitations

Although the results are very promising, there are a number of limitations of the current algorithm. First, quite a few parameters (e.g., candidate threshold, threshold for look-back, step size of the sliding window) need to be determined before online prediction, although they can be manually adjusted at the beginning (for example, candidate threshold). An automatic approach for determining the optimal parameters is needed before the silent speech recognition algorithm can be used in practice.

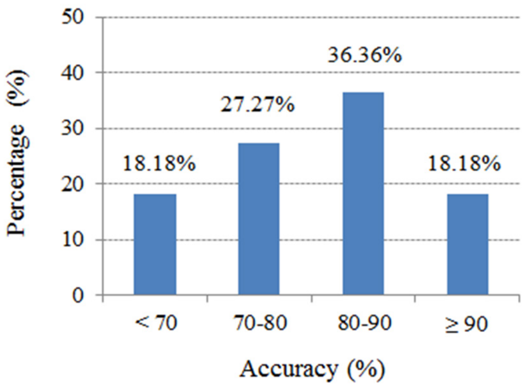

Figure 8. *Distribution of talkers regarding to online recognition accuracy.*

Although the EMA and Wave are able to register 3D tongue motion accurately in real-time, and Wave is lightweight enough to be installed on a wheelchair, they may be still cumbersome in clinical use. An ideal or practical silent speech interface could be a handheld or a wearable device. Fortunately, the electromagnetic motion tracking technology is advancing rapidly. For example, devices that are wearable, and even with wireless sensors are being investigated (e.g., [11, 42, 43]). Our algorithm that uses the sensor coordinates will be seamlessly embedded with those portable systems when they are ready for clinical use.

5. Conclusions & Future Work

Experimental results showed the potential of our word recognition algorithm for building an articulation-based silent speech interface, which can be used in command-and-control systems using silent speech and may even enable voiceless patients to produce synthetic speech using their tongue and lips.

Although the current results are encouraging, future work is required to determine the optimal parameters (e.g., candidate thresholds) automatically for online recognition. In addition, the efficacy of alternative classifiers should be explored such as Hidden Markov Models [44, 45, 46], Fast DTW [47], Dynamic Bayesian Network [48], Random Forest [14]; the current design is easily adapted to classifiers that generate estimated probabilities associated with candidates.

6. Acknowledgments

This work was in part funded by Excellence in Education Fund, University of Texas at Dallas, Barkley Trust, University of Nebraska-Lincoln, and a grant awarded by the National Institutes of Health (R01 DC009890/DC/NIDCD NIH HHS/United States). We would like to thank Dr. Tom Carrell, Dr. Lori Synhorst, Dr. Mili Kuruvilla, Cynthia Didion, Rebecca Hoesing, Kate Lippincott, Kayanne Hamling, Kelly Veys, Toni Hoffer, and Taylor Boney for their contribution to subject recruitment, data collection, data management, and data 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] American Cancer Society, "Cancer Facts and Figures 2012", Atlanta, GA*: American Cancer Society*. Retrieved on December 26, 2012.

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

[4] Sproat, R. (Ed.), "Multilingual text-to-speech synthesis : The Bell Labs approach", in *Computational Linguistics* (1st ed.), vol. 24, p. 328. 1998: Springer.

[5] 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.

[6] 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.

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

[8] Heaton, J. T., Robertson, M., and Griffin, C., "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, 2011.

[9] 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.

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

[11] 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.

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

[13] Livescu, K., Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., Kantor, Lal, P., Yung, L., Bezman, A., Dawson-Haggerty, S., Woods, B., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop", *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, 621-624, 2007.

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

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

[16] Sharma H. V., Hasegawa-Johnson, M., Gunderson, J., and Perlman A., "Universal access: Speech recognition for talkers with spastic dysarthria", *Proc. Interspeech*, 1451-1454, 2009.

[17] Kantor, A., "Pronunciation modeling for large vocabulary speech recognition", PhD Dissertation, Dept. Comput. Sci., University of Illinois, Urbana, 2011.

[18] Akdemir, E., and Ciloglu, T., "The use of articulator motion information in automatic speech segmentation", *Speech Communication*, 50(7):594-604, 2008.

[19] Gilbert, J. M., Rybchenko, S. I., Hofe, R., Ell, S. R., Fagan, M. J., Moore, R.K., and Green, P., "Isolated word recognition of silent speech using magnetic implants and sensors", *Medical Engineering & Physics*, 32(10):1189-1197, 2011.

[20] Green, J.R., Beukelman, D.R., and Ball, L. J., "Algorithmic estimation of pauses in extended speech samples", *Journal of Medical Speech-Language Pathology*, 12, 149-154, 2004.

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

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

[23] Lin, J., Keogh, E., Lonardi, S., and Chiu, B. "A symbolic representation of time series, with implications for streaming algorithms", *Proc. of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. San Diego, CA, 2003.

[24] Mueen, A., Keogh, E., "Online discovery and maintenance of time series motifs", *Proc. 16th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, Washington, DC, 1089-98, 2010.

[25] Wei, L., Kumar, N., Lolla, V. N., Keogh, E., Lonardi, S., and Ratanamahatana, C. A., "Assumption-free anomaly detection in time series", *Proc. 17th International Scientific and Statistical Database Management Conference*, Santa Barbara, CA, 237-240, 2005.

[26] Lin, J., Keogh, E., and Lonard, S. "Visualizing and discovering non-trivial patterns in large time series databases", *Information Visualization*, 4(2):61-82, 2005.

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

[28] 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.

[29] Wang, J., "Silent speech recognition from articulatory motion", Ph.D. dissertation, Dept. Comput. Sci., Univ. of Nebraska-Lincoln, 2011.

[30] Shutts, R. E., Burke, K. S., and Creston, J. E., "Derivation of twenty-five-word PB Lists", *Journal of Speech Hearing Disorders*, 29:442-447, 1964.

[31] Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., and Jackson, M. T. T., "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements", *Journal of Acoustical Society of America*, 92(6):3078–3096, 1992.

[32] Hoole, P., and Zierdt, A., "Five-dimensional articulography", in *Speech Motor Control: New Developments in Basic and Applied Research*, B. Maassen and P. van Lieshout, Eds. Oxford University Press, ch. 20, pp. 331–349, 2010.

[33] Green, J. R., Wang, J., and Wilson, D. L., "SMASH: A tool for articulatory data processing and analysis", *Proc. Interspeech*, 2013 (In press).

[34] 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.

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

[36] 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.

[37] 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.

[38] 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*, 2013 (In press).

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

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

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

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

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

[44] Cai, J., Denby, B., Roussel, P., Dreyfus, G., and Crevier-Buchman, L., "Recognition and real time performances of a lightweight ultrasound based silent speech interface employing a language model", *Proc. Interspeech*, Florence, Italy, 1005-08, 2011.

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

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

[47] Salvador, S., and Chan, P., “Toward accurate dynamic time warping in linear time and space”, *Intelligent Data Analysis*, 11(5):561-580, 2007.

[48] Frankel, J., Wester, M., and King, S., “Articulatory feature recognition using dynamic bayesian networks”, *Computer Speech Language*, 21(4):620-640, 2006.