Speaker-Independent Silent Speech Recognition with Across-Speaker Articulatory Normalization and Speaker Adaptive Training

Jun Wang¹,²,³, Seongjun Hahm¹

¹Speech Disorders & Technology Lab, Department of Bioengineering
²Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
³University of Texas Southwestern Medical Center, Dallas, Texas, United States
{wangjun, seongjun.hahm}@utdallas.edu

Abstract

Silent speech recognition (SSR) converts non-audio information (e.g., articulatory information) to speech. SSR has potential to enable laryngectomees to produce synthesized speech with a natural sounding voice. Despite its recent advances, current SSR research has largely relied on speaker-dependent recognition. High degree of variation in articulatory patterns across different talkers has been a barrier for developing effective speaker-independent SSR approaches. Speaker-independent approaches, however, are critical for reducing the large amount of training data required from each user; only limited articulatory samples are often available for individuals, due to the logistic difficulty of articulatory data collection. In this paper, we investigated speaker-independent silent speech recognition from tongue and lip movement data with two models that address the across-talker variation: Procrustes matching, a physiological approach, to minimize the across-talker physiological differences of articulators, and speaker adaptive training, a data-driven approach. A silent speech data set was collected using an electromagnetic articulograph (EMA) from five English speakers (while they were silently articulating phrases) and was used to evaluate the two speaker-independent SSR approaches. The long-standing Gaussian mixture model-hidden Markov models and recently available deep neural network-hidden Markov model were used as the recognizers. Experimental results showed the effectiveness of both normalization approaches.

Index Terms: silent speech recognition, Procrustes matching, speaker adaptive training, hidden Markov models, deep neural network

1. Introduction

People with speech or voice impairment struggle for social interaction in their daily life. The options for augmenting oral communication of these individuals are currently very limited [1]. Although laryngectomees (who has undergone a surgical removal of the larynx for the treatment of laryngeal cancer) have several options to talk (i.e., esophageal speech, tracheoesophageal speech or TEP, and electrolarynx), these approaches frequently produce abnormal sounding voice that is hard to understand by listeners [2].

Silent speech interfaces (SSIs), a novel technological paradigm, may overcome these limitations by allowing people to generate natural sounding speech from the movements of their tongue and lips [3]. SSIs typically include a speech movement recorder, a speech movement recognizer (silent speech recognition, SSR) [4], and a speech playback device (text-to-speech synthesizer). Another potential of SSI is to drive the speech output using the laryngectomee’s own voice [5]. A number of techniques have been used to record speech movements or other non-audio information including ultrasound [6, 7], surface electromyography [8, 9], and electromagnetic articulograph (EMA) [10, 11]. Text-to-speech synthesis has been well studied and is ready for this application. Therefore, the core problem in current SSI research is developing effective algorithms that convert articulatory information to speech [12, 13].

Despite the recent advances in SSR research, current approaches have largely relied on speaker-dependent recognition models, where training data and testing data are from the same speaker [14]. High degree of variation in the articulatory patterns across different talkers has been a barrier for developing effective speaker-independent SSR approaches. Multiple sources contributed to the inter-talker variation including gender, dialect, individual vocal tract anatomy, and different coarticulation patterns [15]. However, speaker-independent approaches are critical for reducing the amount of training data required from each user. Only limited articulatory data samples are often available, either from healthy subjects or individuals with speech impairments, due to the logistic difficulty of articulatory data collection [16].

To minimize the physiological inter-talker differences, researchers have tried to normalize the articulatory movements by aligning the tongue position when producing vowels [17–19], consonants [20, 21], and pseudo-words [22] to a reference (e.g., palate [17, 18], or a general tongue shape [20]). Procrustes matching, a bidimensional shape analysis technique [23], has been used to minimize the translational, scaling, and rotational effects of articulatory data across speakers [21, 22] and in inter-talker phrase classification [14]. Procrustes matching, however, has rarely been used in continuous silent speech recognition.

Speaker adaptive training (SAT), a data-driven approach for speaker adaption, has been widely used for feature space normalization in speaker-independent (acoustic) speech recognition [24]. However, speaker adaptive training (e.g., using feature space maximum likelihood linear regression, fMLLR) has rarely been used in silent speech recognition.

In this paper, we investigated the use of Procrustes matching (a physiological approach) and SAT (a data-driven approach) in speaker-independent SSR. Tongue and lip movements were recorded using an electromagnetic articulograph (EMA) from multiple speakers while silently producing short phrases. Two recognition models, long-standing Gaussian
mixture model-hidden Markov model (GMM-HMM) and recently available deep neural network-hidden Markov model (DNN-HMM) [25–27], were used in the experiment. Speaker-independent recognition performance were compared with or without speaker normalization to determine the effectiveness of the normalization approaches (Procrustes matching and SAT).

2. Method

2.1. Procrustes matching: A physiological approach

Procrustes matching (or Procrustes analysis [23]) is a robust statistical bidimensional shape analysis, where a shape is represented by a set of ordered landmarks on the surface of an object. Procrustes matching aligns two objects after removing the locational, rotational, and scaling effects [16, 28].

In this project, Procrustes matching was used to match the inter-talker physiological difference (tongue size and orientation). The downsampled time-series multi-sensor and multidimensional articulatory data form articulatory shapes. An example is shown in Figure 1 [14]. This shape contains 40 landmarks that are discretized from the continuous motion paths of four sensors attached on tongue and lips, named as TT (Tongue Tip), TB (Tongue Body Back or Tongue Dorsum), UL (Upper Lip), and LL (Lower Lip). Details on the sensor setup will be provided in Section 3 (data collection). A step-by-step procedure of Procrustes matching between two shapes includes (1) aligning the centroids of the two shapes, (2) scaling the shapes to a unit size, and (3) rotating one shape to match the other [16, 28].

Let \( S \) be a set of landmarks (sensors) as shown below:

\[
S = \{(y_i, z_i)\}, \quad i = 1, \ldots, n
\]  

where \((y_i, z_i)\) represents the \(i\)-th data point (spatial coordinates) of a sensor, and \(n\) is the total number of data points (\(n = 40\) in Figure 1), where \(y\) is vertical and \(z\) is front-back. The transformation in Procrustes matching is described using parameters \((\{e_y, e_z\}, \{\beta_y, \beta_z\}, \theta)\):

\[
\begin{bmatrix}
  y_i \\
  z_i
\end{bmatrix}
= 
\begin{bmatrix}
  \cos \theta & -\sin \theta \\
  \sin \theta & \cos \theta
\end{bmatrix}
\begin{bmatrix}
  \beta_y \\
  \beta_z
\end{bmatrix}
\begin{bmatrix}
  y_i - c_y \\
  z_i - c_z
\end{bmatrix}
\]  

(2)

where \((c_y, c_z)\) are the translation factors (centroids of the two shapes); Scaling factor \(\beta\) is the square root of the sum of the squares of all data points along the dimension; \(\theta\) is the angle to rotate [23]. The parameters were estimated or determined in the procedure described below.

For each participant, the entire articulatory shape (Figure 1) was transformed into an “normalized shape”, which had a centroid at the origin \((0, 0)\) and aligned to the vertical line formed by the average positions (centroids) of the upper and lower lips. Scaling was not used in this experiment, because preliminary tests indicated scaling will cause worse performance in speaker-independent silent speech recognition.

Specifically, the normalization procedure was done in two steps. First, all articulatory data (e.g., a shape in Figure 1) of each speaker were translated to the centroid of that shape (average position of all data points in the shape). This step removed the locational effects between speakers. Second, all shapes of speakers were rotated to make sure the sagittal plane was oriented such that the centroid of lower and upper lip movements defined the vertical axis. This step reduces the variation of rotational effects due to the difference in facial anatomy between speakers. Thus in Eq. 2, \((e_y, e_z)\) are the centroid of shape \(S\); the normalization approaches (Procrustes matching is provided in [14]).

Scaling factor \((\{\beta_y, \beta_z\})\) is one \([1 1]’\); \(\theta\) is the angle of the \(S\) to the reference shape in which upper lip and lower lip form a vertical line. An example of the shape before and after Procrustes matching is provided in [14].

2.2. Speaker adaptive training: A data-driven approach

Speaker adaptive training (e.g., using fMLLR) has been used to reduce speaker variation in acoustic speech recognition [24]. fMLLR (also called CMLLR; constrained maximum likelihood linear regression) is one of the representative approaches for across-speaker feature space normalization.

The procedure of speaker adaptive training is the same as maximum likelihood estimation (MLE) using transformed feature vectors by fMLLR. For each speaker \(s\), a transformation matrix \(A^s\) and a bias vector \(b^s\) are estimated and used for feature vector transformation:

\[
\hat{o}^s(t) = A^s o^s(t) + b^s
\]

(3)

where \(o^s(t)\) is the input feature vector for speaker \(s\) at frame \(t\) and is transformed to \(\hat{o}^s(t)\). This transformed \(\hat{o}^s(t)\) is used for training GMM-HMM or DNN-HMM. Here, speaker labels for observation \(o^s(t)\) is known for training stage. In testing stage, input feature vectors are also transformed using Eq. (3) before they are fed into GMM or DNN. A more detailed explanation of SAT and fMLLR can be found in [24].

2.3. Recognizers and Experimental Setup

The long-standing recognizer GMM-HMM and recently available DNN-HMM were used as the recognizers [25, 26, 29, 30]. In this experiment, frame rate was 10 ms (equivalent to sampling rate of recording: 100 Hz). Two dimensional (vertical and anterior-posterior) EMA data of four sensors (Tongue Tip, Tongue Body Back, Upper Lip and Lower Lip) were used (see Section 3 for details). For each frame, all articulatory movement features plus delta and delta of delta form a 24-dimension vectors that were fed into a recognizer. HMM is left-to-right 3-state with a monophone or a triphone context model. MLE training approach was used for training HMM. The input layer of DNN has 216 \((24 \times 9\) frames \(– 4\) previous plus current plus 4 succeeding frames\) dimensions. The output layer has 116 dimensions \((37\) phonemes \times \(3\) states \(+ 1\) silence \times \(5\) states\) and 305 dimensions for monophone and triphone models, respectively. The output posterior probabilities of DNN are used for decoding. We used 1 to 6 hidden layers and each layer had 1,024 nodes. The best performance obtained using 1 to 6 layers
were reported. In this early stage work, no extra optimization for DNN was used. A bi-gram phoneme-level language model was used. The training and decoding were performed using the Kaldi speech recognition toolkit [31].

Phoneme Error Rate (PER) was used as the measure of silent speech recognition performance. PER is the summation of substitution, insertion, and deletion errors of phonemes divided by the number of all phonemes.

Leave-one-subject-out cross validation was used in the experiment. In each execution, all samples from one subject were used for testing and the samples from the rest subjects were used for training. The average performance of executions was calculated as the overall performance.

3. Data Collection

A silent speech (articulatory) data set was collected from multiple talkers for this speaker-independent silent speech recognition experiment.

3.1. Participants and stimuli

Five American English talkers (3 females and 2 males; Mean of age 25; SD = 3.8) participated in the data collection. No history of speech, language, or cognitive problems were reported. Each subject participated in one session in which he/she silently repeated a sequence of sixty phrases multiple times at their habitual speaking rate. The phrases (e.g., how are you?) were selected from [10].

3.2. Tongue motion tracking device - Wave

The Wave System (Northern Digital Inc., Waterloo, Canada), one of the two commercially available electromagnetic tongue motion tracking devices, was used to collect the movement data of the head, tongue, and lips for all participants (Figure 2a). Wave records tongue movements by establishing a calibrated electromagnetic field that induces electric current into tiny sensor coils that are attached to the surface of the articulators. A similar data collection procedure has been used in [16, 32, 33]. The spatial precision of motion tracking using Wave is about 0.5 mm [34]. The sampling rate of recording was 100 Hz.

3.3. Procedure

Participants were seated with their head within a calibrated magnetic field (right next to the textbook-sized magnetic field generator; Figure 2a). Five sensors were attached to the surface of each articulator using dental glue (PeriAcryl 90, GluStitch) or tape, including one on the head, two on the tongue and two on the lips (Figure 2). A three-minute training session helped the participants to adapt to the wired sensors before the formal data collection.

Figure 2b shows the positions of the five sensors attached to a participant’s head, tongue, and lips. HC (Head Center) was on the bridge of the glasses. The movements of HC were used to calculate the head-independent movements of other articulators. TT (Tongue Tip), TB (Tongue Body Back) were attached at the mid-line of the tongue [16]. TT was about approximately 10 mm from the tongue apex. TB was as far back as possible and about 30 to 40 mm from tongue apex [16]. Lip sensors were attached to the vermilion borders of the upper (UL) and lower (LL) lips at mid-line. Data collected from TT, TB, UL, and LL were used for analysis. The four-sensor set was found optimal and practical for this application [35].

3.4. Data processing

Data processing was applied on the raw sensor position data prior to analysis (i.e., before Procrustes matching or fMLLR). First, the head translations and rotations were subtracted from the tongue and lip data to obtain head-independent tongue and lip movement data. The orientation of the derived 3D Cartesian coordinates system is displayed in Figure 2, in which x is left-right, y is vertical, and z is front-back. Second, a low pass filter (i.e., 20 Hz) was applied for removing noise [16, 33].

In total, 747 sentence samples (for unique sixty phrases) were obtained from the five participants and were used for analysis. It is not expected normal people have significant lateral movement (x in Figure 2b) [16], thus only y and z coordinates (Figure 2) of the tongue and lip sensors were used for analysis.

4. Results & Discussion

Figures 3 and 4 give the phoneme error rates of five configurations using monophone or triphone context models, respectively: (1) speaker-dependent (SD) using GMM-HMM as baseline, (2) speaker-independent (SI) without normalization, (3) speaker-independent with fMLLR, (4) speaker-independent with Procrustes matching, and (5) speaker-independent with Procrustes and fMLLR. These results suggest that both Procrustes matching and fMLLR were effective for speaker-independent silent speech recognition. The best results were obtained when two approaches were used together with DNN-HMM (56.3% in monophone model and 47.5% in triphone model) (Figures 3 and 4). Due to the small data size from each individual speaker, there was no speaker-dependent baseline result using DNN-HMM.

Procrustes matching outperformed fMLLR in all configu-
Healthy [38] and disordered populations [39], and EMA-based imaging speech [16], recognition of speech with articulatory data for may have implications for studies on tongue kinematics during EMA based SSI [37].

We think both approaches can be used in MVOCA, a portable technique that is not a focus of this paper, the results provide a reference for recognizer selection in future studies.

Surprisingly, the speaker-independent monophone recognition (GMM-HMM: 57.0% and DNN-HMM: 56.3%) obtained even better results than the speaker-dependent monophone baseline (58.4%). The speaker-independent triphone recognition (GMM-HMM: 52.5% and DNN-HMM: 47.5%) also obtained better results than the speaker-dependent triphone baseline (55.2%).

Adaptability for online recognition. Although the analysis was offline, both normalization approaches (Procrustes and fMLLR) can be easily integrated into online recognition. Procrustes matching does not require pre-recorded training data and the transformation matrix in fMLLR are obtained in testing stage. Of course, Wave system is currently too cumbersome for clinical applications. However, both approaches are not dependent on a particular data acquisition device. For example, we think both approaches can be used in MVOCA, a portable EMA based SSI [37].

Other implications. The two normalization approaches may have implications for studies on tongue kinematics during speech [16], recognition of speech with articulatory data for healthy [38] and disordered populations [39], and EMA-based speech training [40].

Limitations. Although the experimental results were encouraging, the data set used in the experiment contained only a small number of unique phrases collected from a small number of subjects. Further studies with a larger vocabulary from a larger number of subjects with different genders and ages are necessary to explore the limits of the current approaches.

5. Conclusions & Future Work

This paper presented speaker-independent silent speech recognition using two across-speaker normalization approaches: Procrustes matching, a physiological approach, and fMLLR, a data-driven approach. GMM-HMM and DNN-HMM were used as the recognizers. Experimental results showed the effectiveness of both normalization approaches. The best performance was obtained when the two normalization approaches were used together with DNN-HMM.

Future directions include (1) test of the two normalization approaches using a larger data set collected from more subjects, and (2) evaluating the efficacy of the speaker-independent SSR approaches using data collected from laryngectomees.

6. Acknowledgments

This work was supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under award numbers R03DC013990 and R01DC013547. We thank Dr. Jordan R. Green, Dr. Thomas F. Compbell, Dr. William Katz, Sonya Metha, Vedad Fazel, Marcus Jones, and the volunteering participants.
7. References


