Whole-Word Recognition from Articulatory Movements for Silent Speech Interfaces

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

[1]Department of Special Education & Communication Disorders
[2]Department of Computer Science & Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, United States
[3]Munroe-Meyer Institute, University of Nebraska Medical Center, Omaha, Nebraska, United States
[4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

{junwang,samal}@cse.unl.edu, jgreen4@unl.edu, frank@cs.toronto.edu

Abstract

Articulation-based silent speech interfaces convert silently produced speech movements into audible words. These systems are still in their experimental stages, but have significant potential for facilitating oral communication in persons with laryngectomy or speech impairments. In this paper, we report the result of a novel, real-time algorithm that recognizes whole-words based on articulatory movements. This approach differs from prior work that has focused primarily on phoneme-level recognition based on articulatory features. On average, our algorithm missed 1.93 words in a sequence of twenty-five words with an average latency of 0.79 seconds for each word prediction using a data set of 5,500 isolated word samples collected from ten speakers. The results demonstrate the effectiveness of our approach and its potential for building a real-time articulation-based silent speech interface for health applications.

Index Terms: silent speech recognition, speech impairment, laryngectomy, support vector machine

1. Introduction

Persons who lose their voice due to the surgical removal of the larynx (i.e., laryngectomy) and who have moderate speech impairment struggle with daily communication [1]. Each year, about 15,000 new cases of laryngeal cancer and hyperlaryngeal cancer are diagnosed in the United States [2] and there are an estimated 16,500 tracheo-esophageal speech valve changes every year in the UK [3]. Currently, there are only limited treatment options for these individuals, which include (a) "esophageal speech", which involves oscillation of the esophagus and is difficult to learn; (b) electrolarynx, which is a mechanical device resulting in a robotic-like voice; and (c) 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 these individuals.

Articulation-based silent speech interfaces (SSIs), although still in early development stages [4], provide an alternative interaction modality for individuals with altered or missing larynxes. The common purpose of SSIs is to convert articulatory data (without using audio data) to text that drives a text-to-speech (TTS) synthesizer to output sounds. SSIs may produce a more natural voice (using TTS) than electrolarynx does and SSIs may be more efficient (higher speaking rate) than AAC devices [2, 3, 4].

Two major challenges for developing SSIs are designing (a) hardware devices for articulatory data collection that are suitable for clinical applications and (b) fast and accurate algorithms to convert articulatory observations to text. Different hardware technologies for SSI have been compared in [4], including the promising electromagnetic articulograph (EMA) [3] in which point sensors placed on the lips and tongue are tracked in three dimensions. Recent studies have shown the potential of EMA-based silent speech interfaces with permanently affixed sensors for command-and-control applications [3, 5]. Our study is focused on the development of a fast and accurate algorithm for word recognition from continuous articulatory movements.

Research has shown that articulatory data can improve the accuracy of word recognition from the voiced speech of both healthy [6, 7] and neurologically impaired individuals [8]. 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 often obtain word recognition accuracies less than 50% [6] because articulation can vary significantly within those categorical features depending on the surrounding sounds and the speaking context [9]. These challenges motivate a higher unit level of recognition.

Sentence-level recognition from articulatory data has recently been investigated, which is promising in terms of accuracy [2]. However, sentence-level recognition lacks the scalability of phoneme- and word-level recognition, because all sentences are required to be known prior to prediction. Word-level recognition trades off relatively good scalability and the potential of higher accuracy than phoneme-level recognition.

Word-level recognition from acoustic data has been investigated, which showed that word-level (and triphone-level) models outperform monophone models with approximately 25% in the relative reduction of error rate [10, 11]. However, whole-word recognition has rarely been investigated in articulatory data due to logistic difficulty to collect articulatory data [5].

This paper investigates whole-word recognition from articulatory data for silent speech interfaces by applying a newly developed whole-unit articulatory recognition algorithm [2]. The algorithm is adapted to this project and is characterized by the following features: (1) recognition is at the whole-word level rather than the phoneme- or sentence-level; (2) recognition is based on articulatory movements rather than on derived AFs; (3)

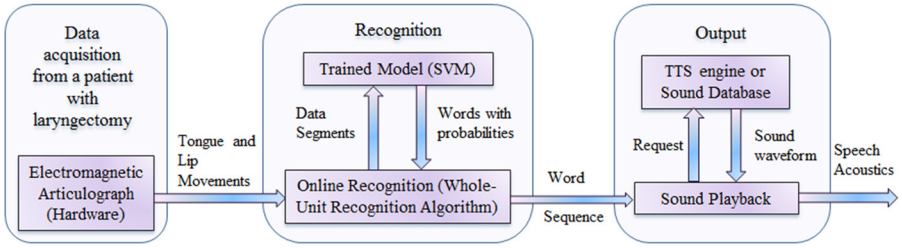

Figure 1. Design of the EMA-based silent speech interface

recognition is based on a dynamic thresholding technique based on patterns in the probability change returned by a classifier; and (4) the algorithm is extensible (i.e., it can be embedded with a variety of classifiers). In the future, this algorithm will serve as the recognition component of our articulation-based SSI. A phonetically-balanced and isolated word dataset was collected and used to evaluate the effectiveness and efficiency of this algorithm.

2. Design & Method

The design of our articulation-based SSI is illustrated in Figure 1, which contains three components: (a) data acquisition, (b) online (whole-word) recognition, and (c) output (playback or synthesis) [2]. Data acquisition is performed using an EMA machine that tracks the motion of sensors attached on a speaker's tongue and lips. The focus of this paper is the online recognition component whose goal is to recognize a set of phonetically-balanced and isolated words from articulatory data only. The core recognition problem is to convert a time-series of spatial configurations of multiple articulators to time-delimited words. Here, a spatial configuration is an ordered set of 3D locations of the sensors. In the whole-word recognition algorithm, segmentation and identification are conducted together in a variable-size moving 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 classifier that derives these matching probabilities is embedded into the algorithm, as described in the rest of this section.

2.1. Model training (classification)

A support vector machine (SVM) [12] was trained using pre-segmented articulatory movement data from multiple sensors associated with known words. SVMs are widely used soft margin classifiers that find separating hyperplanes with maximal margins between classes in high dimensional space [13]. A kernel function is used to describe the distance between two data points (i.e., x and y in Equation 1). A radial basis function (RBF) was used as the kernel function in this experiment, where λ is an empirical parameter:

$$K_{RBF}(x, y) = \exp(1 - \lambda \| x - y \|) \quad (1)$$

The training component was developed off-line before the SSI was deployed in a real-time application. Therefore, we do not consider the time required to build the model as 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, a direct mapping strategy was used, in which the input data was minimally processed before being fed into the SVM (directly mapped to words). 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, which formed a word sample. Furthermore, to determine the relative accuracy of our SVM classification to another commonly-used time-series classification approach, we also tested classification using dynamic time warping (DTW), which were used for the same application [3, 5].

2.2. Online recognition

The trained classifier was then used to recognize words from continuous (unsegmented) tongue and lip movement data. A prediction window with variable boundaries was used to traverse the sequence of tongue and lip movement data to recognize words and their locations within the window based on the probabilities returned by LIBSVM, which extends the generic SVM by providing probability estimates transformed from SVM decision values [12]. Pseudo-code of the whole-unit recognition algorithm is in [2].

In step 1 (Figure 2), word candidates are identified within the window according to probabilities returned from the trained SVM. All possible word lengths (within the length range of training words with a step size) are considered and the maximum probability is returned as the probability for a time point. The word length in our list ranges from 370 to 608 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 exceeded a candidate threshold obtained empirically from training data.

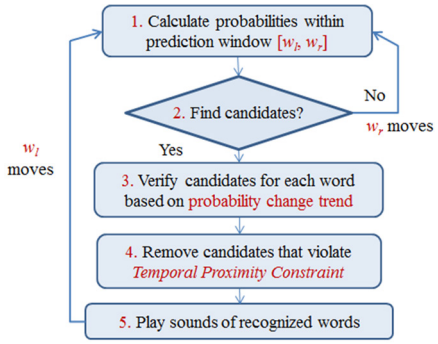

Figure 2. A schematic of the whole-word recognition algorithm from articulatory movement data.

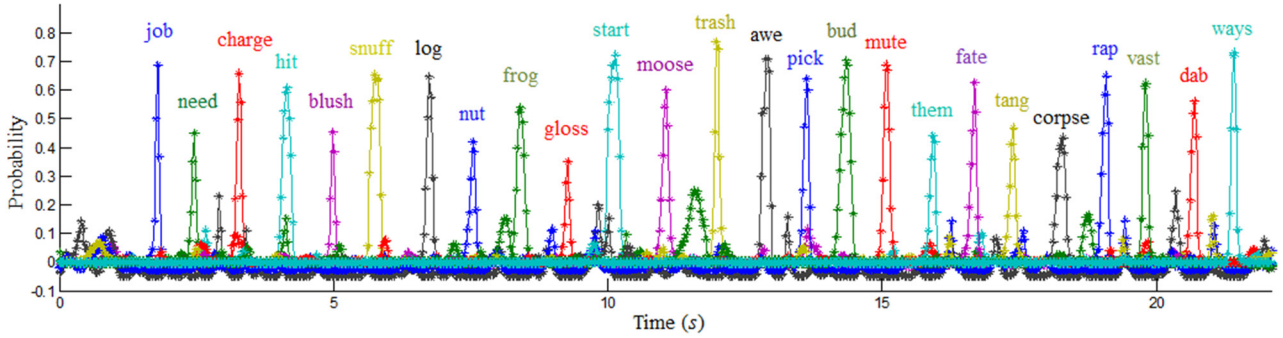

Figure 3. Example of probabilities (baseline removed) of 25 words on a test sequence.

Each word has its own threshold in design [2], but in the early stage of this work, a single threshold 0.20 was used for all words. If no candidates are found, w_r moves forward (to get more data), and the process returns to step 1. If a candidate is found, it is verified based on trends in its probability function (Step 3); 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. In step 4, *Temporal Proximity Constraint* allows only one word to occur within each time span. 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 step 5, after playing prerecorded audio samples of recognized words, the left boundary of the prediction window (w_l) moves forward. The whole procedure is repeated until the rightmost boundary of the prediction window (w_r) reaches the end of the input sequence. A word is recognized correctly if the word is identified within 100 ms of its actual occurrence time. Figure 3 illustrates the word probability distribution on a selected sequence.

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. This provides a rough 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, stimuli, and procedure

Ten healthy native female English speakers participated in data collection. Each speaker participated in one session in which she repeated a sequence of twenty-five words (i.e., one of the four phonetically-balanced word lists in [14], see Figure 3) multiple times. In all, 5500 word samples (in 220 sequences) were obtained and used in this experiment.

The electromagnetic articulograph (EMA) AG500 (Carstens Inc., Germany) was used to register the 3-D movements of the tongue, lips, and jaw when a subject was talking. The EMA records 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. Dental glue was used to attach the sensors. The spatial precision of motion tracking using EMA (AG500) is approximately 0.5 mm [15].

Figure 4 shows the positions of 12 sensors attached to the head, face, and tongue. Movement of the three head sensors (Head Center, Head Left, and Head Right) were collected and used to perform head-orientation calibration. Data from the four tongue sensors (named T1, T2, T3, and T4, or Tongue Tip, Tongue Blade, Tongue Body Front, and Tongue Body Back) and two lip sensors (Upper Lip and Lower Lip) were used for analysis. The movements of three jaw sensors, JL (Jaw Left), JR (Jaw Right), and JC (Jaw Center), were recorded for future use.

3.2. Data processing

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 4. Second, a zero phase lag low pass filter (i.e., 10 Hz) was applied for removing noise. Third, all sequences were manually segmented and annotated with words. Finally, only the y and z dimensions (Figure 4) of the sensors (i.e., T1, T2, T3, T4, UL, LL) were used for analysis because the movement along the side-to-side x axis is not significant in normal speech production.

4. Results & Discussion

Leave-one-out cross validation was conducted on the dataset from each subject in both training and online recognition, where one sequence of 25 words was used for testing and the rest sequences were for training. The classification (training) accuracy using our approach (direct mapping using SVM) was

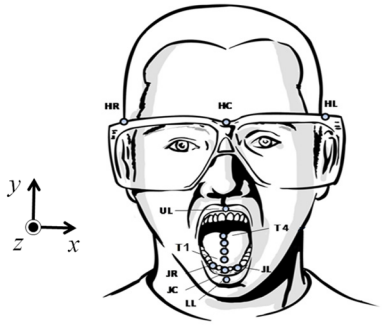

Figure 4. Positions of sensors attached on the subject's head, face and tongue in data collection.

93.71% and was not significantly different with that obtained using DTW (92.69%), which showed that a direct mapping is probably sufficient to capture the speaker-dependent articulatory movement patterns of isolated words. However, a *t*-test showed our approach (which took 0.5 ms, in average, for a single word classification) was significantly faster ($p < 0.00001$) than DTW (which took 1287.5 ms) in the same setting. Thus, only the direct mapping approach using SVM was used in online recognition.

On average, our online recognition algorithm failed to recognize 1.93 words (std. dev. σ=1.01) in a sequence of twenty-five words. The average difference of predicted word locations and their actual locations was 45 ms (σ=13) for correctly predicted words. On average, 8.08 (σ=4.45) false positives (words identified when there was no word) were reported in a sequence. Average prediction location offset and latency were 142 ms (σ=44) and 786 ms (σ=387) for a word prediction, respectively. The high accuracy shows the effectiveness of our proposed algorithm and the low prediction location offset and latency demonstrate the potential of our approach for real-time applications. Latency was measured on a PC with 3.1 GHz CPU and 8GB memory. The low standard deviations of the accuracy and other measures across subjects indicate that our approach can be applied generally with multiple subjects.

5. Conclusion & Future Work

Experimental results showed the potential of our whole-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, the recognition algorithm still has room for improvement. First, an automated approach to determine the optimal parameters (e.g., candidate thresholds) in online recognition can be developed. Second, the decision making in online prediction may be improved using some alternate strategies (e.g., area under curve and adaptive thresholds), which are particularly needed for reducing the number of false positives. Third, the recognition accuracy is dependent on the model training accuracy. Fortunately, our design is easily adapted to classifiers that associate candidates with probabilities. Faster DTW algorithms [16] and other classifiers (e.g., HMM [17, 18]) will be investigated.

Additional future work will be conducted including collecting a larger dataset with more functional words (e.g., a core vocabulary of 250 words used by augmented communicators, which represent 78% of the total words used [19]) to determine the scalability of our design. Although the EMA was used for collecting data and evaluating our algorithm, it may be cumbersome in practice. A portable EMA (e.g., Wave Speech Research System, NDI Inc., Canada) will therefore be investigated for practical use. Finally, four tongue sensors may cause discomfort for users (patients) after extended use, although the sensors are tiny. A further study will determine the minimum number of tongue sensors that still maintain a high level of silent speech recognition accuracy.

6. Acknowledgments

This work was in part funded by a grant awarded by the National Institutes of Health (R01 DC009890/DC/NIDCD NIH HHS/United States) and the Barkley Trust, University of Nebraska - Lincoln. We would like to thank Dr. Tom D. Carrell, Dr. Lori Synhorst, Cynthia Didion, Rebecca Hoesing, Kate Lippincott, Kayanne Hamling, Kelly Veys, and Taylor Boney for their contribution to subject recruitment, data collection, 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. (2006), pp. 1779-1780.

[2] Wang, J., Samal, A., Green, J. R., and Rudzicz, F., "Sentence recognition from articulatory movements for silent speech interfaces", *Proc. ICASSP*, 4985-4988, 2012.

[3] 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.

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

[5] 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.

[6] 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. ICASSP*, 2007.

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

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

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

[10] 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.

[11] Kantor, A., "Pronunciation Modeling For Large Vocabulary Speech Recognition", PhD Dissertation, Dept. Comput. Sci., University of Illinois, Urbana, 2011.

[12] Chang, C. -C., and Lin. C. -J., "LIBSVM: a library for support vector machines", *ACM Trans. Intell. Syst. Technol.*, 2(27):1-27, 2011.

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

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

[15] Yunusova, Y., Green, J. R., and Mefferd, A., "Accuracy assessment for AG500 electromagnetic articulograph", *J. Speech, Lang., and Hearing Research*, 52:2, 547-555, 2009.

[16] Salvador, S., and Chan, P., "Toward accurate dynamic time warping in linear time and space", *ACM Intell. Data Anal.* 11(5): 561-580, 2007.

[17] Heracleous, P., and Hagita, N. Automatic recognition of speech without any audio information, *Proc. ICASSP*, 2392-2395, 2011.

[18] 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*, 1005-1008, 2011.

[19] Stuart, S., Beukelman, D. R., and King, J. "Vocabulary use during extended conversations by two cohorts of older adults", *Augmentative and Alternative Communication*, 13:40-47, 1997.