8-2010

Vowel Recognition from Continuous Articulatory Movements for Speaker-Dependent Applications

Jun Wang
University of Nebraska - Lincoln, jwang3@unl.edu

Jordan R. Green
University of Nebraska-Lincoln, jgreen4@unl.edu

Ashok Samal
University of Nebraska - Lincoln, asamal1@unl.edu

Tom D. Carrell
University of Nebraska - Lincoln, tcarrell@unl.edu

4th International Conference on Signal Processing and Communication Systems (ICSPCS), 2010
doi: 10.1109/ICSPCS.2010.5709716

Vowel Recognition from Continuous Articulatory Movements for Speaker-Dependent Applications

Jun Wang[1], Jordan R. Green[2], Ashok Samal[1], Tom D. Carrell[2]
[1]Department of Computer Science & Engineering, {junwang, samal}@cse.unl.edu
[2]Department of Special Education & Communication Disorders, {jgreen4, tcarrell}@unl.edu
University of Nebraska-Lincoln, Lincoln, NE, United States

***Abstract*—A novel approach was developed to recognize vowels from continuous tongue and lip movements. Vowels were classified based on movement patterns (rather than on derived articulatory features, e.g., lip opening) using a machine learning approach. Recognition accuracy on a single-speaker dataset was 94.02% with a very short latency. Recognition accuracy was better for high vowels than for low vowels. This finding parallels previous empirical findings on tongue movements during vowels. The recognition algorithm was then used to drive an articulation-to-acoustics synthesizer. The synthesizer recognizes vowels from continuous input stream of tongue and lip movements and plays the corresponding sound samples in near real-time.**

Keywords-articulation; recognition; machine learning; support vector machine

I. INTRODUCTION

Oral communication is arguably the most natural and efficient mode of human communication. Currently, there are only limited options to maintain oral communication for individuals with severe speech motor impairments or laryngectomy (surgical removal of larynx due to the treatment of cancer). It is estimated that 1.5 - 2.0 million children or adults suffer from cerebral palsy, which is often associated with significant speech motor impairment in the United States [1]. Each year, about 12,500 new cases of laryngeal cancer [2] and 2,500 new cases of hyperlaryngeal cancer [3] are diagnosed in the United States. In the absence of good options for oral communication, patients communicate via other modalities with the assistance of Augmented and Alternative Devices (AAC), e.g., a text-to-speech synthesizer by typing. Currently, AAC devices are limited to text input or relatively slow forms of manual input.

Our long-term goal is to develop a real-time articulation-driven speech synthesizer that can compensate for aphonia and poor speech motor control, enabling the production of speech using movements of the tongue and lips for individual patients. The need for this technology was discussed previously by Paush [1], who wanted to improve oral communication in patients with cerebral palsy by playing synthesized speech acoustics from articulatory movement directly, a goal that proved to be extremely challenging because human can produce the same sound in different ways of articulation. That is, the mapping between articulatory movements to speech is many-to-one [4], [5], [6].

Additional major challenges to this research include limited options for tracking tongue movements, high degree of variability in speech movements. Most published work in this domain has used only lip or facial data, so-called visual speech recognition, or automatic lip reading [8], because recording tongue motion is logistically difficult. The lip and facial data are also commonly used as an extra input source for acoustic speech recognition in so-called articulatory speech recognition [9] or audio-visual speech recognition [7], [8]. However, the tongue is a very important articulator, particularly, for vowels. Without tongue information, a high recognition accuracy (e.g., greater than 90%) is unlikely. Fortunately, recently developed electromagnetic articulography devices provide a reasonably affordable, noninvasive, and accurate way to track the 3D motions of tongue [10].

During speech, the spatial and temporal characteristics of articulatory movements for a given sound can vary considerably [11]. To address the variation of speech movement problem, most prior work on articulatory movement-based vowel recognition has focused on extracting articulatory features such as lip opening and tongue position. This approach is based on the assumption that a small set of articulatory features can be used to distinguish vowels. These features include lip rounding/lip opening, tongue tip position [12], [13], [14], [16], lip contour or area [8], [15], [16], visemes [17], vertical and horizontal lip apertures, angles of lips [18], lip opening height and width, velocity of lip opening/closing, acceleration of lip movement [19]. However, seldom have these features resulted in a recognition accuracy greater than 90%. The recognition rates of most of approaches range from 20s to 80s in percentage.

A final major challenge for recognition of continuous speech is to identify individual speech segments (e.g., vowels) in the continuous articulatory movements (i.e., the segmentation problem [7]). Most of the previous work for vowel recognition has focused on the recognition from pre-

This work was in part funded by the Barkley Trust, Barkley Memorial Center, University of Nebraska-Lincoln and a grant awarded by the National Institutes of Health (R01 DC009890/DC/NIDCD NIH HHS/United States).

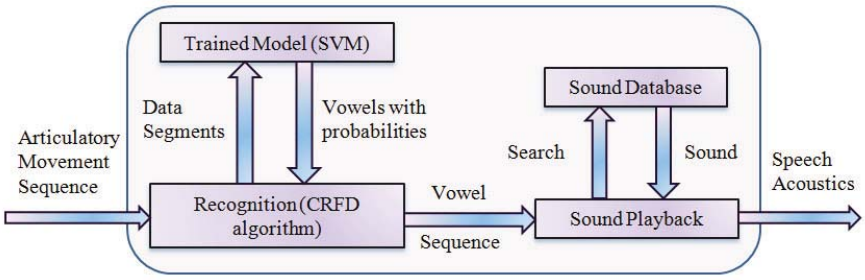

Figure 1. Design of the proposed articulation-to-acoustics synthesizer.

segmented data when the onset and offset of vowels are known.

The goal of this research is to obtain accurate recognition of vowels from continuous (unsegmented) tongue and lip movement, without using acoustic information. To address the challenges posed by the high spatial and temporal variation of articulators and the segmentation issue, we have developed an algorithm based on tongue and lip movement pattern classification (rather than on articulatory features) using a machine learning classifier (i.e., Support Vector Machine, or SVM [20]). At the same time, an algorithm called CRFD (Continuous Recognition with Fixed Delay) was developed to identify the individual vowel segments and their locations from the continuous articulatory movements by analyzing the probabilities from the output of the trained classifier.

The recognition algorithm then served as the recognition component of an articulation-to-vowel sound synthesizer. The synthesizer recognized vowels from continuous tongue and lip movements first, then played the corresponding vowel samples. A single-speaker dataset consisting of eight major English vowels was collected from a healthy native English speaker and used to evaluate the feasibility of our proposed approach. This research will serve as the foundation for developing a real-time word-level articulation-driven speech synthesizer for clinical applications.

II. Design & Method

A. Problem

The goal of this work is to recognize speech patterns from the time-series sequence of spatial coordinates. It is essentially is a time-series data classification problem, one of the top challenging topics in data mining research [29]. That is, given a dataset, D, of time-series sequences of 3D spatial coordinates of landmarks on tongue and lips, and a set of possible vowels, V, contained in those sequences, the research questions are (1) what vowels are in the sequences? and (2) when are the vowels produced? The synchronously recorded acoustic data are provided for segmenting the training data to segments associated with vowels, but not used for recognition. The following gives the formal definitions of the dataset D, vowel set V, and the research questions.

$D = <A_1, A_2, ..., A_K>$, is the dataset, where K is the number of articulators (Section III will give details of those articulators), and

$A_i = <X_i, Y_i, Z_i>$, $1 \le i \le K$, where X_i, Y_i, and Z_i are the time-series sequences of 3D spatial coordinates.

$X_i = <X_1, X_2, X_3, ..., X_n>$, is a time-series sequence of x coordinate, where n is the length of the sequence;

$Y_i = <Y_1, Y_2, Y_3, ..., Y_n>$, is a time-series sequence of y coordinate, where n is the length of the sequence;

$Z_i = <Z_1, Z_2, Z_3, ..., Z_n>$, is a time-series sequence of z coordinate, where n is the length of the sequence;

$V = <v_1, v_2, ..., v_m>$, is the possible vowel set, where m is the number of vowels. Here, x, y, and z are axes of a 3D Cartesian coordinate system. The orientation of x, y, and z will be given in Section III.

Research Questions: Give a dataset D ($|D| = K$), vowel set V, what vowels v_j ($1 \le j \le |V|$) are in D? When are the vowels v_j ($1 \le j \le |V|$) are produced?

B. Design

Our approach to continuous vowel recognition (i.e., when onset and offset of vowels are not known) is based on a prior data-driven approach we developed to recognize vowels from pre-segmented articulatory movements [21]. First, the dataset D is partitioned into training data and testing data. In the training procedure, we manually segment the sequences of articulatory movements to segments associated with vowels by aligning them to synchronously recorded acoustic data. These segments are used to train a classifier (i.e., SVM). Then CRFD is used to identify the vowels and their occurrence times by analyzing their associated probabilities determined by the classifier. After the vowels are recognized, the corresponding pre-recorded sounds are played back in the same order in which they are recognized.

These steps, i.e., training, recognition, and playback form the three components of the articulation-to-acoustics synthesizer or converter, as illustrated in Fig. 1.

C. Model Training & Parameter Estimation

The training module (in Fig. 1) consists of a computational model (a classifier, SVM) tuned to recognize vowels from segmented articulatory movement data [21]. That is, the model takes the beginning and end of a segment as the onset

and offset of a vowel production and computes the probability of the segment being a specific vowel.

Efforts were made to minimize the stages of data processing because our long-term goal is to develop real-time applications, which will require very rapid on-line recognition. This approach was designed so that the time-intensive calculations required for training are done off-line, prior to recognition. In addition, the trained classifier recognizes candidate vowels directly from minimally processed articulatory movement time-series data, rather than derived articulatory features (e.g., lip opening).

Training. First, the movement data for training are manually segmented based on synchronously recorded sounds (acoustic waveforms). Then the segmented data are time-normalized and sampled to fixed-width (the classifier requires fixed-width input) vectors of attributes for each articulator. An attribute represents the location of an articulator at a given time. Third, vectors for all articulators are concatenated as a composite vector, to represent a vowel sample. Finally, the classifier is trained using these vectors with their associated vowels (labels). Here an articulator is a sensor attached on the surface of tongue or lips. Section III will give details of the articulators and how data are preprocessed for training.

In addition to training a classifier, several important parameters which will be used in CRFD (Section II-D) are obtained during training: (a) *minlen* (minimum vowel length), (b) *maxlen* (maximum vowel length), and (c) *thresholds*, an array of minimum probabilities of correct prediction for all vowels that are used to select the candidate vowels in CRFD. The first two parameters are easily obtained by checking the lengths of all segmented vowel data. The array *thresholds*, minimum of maximum probabilities of vowels across all training sequences, is obtained in the procedure as following.

A variable length sliding window approach is used to find the *thresholds* from training sequences. At each time *t*, the window size is estimated to be the length of a vowel and is varied from *minlen* to *maxlen* with a step size Δ*len*. Δ*len* is a user-defined parameter. There is a trade-off in real-time applications. The smaller the Δ*len* is, the more values can be obtained, but more time is needed. From our experience, we have verified a value of 50 *ms* produces accurate results with acceptable high speed. The data within the sliding window at all time *t* are sent to the training model for probability calculation. Maximum of probabilities in each training sequence are saved for all vowels. Then, the minimum of the maximum probabilities across all training sequences for all vowels are saved as *thresholds*, where *thresholds(v), v* $\in$ *V*, means the least probability value that vowel *v* can be recognized. Section II-D will give details how *thresholds* is used and Section IV will give the values of *thresholds* in the experiment.

D. Recognition (CRFD Algorithm)

The recognition component (in Fig. 1) is the focus of this paper; it recognizes vowels from an unsegmented sequence of articulatory motion by analyzing the probability values returned from the trained model. In addition to recognizing the vowels, it also determines when the vowels are produced in test sequences.

The rationale of CRFD is that the correct vowel should have higher probabilities than any other vowels at the time when it occurs. However, the onset and length of each vowel is unknown. Therefore we need to determine the location (in time) and estimate the length for each vowel.

The core idea of the CRFD algorithm is to progressively examine a given sequence and determine the vowels with highest probabilities along the way. Two user-specified constant parameters, *delay* and *HighPeakThreshold*, are used to optimize the execution of CRFD. The parameters are explained in the following step-by-step description of CRFD. The schematic of CRFD algorithm is given in Fig. 2.

Steps 1 and 2 (Fig. 2) are used for probability calculation. A sliding window is used to go through a test sequence. Data within the window are continuously sent to the trained model for probability calculation (Step 1), until it reaches a *delay* (Step 2), which defines how late CRFD executes after the beginning of the algorithm or after the previous execution of probability analysis (Step 3). There is a *delay*, because the algorithm has to wait until the speaker finishes articulating the vowel. In this early work, *delay* is a user-specified parameter.

Step 3 is for picking up candidate vowels by probability analysis. When a *delay* is reached, candidate vowels are found within the prediction range (between the previous *delay* time and the current *delay*), and returns a list of candidates, sorted by time. A candidate vowel *v* at time *t* must satisfy the following two conditions. First, the probability at time *t* is greater than *thresholds(v)*, that is

$$prob(v,t) \geq thresholds(v) \qquad \text{(condition 1)}$$

Second, the probability value is close to the maximum value from the beginning of this sequence to the current time location *t*. A vowel may occur more than once in a test sequence. So it cannot just simply find the vowel with maximum probability. Instead, those vowels with probabilities which are close to the maximum probability are all considered. There is a user-defined *HighPeakThreshold*,

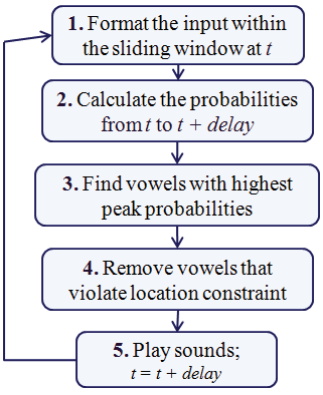

Figure 2. Schematic of the Continuous Recognition with Fixed Delay (CRFD) algorithm.

Attributes						Label
ULy_1, ULy_2,... $ULy_{n'}$	ULz_1, ULz_2,... $ULz_{n'}$	…	$T1y_1$, … $T1y_{n'}$	…	$T4z_1$.. $T4z_{n'}$	Vowel

Figure 4. Format of a vowel sample for classification (n'=10). The label is filled for training and empty for testing.

which defines the threshold that two probability values can be considered close or similar. Those peaks that has a difference with the maximum probability less than the *HighPeakThreshold* are considered as candidates, that is

$$|prob(v,t)-maxprob(v)| \leq HighPeakThreshold \quad \text{(condition 2)}$$

In this experiment, *HighPeakThreshold* is empirically given 0.05.

In Step 4, those candidates that violate Location Constraint (or Time Constraint) are removed. Location Constraint means that at most one vowel can be present at the same time. If two candidate vowels, v_1 and v_2, are recognized at time t_1 and t_2 with $| t_2 - t_1 | < minlen$ (means the two vowels are actually at the same time), the vowel with the lower probability is removed. Here, *minlen* is the minimum vowel length.

Finally, the corresponding sound samples of the recognized vowels are fetched from the sound database (Fig. 1) and played back in the order which the vowels are predicted.

The algorithm then clears the candidate list and repeats the procedure (Steps 1 - 4) until the sliding window reaches the end of the test sequence.

The time complexity of CRFD is $O(n \times l + n \times p \times |V|)$, where n is the length of the input sequence in time; l is a constant determined by (*maxlen* - *minlen*) / Δlen; $|V|$, number of possible vowels, is a constant for a given dataset; $p = \lfloor n / delay \rfloor$, number of executions of prediction (probability analysis). Thus, the overall time complexity of CRFD is $O(n^2)$.

III. Data Collection & Preprocessing

A. Participant, Stimuli, Device and Procedure

A single-speaker dataset of eight major English vowels in CVC (consonant-vowel-consonant) form, /gɑg/, /gig/, /gɐg/, /gæg/, /gʌg/, /gɔg/, /gog/, /gug/, was collected in this experiment. The speaker, a female native English-speaking college student produced the eight vowels sequentially at a normal speaking rate. The procedure was repeated 23 times, generating 23 productions of each vowel.

TABLE I. Articulators Used For Recognition

Articulator ID	Articulator Name	Location
1	UL	Upper Lip
2	LL	Lower Lip
3	T1	Tongue tip
4	T2	Tongue Body Front
5	T3	Tongue Body Back
6	T4	Tongue Back

Electromagnetic Articulograph (EMA) AG500 was used to record the 3D movements of tongue, jaw and lips during vowel production. Compared with X-ray and MRI, EMA is much more affordable while maintaining high resolution. The spatial precision of motion tracking using EMA (AG500) is approximately 0.5 *mm* [10]. The subject with attached sensors was seated with her head within an electromagnetic cube. When she spoke, the 3D coordinates of the sensors were recorded to a desktop computer connecting to the cube. The orientations of x, y, and z axes of the anatomically based coordinated system are illustrated in Fig. 3. Here, x, y, and z are defined as spatial dimensions width (left-right), height (up-down) and length (front-back) in the coordinate system.

Table I lists the names of the six articulators (sensors) which are used for recognition. Fig. 3 shows all twelve sensors (including the six articulators listed in Table I) attached on the subject's head, face, and tongue. HC (Head Center), HL (Head Left) and HR (Head Right) were attached to a pair of rigid glasses to avoid skin motion artifact [25]. The motion of HC, HL and HR were used to derive lip and tongue movement data that were independent from head motion. UL (Upper Lip) and LL (Lower Lip) were attached on the middle position of upper and lower lip. T1 (Tongue Tip), T2 (Tongue Body Front), T3 (Tongue Body Back) and T4 (Tongue Root) were attached on the midsagittal line on the tongue surface. The distance between adjacent tongue sensors was approximately 10 *mm* [23]. Three of the sensors, JL (Jaw Left), JR (Jaw Right) and JC (Jaw Center), are attached on the canines and one of the incisors. JL, JR, and JC were prepared for future use only.

B. Data Preprocessing

The time-series data of sensor locations derived from EMA

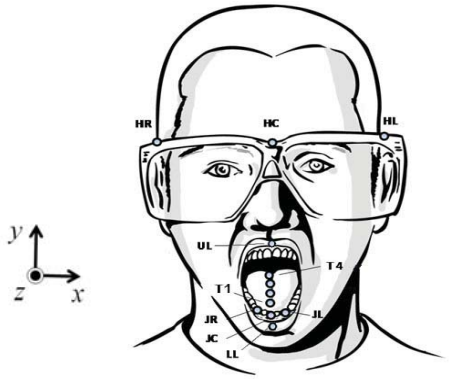

Figure 3. Sensor positions in data collection

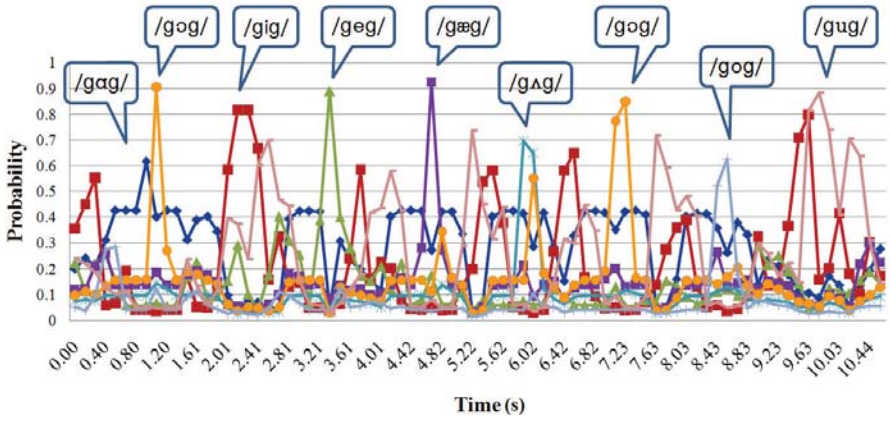

Figure 5a. Probability distribution of all vowels in a test sequence.

Expected Vowel	/gɑg/	/gig/	/geg/	/gæg/	/gʌg/	/gɔg/	/gog/	/gug/
Expected Occurrence Time	0.856	1.364	3.150	4.418	5.613	6.701	7.970	9.090
Actual Vowel	/gɔg/	/gig/	/geg/	/gæg/	/gʌg/	/gɔg/	/gog/	/gug/
Actual Occurrence Time	1.004	2.008	3.129	4.376	5.507	6.754	8.005	9.122

Figure 5b. Recognized vowels and their occurrence time (s) in the same test sequence.

need to be preprocessed prior to analysis. First, the head movements were subtracted from the lip and tongue data. Second, a low pass filter of 10 *Hz* was applied to the motion data for removing noise. Third, all sequences were segmented for each vowel by manually aligning the motion data with acoustic data recorded synchronously. To reiterate, the acoustic data were used only for segmenting training data and never used for recognition.

Only *y* and *z* coordinates were used in this research since the movement along the *x* axis is not significant in normal speech production [22].

After removing the mean of each dimension of articulators, all sampled frames of all articulators were concatenated as a vector of 120 (6 articulators × 2 dimensions × 10 frames) attributes that is used for classification, as illustrated in Fig. 4. Based on our previous work [21], 10 frames are sufficient to capture the motion patterns for vowels. Thus, formally, a subset (2D) of $A_i = <X_i, Y_i, Z_i>$, $1 \le i \le K$ in Section II are transformed to A'_i as following for classification.

$A'_i = <Y'_i, Z'_i>$, $1 \le i \le K$ ($K = 6$), *where*
$Y'_i = <Y_1, Y_2, Y_3, ..., Y_{n'}>$, $n' = 10$;
$Z'_i = <Z_1, Z_2, Z_3, ..., Z_{n'}>$, $n' = 10$.

IV. Results

A. Recognition Accuracy and Latency

The performance of the CRFD algorithm was measured in terms of its recognition accuracy and latency. Leave-One-Out (LOO) cross validation was conducted for measuring the accuracy. In each execution, one sequence was chosen for testing, and the rest were for training. There were 23 executions in total for 23 sequences. In each execution, a prediction was deemed correct only when both the predicted vowel was correct and if its occurrence time was close to the expected time (less than *minimum vowel length*, 0.375 *s* in this dataset). The average recognition rate of all executions was considered as the recognition accuracy of the CRFD algorithm. Latency was defined by the time between the onset of the recognition algorithm and the start of vowel sound playback.

Two *delay* values were tested on each sequence, 5 seconds and 10 seconds. The experimental results are summarized in Table II. In this experiment, CRFD was implemented using Matlab (MathWorks Inc.) with LIBSVM [20]. The experiment was executed on a laptop with 2.5G duo processor and 2G memory.

As anticipated, when the *delay* was longer, accuracy was higher and latency was longer. The latency was less than 1 second for both *delay* values. Most errors in CRFD, if not all, were caused by the unexpected probabilities returned by the trained classifier.

Fig. 5 gives the result on a selected sequence (length = 10.57 *s*), with a five-second *delay*. Fig. 5a illustrates the probability distribution of vowels. Fig. 5b gives the results of recognized vowels and their occurrence time. There is one error at the beginning. The highest probability of /gɑg/ (0.62) occurred at time 0.856. However, at time 1.004, /gɔg/ had a greater probability (0.90). The two time locations 0.856 and

TABLE II. Recognition Accuracy and Latency

Delay (s)	Accuracy	Latency (s)
5	83.15%	0.69
10	94.02%	0.73

1.004 are considered close (the difference of them is less than *minimum vowel length*, 0.375 *s*). Thus, the algorithm considers there is a /gɔg/ at time 1.004 (based on the Time Constraint).

B. Vowel Articulation Variation

The results also identified another interesting pattern. Table III gives the mean and standard deviation of maximum probability across test sequences for all vowels. Means of probabilities indicates that high tongue vowels are easier to distinguish than low tongue vowels. Standard deviations of probabilities show that high tongue vowels have lower prediction probability variation (means lower articulation variation) than low tongue vowels, which is consistent with previous empirical findings in phonetics [11], [26]. High tongue vowels (e.g., /i/, /u/) and low tongue vowels (e.g., /ɑ/), are categorized by the target position of tongue dorsum when the vowels are produced.

The column Min in Table III gives the actual values of *thresholds* in CRFD algorithm in the experiment.

TABLE III. MAXIMUM PROBABILITIES ACROSS TEST SEQUENCES FOR ALL VOWELS

Probability	Min (*thresholds*)	Mean	Std.Dev.	Tongue Position
/gɑg/	0.54	0.76	0.12	Low
/gig/	0.73	0.90	0.05	High
/geg/	0.46	0.85	0.12	Middle High
/gæg/	0.30	0.77	0.22	Low
/gʌg/	0.53	0.82	0.13	Middle Low
/gɔg/	0.40	0.83	0.14	Middle Low
/gog/	0.51	0.73	0.12	Middle High
/gug/	0.76	0.91	0.06	High

V. DISCUSSION & FUTURE WORK

This investigation developed and tested a novel algorithm for detecting vowels from continuous recordings of tongue and lip movements during vowel production. Recognition accuracy on the single-speaker dataset of eight major English vowels in CVC form, /gɑg/, /gig/, /geg/, /gæg/, /gʌg/, /gɔg/, /gog/, /gug/, was 94.02% with a very short latency. As expected, better results were obtained with the longer *delay* (10 seconds) than shorter *delay* (5 seconds).

The approach is unique in that it identifies the vowels using a direct mapping of the articulatory movements (rather than on derived articulatory features) to vowels. While we use a support vector machine for classification, the approach can be easily adapted to use other classifiers; any classifier (e.g., Hidden Markov Model) that gives probability or confidence can be seamlessly integrated into CRFD. Because a direct mapping approach is used, there is no computational cost of deriving features during recognition. This approach, therefore, may be ideally suited for real-time applications. Moreover, because the training is based on the motion patterns of the articulators, this approach should also apply to the recognition of other speech units including consonants and words [23].

The playback component (see Fig. 1) plays corresponding vowel samples in this prototype implementation. Other synthesis-based approaches to the output will be explored in the future [24]. For example, a text-to-speech synthesis (TTS) engine could be used to produce synthesized speech with different sounding voices, even using the patient's own voice recorded pre-surgery [27].

The algorithm is intended to eventually serve as the speaker-dependent recognition component of a real-time articulation-to-speech synthesizer. The articulation-driven synthesizer may provide an efficient mode of communication for individuals who rely on Augmentative and Alternative Communication (ACC) devices.

However, before our algorithms can be implemented for such purposes, the portability of tongue tracking device needs to improve. Fortunately, motion tracking technologies are improving rapidly and becoming increasingly affordable, more accurate, and smaller in size. For example, NDI Inc. (www.ndigital.com) has recently developed a relatively small and portable electromagnetic tracking device, Speech Wave System. These rapid advances in tongue motion tracking technologies suggest that barrier to progress toward developing a functioning real-time articulatory-movement based synthesizer will lie primarily in algorithms rather than hardware development.

Although these results obtained in this paper are very encouraging, future work is required (1) to improve the recognition accuracy for shorter *delay* values with minimized number of user-define parameters, (2) to extend recognition and test the approach using larger datasets of more vowels, consonants, words, and even sentences, and (3) to automatically segment training data [7], [28], which is necessary when larger datasets are available in the future.

ACKNOWLEDGMENT

We would like to thank Dr. Mili Kuruvilla for her technical support in data collection.

REFERENCES

[1] R. Paush and R. D. Williams, "Giving CANDY to children: User-tailored gesture input driving an articulator-based speech synthesizer," Univ. of Virginia Computer Science Report, TR-91-23, 1991.

[2] W. Samuel and M. D. Beenken, "Laryngeal cancer (cancer of the larynx)," *Armenian Health Network.* Online: http://www.health.am/cr/laryngeal-cancer/. Retrieved on 2009-06-23.

[3] W. M. Mendenhall, C. E. Riggs, and N. J. Cassisi, "Treatment of head and neck cancers," in Cancer: Principles and Practice of Oncology, 7th Edition, Chapter 26.2, V. T. DeVita, S. Hellman, S. A. Rosenberg Eds. Philadelphia: Lippincott, Williams, & Wilkins, 2005, pp. 642-743.

[4] J. Jiang, A. Alwan, P. A. Keating, E. T. Auer, and L. E. Bernstein, "On the relationship between face movements, tongue movements and speech acoustics," *Journal on Applied Signal Processing(EURASIP)*, vol. 11, pp. 1174--1188, 2002.

[5] C. T. Kello, and D. C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *Journal of Acoustic Society of America*, vol.

116, pp. 2354--2364, 2004.

[6] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, pp. 555-568, 2002.

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

[8] G. Potamianos, C. Neti, G. Gravier, A. Garg and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. of IEEE*, vol. 91, no. 9, 2003.

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

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

[11] J. S. Perkell, and M. H. Cohen, "An indirect test of the quantal nature of speech in the production of the vowels /i/, /ɑ/ and /ʉ/," *Journal of Phonetics*, vol. 17, pp. 123-133, 1989.

[12] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-Articulator Markov Models for speech recognition," *ISCA Tutorial and Research Workshop on Automatic Speech Recognition*, pp. 133-139, 2000.

[13] K. Saenko, T. Darrell and J. R. Glass, "Articulatory features for robust visual speech recognition," In *Proceedings of the 6th international Conference on Multimodal interfaces (ICMI04)*, State College, PA, USA, Oct 2004.

[14] K. Saenko, K. Livescu, J. Glass, and T. Darrell, "Multistream Articulatory Feature-Based Models for Visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(9), pp. 1700-1707, 2009.

[15] T. Shinchi, Y. Maeda, K. Sugahara, and R. Konishi, "Vowel recognition according to lip shapes by using neural network," *Proc. of IEEE*, 1998.

[16] A. J. Goldschen, O. N. Garcia, and E. Petajau, "Continuous optical automatic speech recognition by lipreading," *The 28th Annual Asilomar conference on Signals, Systems, and Computers*, pp. 572-577, 1994.

[17] M. Visser, M. Poel, and A. Nijholt, "Classifying visemes for automatatic lipreading," *LNCS*, vol. 1692, pp. 349--352, Springer, Heidelberg, 1999.

[18] M. N. Kaynak, Zhi Qi, A. D. Cheok, K. Sengupta, Jian Zhang, Ko Chi Chung, "Analysis of lip geometric features for audio-visual speech recognition," IEEE Trans. on Systems, Man and Cybernetics, Part A, vol. 34(4), pp. 564-570, 2004.

[19] P. Cosi, and E. Caldognetto, "Lips and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications," *NATO ASI Series F Computer And Systems Sciences*, vol. 150, pp. 291-314, 1996.

[20] C. C. Chang and C. J. Lin, "LIBSVM:a library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[21] J. Wang, A. Samal, J. R. Green, and T. D. Carrell, "Vowel recognition from articulatory position time-series data," *IEEE Intl. Conf. on Signal Processing and Communication Systems (ICSPCS)*, pp. 1-6, 28-30, September 2009.

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

[23] J. R. Green, and Y. T. Wang, "Tongue-surface movement patterns during speech and swallowing," *Journal of Acoustical Society of America*, vol. 113(5), pp. 2820-2833, 2003.

[24] J. R. Green, L. Phan, I. Nip, and A. Mefferd, "A real-time articulatory controlled vowel synthesizer for research on speech motor learning, " *Stem-, Spraak- en Taalpathologie,* 14(Supp.), 46, 2006.

[25] J. R. Green, E. M. Wilson, Y. Wang, and C. A. Moore, Estimating mandibular motion based on chin surface targets during speech, *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 928-939, 2007.

[26] J. S. Perkell and W. L. Nelson, "Variability in production of the vowels /i/ and /ɑ/," *Journal of Acoustic Society of America*, vol. 77, pp. 1889-1895, 1985.

[27] D. Yarrington, J. Gray, C. Pennington, H. T. Bunnell, A. Cornaglia, J. Lilley, K. Nagao, and J. B. Polikoff, "ModelTalker Voice Recorder – An Interface System for Recorded a Corpus of Speech for Synthesis," *Proceedings of the ACL-08: HLT Demo Session*, Columbus, OH, pp. 28-31, 2008.

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

[29] Q. Yang and X. Wu, "10 challenging problems in data mining research," *International Journal of Information Technology and Decision Making*, vol. 5, no. 4, pp. 597-604, 2006.