Published in final edited form as:

IEEE/ACM Trans Audio Speech Lang Process. 2017 December ; 25(12): 2323–2336. doi:10.1109/TASLP.2017.2758999.

Speaker-Independent Silent Speech Recognition from Flesh-Point Articulatory Movements Using an LSTM Neural Network

Myungjong Kim [Member, IEEE],
Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080, United States

Beiming Cao,
Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080, United States

Ted Mau, and
Department of Otolaryngology-Head and Neck Surgery, University of Texas Southwestern Medical Center, Dallas, TX 75390, United States

Jun Wang [Member, IEEE]
Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080, United States

Abstract

Silent speech recognition (SSR) converts non-audio information such as articulatory movements into text. SSR has the potential to enable persons with laryngectomy to communicate through natural spoken expression. Current SSR systems have largely relied on speaker-dependent recognition models. The high degree of variability in articulatory patterns across different speakers has been a barrier for developing effective speaker-independent SSR approaches. Speaker-independent SSR approaches, however, are critical for reducing the amount of training data required from each speaker. In this paper, we investigate speaker-independent SSR from the movements of flesh points on tongue and lip with articulatory normalization methods that reduce the inter-speaker variation. To minimize the across-speaker physiological differences of the articulators, we propose Procrustes matching-based articulatory normalization by removing locational, rotational, and scaling differences. To further normalize the articulatory data, we apply feature-space maximum likelihood linear regression and i-vector. In this paper, we adopt a bidirectional long short term memory recurrent neural network (BLSTM) as an articulatory model to effectively model the articulatory movements with long-range articulatory history. A silent speech data set with flesh points was collected using an electromagnetic articulograph (EMA) from twelve healthy and two laryngectomized English speakers. Experimental results showed the effectiveness of our speaker-independent SSR approaches on healthy as well as laryngectomy speakers. In addition, BLSTM outperformed standard deep neural network. The best performance was obtained by BLSTM with all the three normalization approaches combined.

Index Terms

Articulatory normalization; long short term memory; Procrustes matching; silent speech recognition

I. INTRODUCTION

Laryngectomy is a surgical removal of the larynx for the treatment of laryngeal or other oral cavity cancers [1]; therefore, persons after laryngectomy lose their ability to produce speech sounds and suffer in their daily communication [2], [3]. Although there are several options to assist the speech communication for those patients such as esophageal speech, tracheaesophageal speech, and electrolaryngeal speech, these approaches generally produce an abnormal sounding voice with a pitch that is aberrantly low and limited in range [4]. Thus, it is necessary to develop an alternative technology with natural output voice for persons with laryngectomy.

Silent speech interfaces (SSIs) [6], [7], which represent a novel technological paradigm, have the potential to provide an alternative way to assist those patients to produce speech with natural sounding voices from the movements of their articulators such as the tongue and lips. SSIs typically include 1) an articulatory movement recorder, 2) a silent speech recognizer [7], and 3) a text-to-speech synthesizer [8], [9]. A variety of techniques have been used to record articulatory movements including ultrasound [10], [11], surface electromyography [12], [13], and electromagnetic articulograph (EMA) [14], [15] (portable EMA in [47]). Each technique has its own advantages in acquiring the movements of articulators [6]. In particular, EMA is a direct and attractive approach to record the flesh point articulatory movements since it captures the 3D motion of several sensors adhered to the tongue and lips [6]. We used EMA sensors in this work to track the precise Cartesian coordinates of articulators. Text-to-speech synthesis (TTS) then plays synthesized sounds based on the recognized text, which is well studied and is ready for this application (e.g., [15]). Researchers on TTS are currently exploring how to restore the laryngectomee's own voice [5, 55] with limited training data. Thus, the core problem in current SSI research is developing effective algorithms of silent speech recognition (SSR) that map articulatory movements to text.

One of the greatest challenges in SSR is speaker independence, primarily because only limited articulatory data samples from few speakers are often available due to the logistic difficulty and expense to collect articulatory data [17]. Moreover, the articulatory data are directly affected by the speaker's anatomy, such as the shape and size of articulators [6], [16], and the speaker's articulatory patterns [16]. This results in high variability of articulatory movements between speakers. Therefore, most SSR studies have focused on developing speaker-dependent recognition models [7], [10], [31], where training data and testing data are recorded from the same speaker. The high degree of variability in articulatory patterns across different speakers has been a barrier for developing effective speaker-independent SSR approaches. Speaker-independent approaches, however, are critical for reducing the amount of training data required from each speaker.

Two types of articulatory normalization approaches have been investigated for speaker-independent SSR: *physiological* or *data-driven*. In a *physiological approach*, articulatory data are normalized based on the physiological characteristics of articulators. Researchers have tried to normalize the articulatory movements by aligning the tongue position when producing vowels [18], [19], [20], consonants [21], [22], and pseudo-words [23] to a reference (e.g., palate [18], [19] or a general tongue shape [21]). Procrustes matching, a robust shape analysis technique [24], has been used to reduce the translational, rotational, and scaling effects of articulatory data across speakers [25], [37].

For a *data-driven approach*, which depends on the data variation, speaker normalization techniques have been widely studied for speaker-independent acoustic speech recognition in a transformation or footprint approach. In a transformation approach, the input features are transformed onto general feature space using a transformation matrix estimated from the data, e.g., feature-space maximum likelihood linear regression (fMLLR) [26]. In a footprint approach, speaker-specific information estimated from speech data is used as an additional input to the speech recognition system so as to adjust the model parameters to exploit the speaker information. Recently, i-vectors, which are low-dimensional vectors characterizing speakers, have been successfully used as additional speaker information for speaker-independent acoustic speech recognition [27], [28]. Despite the advances in data-driven speaker normalization techniques in acoustic speech recognition, it has rarely been studied in silent speech recognition.

Deep learning-based acoustic models such as deep neural networks (DNNs) have been widely applied in acoustic speech recognition [29], [30], showing significant improvements over the long-standing Gaussian mixture model (GMM)-based acoustic model. The promise of deep acoustic models to improve the speech recognition accuracy motivates the application of deep articulatory models in silent speech recognition. Deep articulatory models, however, have rarely been studied in silent speech recognition (i.e., without acoustic information). Hahm and Wang [31] applied DNN-based articulatory models to speaker-dependent silent speech recognition with EMA data. They reported that DNN-based articulatory models outperform conventional GMM-based articulatory models. However, it has limitations to model long-range temporal dependencies of articulatory movements.

This paper addresses the problem of speaker-independent automatic recognition of silent speech using only articulatory movement data collected from EMA sensors, focusing on articulatory normalization and recognition models. In our preliminary work [25], [37], articulatory normalization approaches based on Procrustes matching and fMLLR showed promising results with a small data set with GMM and standard DNN. This paper extends our prior work and the detailed contributions of the paper include the following:

- Articulatory normalization based on physiological and data-driven approaches. To this end, we apply three different normalization techniques to articulatory movement data: *Procrustes matching, fMLLR*, and *i-vectors*. First, Procrustes matching is used to normalize the articulatory difference resulting from speaker's anatomy. Then, fMLLR is further applied to normalize the articulatory movement patterns during articulation. Finally, i-vectors are extracted from fMLLR-normalized articulatory features to capture remaining speaker information and used as additional input to silent speech recognition. The three different normalization methods have their own purposes, and therefore, it is expected that the methods can successfully complement each other.

- Silent speech recognition based on Bidirectional LSTM-based deep articulatory models. Long short term memory (LSTM) recurrent neural networks [32] can capture long-range temporal dependencies by overcoming the vanishing gradient problem in conventional recurrent neural networks (RNN). It has shown excellent performance in many machine learning applications [33], [34]. Bidirectional LSTM (BLSTM) has been recently introduced to model temporal dependencies in forward and backward direction using two hidden layers that are connected to the same output layer [35], [36]. This model is able to effectively represent articulatory history in bidirectional context, because articulatory movements show strong temporal dependency. Therefore, it is expected that BLSTM is appropriate for silent speech recognition by jointly modeling articulatory movements in bidirectional context. To our knowledge, this paper is the first to apply BLSTM for silent speech recognition.
- The articulatory normalization techniques and the deep articulatory models are evaluated on silent speech recorded from healthy speakers as well as laryngectomy patients to verify the effectiveness of the methods.

The remainder of the paper is organized as follows: The data collection including participants, device, and data preprocessing is described in Section II. In Section III, we present the proposed method including articulatory normalization and BLSTM-based articulatory model in detail. Section IV shows experimental results and analyses demonstrating the effectiveness of the proposed method. Finally, our conclusions are summarized in Section V.

II. Data Collection

A. Participants & Speech Tasks

Twelve American English speakers (8 females and 4 males) participated in the data collection. The mean age of the participants was 25.1 ± 3.5 years. The range of the age was from 21 to 31. No history of speech, language, or cognitive problem from any participant was reported. Each subject participated in one session in which he or she repeated a list of phrases multiple times at their habitual speaking rate. Five of them repeated a list of 60 phrases a few times. Eight of them repeated a sequence of 132 phrases twice. The 60-phrase list is part of the 132-phrase list. The phrases (e.g., *how are you doing*?) were selected from [15] and [37].

In addition, we collected the silent speech data from two male laryngectomy patients (mean age 54). They are also American English speakers with no history of language or cognitive problems. They had their surgery four or five years ago. They both use trachea-esophageal (TE) speech as their current speech option. Each subject participated in one session in which he produced a sequence of 132 phrases at his habitual speaking rate.

B. Tongue and Lip Motion Tracking Device: Wave

The Wave system (Northern Digital Inc., Waterloo, Canada), a commercially available electromagnetic tongue and lip motion tracking device, was used to collect the movement data of the head, tongue, and lips for all participants as in Fig. 1(a). Wave records flesh point articulatory movements by establishing a calibrated electromagnetic field that induces electric current into tiny sensor coils that are attached to the surface of the tongue and lips. A similar data collection procedure has been used in [17], [38]. The spatial precision of motion tracking using Wave is about 0.5 mm [39]. The sampling rate of recording was 100Hz.

C. Procedure

Participants were seated with their heads within a calibrated magnetic field (right next to the textbook-sized magnetic field generator) as in Fig. 1(a). 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. A three minutes training session helped the participants to adapt to the wired sensors before the formal data collection.

Fig. 1(b) illustrates the positions of the five sensors attached to a participant’s head, tongue, and lips. HC (Head Center) was on the bridge of a set of glasses. We used glasses, rather than taping the sensor to the skin directly, to avoid skin artifact during speaking. The movement data of HC were used to calculate the head-independent data of other sensors. TT (Tongue Tip) and TB (Tongue Body Back) were attached at the mid-line of the tongue [17]. TT was 5–10 mm from the tongue apex. TB was as far back as possible and about 30–40 mm from TT [17]. Lip sensors were attached to the vermilion borders of the upper (UL) and lower (LL) lips at mid-line. The flesh point articulatory data collected from TT, TB, UL, and LL were used for analysis. Our prior work has shown the four-sensor set is an optimal set for this application, balancing the number of sensors and recognition performance [40], [48], [49]. Therefore, we used the four sensors in this work.

D. Data Processing

Raw sensor position data were processed prior to analysis (i.e., before Procrustes matching, fMLLR, or i-vector normalization). First, the head translation and rotation 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 Fig. 1(b), in which x is left-right, y is vertical, and z is front-back. Second, a low pass filter with cutoff frequency of 20 Hz was applied for removing noise [17], [39].

In total, 2,858 utterances (44,004 phonemes/11,865 words) and 264 utterances (4,512 phonemes/1,181 words) for a set of unique 132 phrases were collected from the twelve healthy participants and two laryngectomy patients, respectively, and were used for analysis. Here, the number of unique phonemes is 39 and the number of unique words is 278. It is not expected that articulators have significant lateral movement (x in Fig. 1(b)) [17], thus only y and z coordinates of the tongue and lip movement data were used for analysis.

III. Proposed Method

In this section, we explain the proposed silent speech recognition approach, including articulatory normalization based on Procrustes matching, fMLLR, and i-vector methods, as well as BLSTM-based deep articulatory model. The schematic diagram of the method is depicted in Fig. 2.

A. Procrustes Matching: A Physiological Approach

Procrustes matching or Procrustes analysis [24] is a robust statistical bidimensional shape analysis technique, which has been widely used in other fields such as human motion recognition [41], [42]. In Procrustes matching, a shape is represented by a set of ordered landmarks on the surface of an object. Procrustes matching aligns two objects by removing the locational, rotational, and scaling effects [17], [23].

In this work, we used Procrustes matching for normalizing the articulatory shapes due to the inter-talker physiological difference (tongue and lip orientation). The time-series multi-sensor and multi-dimensional articulatory data form articulatory shapes as in Fig. 3. This shape contains trajectories of the motion paths of four sensors attached on tongue and lips, i.e., TT, TB, UL, and LL. 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 [17].

Let S be a set of discrete landmarks from sensors on tongue and lips as shown below:

$$S = \left\{o_t = \left(y_t, z_t\right)\right\}, \quad t = 1, \ldots, T \qquad (1)$$

where (y_t, z_t) represents the tth data point (spatial coordinates) of a sensor, and T is the total number of data points in time, where y is vertical and z is front-back. The transformation in Procrustes matching is described using parameters $\{(c_y, c_z), (\beta_y, \beta_z), \theta\}$:

$$\begin{bmatrix}\bar{y}_t \\ \bar{z}_t\end{bmatrix} = \begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}\left(\begin{bmatrix}\beta_y \\ \beta_z\end{bmatrix} \mathrm{o} \begin{bmatrix}y_t - c_y \\ z_t - c_z\end{bmatrix}\right) \qquad (2)$$

where (c_y, c_z) are the translation factors (centroids of the two shapes); Scaling factor β is the inverse of the square root of the sum of the squares of all data points along the dimension; θ is the angle to rotate [24]; “o” is the element-wise multiplication. The parameters are estimated or determined in the procedure described below.

Each subject’s articulatory shape is transformed into a “normalized shape”, which has a centroid at the origin (0,0) and aligned to the vertical line formed by the average positions of the upper and lower lips. Specifically, the normalization is performed as follows: First, all the articulatory data, e.g., a shape in Fig. 3, of an utterance of each speaker are translated to the origin by the centroid, i.e., average position of all the data points in the shape. This step can remove the locational differences between speakers. Second, all the articulatory shapes are scaled to unit size. This step can reduce the scaling effects due to the difference resulting from the size of speaker's anatomy. Third, all the shapes of speakers are rotated to make sure the sagittal plane is oriented such that the centroid of lower and upper lip movements defined the vertical axis. This step can reduce the variation of rotational effects due to the difference in facial anatomy between speakers.

Thus in eq. (2), (c_y, c_z) are the centroid of shape S; Scaling factor (β_y, β_z) is set to the inverse of the square root of the sum of the squares of all data points along the dimension of shape S; θ is the angle of the S to the reference shape in which upper and lower lips form a vertical line. Fig. 3 shows an example of Procrustes matching of articulatory data from two healthy speakers when producing "*How are you doing*?". As can be seen, the original articulatory data in Fig. 3(a) are on different positions. After Procrustes matching as in Fig. 3(d), we can see that the articulatory shapes from two speakers are normalized by reducing locational, scaling, and rotational differences. The effect of each component of Procrustes matching will be discussed in detail in Section IV-A.

B. fMLLR: A Data-Driven Transform Approach

Feature-space maximum likelihood linear regression (fMLLR) is one of the representative data-driven approaches for feature space normalization in acoustic speech recognition [27]. In this method, a transformation matrix **A** is estimated[1] from input features and used for transforming the features onto general feature space as follows:

$$\hat{o}_t = \mathbf{A} o_t \quad (3)$$

where o_t is the input feature vector at frame t and is transformed to $\hat{o}_t$ using the transform matrix **A**. This transformed vector $\hat{o}_t$ is used for training and also for decoding.

In this work, the transformation matrix was estimated using articulatory movement data from an utterance of each speaker and applied to the articulatory data. An expectation-maximization (EM) algorithm [26] is normally used to estimate the transformation matrix. However, when the length of an utterance is short (e.g., less than 5 seconds), this does not always lead to improvements in speech recognition performance [43]. Alternatively, a basis representation of the transformation matrix can be used to robustly estimate the matrix when short duration utterance is available as follows:

$$\mathbf{A} = \sum_{n=1}^{N} d_n \mathbf{A}_n \quad (4)$$

where N is the basis size, $\mathbf{A}_n$ is the basis matrices, and d_n is speaker-specific coefficients. The optimal basis matrices and the optimal basis size can be found during training phase. Also, the speaker-specific coefficients are found for each utterance. Thus, we used basis

[1] The bias term is omitted for concise description in this paper.

representation-based fMLLR in this work and estimated the matrix in each utterance. A more detailed explanation of the basis representation of fMLLR can be found in [43].

C. i-vector: A Data-Driven Speaker Footprint Approach

The i-vector is a low-dimensional vector that effectively characterizes a speaker [44]. In recent years, i-vector approaches have become the state-of-the-art in the speaker verification field [44]. Also, i-vector approaches have been widely used in acoustic speech recognition to normalize the speaker effect. In this method, the i-vector is used as an additional input together with acoustic features to deep neural network-based acoustic model so as to adjust the model parameters to exploit the speaker information [27], [28]. In other words, the weights of neural networks are trained to represent the relationship between phonetic information (output) and both acoustic features and i-vectors (input). Therefore, the neural network can be robust to speaker variation.

Given GMM, which is adapted from GMM-based universal background model (UBM) trained on a large number of utterances and speakers, the corresponding mean super-vector M can be approximated by

$$M = m + \mathbf{T}w \quad (5)$$

where m is the mean super-vector taken from the GMM-UBM; **T** is a low-rank rectangular matrix spanning the subspace covering the total variability; w is a low-dimensional vector with a normally distributed prior $N(0, \mathrm{I})$. After iteratively estimating matrix **T** over training data, the lower dimensional vector w can be used as abstracted speaker information instead of large GMM and w is referred to as the i-vector. In this work, we extracted an i-vector from articulatory data of each utterance and the i-vector is combined with articulatory data as input to the silent speech recognition system.

D. Combination of Normalization Approaches

The articulatory normalization methods explained in previous sections show distinctive characteristics, and therefore, they may effectively complement each other. Actually, a combination of fMLLR and i-vector methods shows better performance than using each normalization method in acoustic speech recognition [28]. In silent speech recognition, however, the physiological variability of articulators between speakers is huge. Thus if the data-driven normalization methods (i.e., fMLLR) are directly applied to articulatory data, it may provide unsatisfactory performance.

To overcome this limitation, we first applied Procrustes matching to normalize the articulatory difference caused by speaker's anatomy. Then, the resulting articulatory data were transformed using an fMLLR transformation matrix to normalize the articulatory movement patterns during articulation in a data-driven way. Next, the i-vectors were extracted from fMLLR transformed articulatory data. fMLLR tries to remove the speaker variation, but practically still speaker-specific articulatory information remains due to the huge variation of articulatory movements across speakers. Therefore, it is expected that combining three different normalization approaches may be effective in achieving a better speaker-independent SSR system. Finally, the fMLLR-normalized articulatory data and i-vector were concatenated and fed into the BLSTM-based deep articulatory model described in the next section.

E. BLSTM-based Deep Articulatory Model

A long short term memory (LSTM) recurrent neural network is a type of a recurrent neural network, which has memory blocks containing a set of recurrently connected subnets [32]. In other words, an LSTM network is formed like a simple recurrent neural network (RNN) by replacing the nonlinear units in the hidden layers by memory blocks. It has been proven that LSTM can successfully address the problem of the vanishing gradients for RNN by using the memory blocks. Each memory block includes one or more self-connected memory cells and multiplicative units including input, output, and forget gates. These multiplicative gates provide continuous analogues of write, read, and reset operations for the cells, and therefore, it allows LSTM memory cells to store and access information over long periods of time.

A conventional LSTM network can make use of previous context with forward direction, which limits the prediction performance. Bidirectional LSTM (BLSTM) considers both the previous context with forward direction and the following context with backward direction by processing the data in both directions with two separate hidden layers, which are then fed forwards to the same output layer as in Fig. 4. Therefore, BLSTM can effectively access long-range context in both input directions. In this work, the fMLLR-normalized articulatory data and i-vectors were used as input units and context-dependent tied triphone states (i.e., senones) were used as output units as in Fig. 4. For hidden layers, we used two forward direction layers and two backward direction layers to capture the complex structure of articulatory movements. Therefore, it is expected that BLSTM is appropriate for silent speech recognition by jointly modeling articulatory movements in bidirectional temporal context.

F. Experimental Setup

The experimental setup is summarized in Table I. We used 24 dimensional EMA data consisting of 8 static data and their first and second order derivatives with shift size of 10 milliseconds as input features to silent speech recognition systems as baseline. For the experiments using baseline EMA alone, we used mean normalization along each dimension as a default setting. For articulatory normalization, Procrustes matching was first applied to the EMA data and then mean normalization along each dimension was used. Next, fMLLR transformation was performed to the Procrustes matched data on an utterance basis. A 20 dimensional i-vector[2] was extracted from each utterance using fMLLR-transformed EMA data to avoid overfitting [27]. For i-vector extraction, we used GMM-UBM with 2048 Gaussians and the total variability matrix is iteratively estimated using the Expectation Maximization algorithm on the training dataset using the Kaldi toolkit [46]. Finally, the i-vector was concatenated with fMLLR-transformed features as input.

[2] We have performed experiments on a variety number of dimensions of i-vectors and obtained the best performance on 20 dimensional i-vectors.

We compared following SSR systems[3] to evaluate the proposed method: *GMM-HMM, DNN-HMM, LSTM-HMM*, and *BLSTM-HMM*. We used hidden Markov model (HMM)-based silent speech recognition systems where each state can be modeled by GMM or neural network. It consists of 773 tied-state (senone) left-to-right triphone HMMs, where each HMM has 3 states. The senones were obtained using decision tree-based state tying methods [46]. GMM was trained using maximum likelihood estimation. DNN was trained using 24 dimensional EMA data with a context window of 9 frames as well as 20 dimensional i-vectors. The DNN has 3 hidden layers with 512 hidden units at each layer and the 773 dimensional softmax output layer, corresponding to the senones of the GMM-HMM system. The parameters were initialized using layer-by-layer generative pre-training based on restricted Boltzmann machines (RBM)[4] and the network was discriminatively trained using backpropagation [45]. LSTM was trained using 24 dimensional EMA data plus 20 dimensional i-vectors. The LSTM has 2 hidden layers with 640 LSTM cells at each layer and the 773 dimensional softmax output layer. The parameters were trained using backpropagation through time (BPTT). For BLSTM, we used 2 forward hidden layers with 320 LSTM cells and 2 backward hidden layers with 320 LSTM cells. Bigram phoneme language model was used for the phoneme sequence recognition system. For the word sequence recognition system, we used unigram and bigram word language model. The language model is trained using the text from training data. The training and decoding were performed using the Kaldi speech recognition toolkit [46].

Phoneme error rates (PERs) and word error rates (WERs) were used as the measure of silent speech recognition performance. Leave-two-subjects-out (6 fold) cross validation was used in the experiment for healthy speakers. In other words, the samples from ten subjects were used for training GMM-UBM and neural network while the samples from two subjects were used for testing for each cross validation. The average performance of cross validations was reported as the overall performance. For the test of laryngectomy patients in Section IV-D, all samples from two laryngectomy patients were used for testing and the samples from the twelve healthy speakers were used for training.

IV. Results and Discussion

A. Effectiveness of Procrustes Matching

We first investigated the effectiveness of Procrustes matching on speaker-independent (SI) GMM for healthy speakers in Table II. Here, we applied Procrustes matching-based articulatory normalization only to check the effectiveness of Procrustes matching itself in speaker-independent silent speech recognition. Note that “baseline” is the original EMA data before normalization. As shown in Table II, rotation was more effective than translation and scaling for the comparison of each component. When we applied translation and rotation together, we were able to obtain the best performance for triphone systems. Interestingly, scaling was not appropriate for the silent speech recognition. It is found that scaling may

[3] We tried to find the best configuration on each SSR system and set the configuration in Table I.

[4] We have conducted experiments with/without RBM pretraining on SI-DNN using our database. DNN without pretraining produced PERs of 54.1% and 52.6% for Procrustes matching and Procrustes matching+fMLLR, respectively. DNN with pretraining gave 52.3% and 51.7%, respectively, which show better results. Thus, we used RBM pretraining in this work.

reduce the discriminability between phonemes. It is also observed that the triphone system is better than the monophone system as in acoustic speech recognition. Therefore, we used translation and rotation as default Procrustes matching in the remainder of the experiments. Also, we performed experiments on triphone context systems in the following experiments.

B. Effectiveness of the Proposed Method

Fig. 5 shows the performances of articulatory normalization methods on speaker-independent (SI) GMM, DNN, LSTM, and BLSTM based silent speech recognition systems for healthy speakers. Speaker-dependent (SD) GMMs are also presented to compare with speaker-independent systems with articulatory normalization. For SD GMM, 4-fold cross validation for each speaker was performed. Note that “baseline” is the original EMA data before normalization. The performances of speaker-independent systems with articulatory normalization were comparable or even better than with the GMM-based speaker-dependent system. These results suggest that all the articulatory normalization methods are effective for speaker-independent silent speech recognition. Specifically, Procrustes matching outperformed fMLLR and i-vector methods for all the articulatory models. When we combined the fMLLR and i-vector with Procrustes matching, the performance was improved. Although fMLLR and i-vector methods even degrade the performance compared to baseline on the LSTM and BLSTM systems, the performance was improved by using Procrustes matching together with fMLLR and i-vector. Finally, when we used all the normalization methods together, we were able to obtain the best performance on all the articulatory models. This indicates that Procrustes matching (or physiological normalization) is critical for speaker-independent silent speech recognition. Also, these articulatory normalization approaches complement each other.

For the articulatory models, the performances of deep articulatory models, such as DNN, LSTM, and BLSTM, were better than with GMM-based shallow generative models. This indicates that the complex structure of articulatory movements can be properly modeled on the deep models. Also, LSTM provided better recognition performance than DNN for all the corresponding normalization conditions. In addition, BLSTM was always better than one directional LSTM. Finally, we obtained the best results (40.3% in PER) on BLSTM with all the normalization methods, showing 14.4% and 20.0% relative improvements in the PER reduction over LSTM and DNN with all the normalization methods, respectively. Also, BLSTM with all the normalization methods gave 19.2% relative PER reduction over baseline BLSTM.

To show the performance of the proposed method for each speaker, PERs of the selected models with/without normalization methods for each individual are presented in Table III. For articulatory normalization, we used three normalization methods together. As can be seen, DNN with articulatory normalization gave better results than speaker-dependent baseline GMM and speaker-independent baseline DNN (without normalization) for almost all the speakers. Normalized LSTM further improved the recognition performance for all the speakers. Finally, we obtained the best performance on BLSTM with articulatory normalization for all the speakers.

Table IV shows the WERs of the selected models with/without normalization methods using word-level unigram and bigram language models. As shown in Table IV, DNN with Procrustes matching provided better results than with baseline DNN. When we used other normalization methods together with Procrustes matching (Proc.+fMLLR and Proc.+fMLLR +ivector), the performance was further improved on DNN as in the experiments using PERs. LSTM with three normalization methods was slightly better than normalized DNN. We also obtained the lowest WERs on BLSTM with three normalization methods for both unigram and bigram language models.

C. Analysis Using Phoneme Confusion Matrix

Next, we analyzed the phoneme confusion matrices resulting from baseline DNN, normalized DNN, normalized LSTM, and normalized BLSTM to find the cause of the performance improvement in Fig. 6. Here, we used the three normalization methods together. To this end, actual phoneme sequences and predicted phoneme sequences were aligned using the Levenshtein edit distance and then the number of the aligned phoneme pairs was calculated. The numbers of deleted and inserted phonemes were also computed. The matrices show relative accuracies in percent, i.e., the sum of each row is 100%, as in [50].

In the phoneme confusion matrix from baseline DNN in Fig. 6(a), there are various phoneme confusions against target phonemes including unnatural phonetic substitutions, e.g., consonants to vowels (left-bottom part in Fig. 6(a)) and vowels to consonants (right-top part), and a large number of deletions. This may be due to the huge variation of articulatory movements between speakers. In other words, the articulatory variability produces high confusability between a lot of phoneme classes. When we applied our normalization approaches on DNN in Fig. 6(b), on the other hand, phoneme confusions were reduced over all the target phonemes. Interestingly, deletion errors were much reduced while understandable substitutions, which are within-manner and within-place phoneme substitutions, were still challenging. Specifically, the bilabial consonants /b/ sound was confused with the same place group of sounds such as /p/ and /m/ sounds. Also, voiced and voiceless pairs, such as the /f/ and /v/ sounds and the /th/ and /dh/ sounds, were confusable each other.

For the confusion matrix from LSTM with articulatory normalization in Fig. 6(c), unnatural substitutions in the left-bottom and right-top parts in the matrix were much reduced. Also, some understandable substitutions were reduced. In particular, within-place substitutions for the bilabial sounds (e.g., the /p/, /b/, and /m/ sounds) and alveolar sounds (e.g., the /t/, /d/, / n/, /s/, and /z/ sounds) were much reduced. Finally, when we used BLSTM with articulatory normalization in Fig. 6(d), the confusion matrix shows a clearer pattern of correct recognition and a few confusions. In particular, the confusions between velar sounds, such as the /k/, /g/, and /ng/ sounds, were further reduced. Throughout the confusion matrices, the /ao/ and /zh/ sounds were always misrecognized. We found that there were only a few samples for those sounds and therefore it might be hard to train those sounds. Nonetheless, the most confusable sound was associated with same phonological attribute, e.g., both the /ao/ sound and the /aa/ sound are associated with back vowels and both the /zh/ sound and the /z/ sound are voiced alveolar sounds. From these analyses, we can conclude that the proposed method (BLSTM with articulatory normalization) is effective in distinguishing phonemes, even for phonemes that are in the same category of the manner and place of articulation.

Fig. 7 shows an example of alignment between actual and predicted phoneme sequences when producing "*I have a speech problem*" by a healthy speaker (SPK9). Here, we compared the results from baseline DNN, normalized DNN, normalized LSTM, and normalized BLSTM. As can be seen, baseline DNN produced three deletions and four substitutions including unnatural substitutions (e.g., /b/ to /ah/). We found that when substitutions occur with deletions or insertions successively, it is likely to incur unnatural substitutions. Since the standard Levenshtein distance takes all the phoneme errors as uniform weights, the aligned results may be unnatural[5]. Similarly, when there are successive deletions and substitutions for normalized DNN, we observed unnatural substitutions (e.g., /ah/ to /p/). By contrast, we obtained a quite reasonable substitution error (e.g., /p/ to /m/) for most isolated substitutions. For normalized LSTM, all the consonants were correct but one vowel substitution was produced. When we used normalized BLSTM, all the phoneme alignments were correct. This analysis implies that unnatural substitutions may be observed when there are successive phoneme errors and it can be reduced by using our proposed method.

D. Evaluation of the Proposed Method on Laryngectomy Patients: A Case Study

Finally, the proposed speaker-independent silent speech recognition method was evaluated on the articulatory data collected from two laryngectomy patients in Fig. 8. We used the data from all the twelve normal speakers to train the articulatory model. For speaker-dependent (SD) GMM, 4-fold cross validation for each speaker was performed. As shown in Fig. 8, this approach obtained the similar results on recognition performance improvement as on healthy speakers, which have been shown in Fig. 5. Specifically, Procrustes matching produced lower PERs than fMLLR and i-vector methods on LSTM and BLSTM systems. Also, the performances were improved by combining the fMLLR and i-vector with Procrustes matching. The best performance was obtained when all the normalization methods were used together with the BLSTM-based SSR system, achieving a PER of 56.0%.

Table V shows the WERs of the selected models with/without normalization methods using word-level unigram and bigram language models. As shown in Table V, DNN with Procrustes matching provided slightly worse results than with baseline DNN. When we used other normalization methods together with Procrustes matching (Proc.+fMLLR and Proc.+fMLLR+ivector), the performance was better than with baseline DNN as in the experiments using PERs. LSTM with three normalization methods was slightly better than normalized DNN. We also obtained the lowest WERs on BLSTM with three normalization methods for both unigram and bigram language models.

[5] If we consider the relationship between phonemes using weighted finite state transducer [54], which imposes different weights to phoneme mapping, we can obtain more reasonable phoneme alignment results even when we have a lot of phoneme errors.

Experimental results (Fig. 8 and Table V) showed the potential of our proposed speaker-independent SSR approach (articulatory normalization + BLSTM) on laryngectomy patients as well as healthy speakers. These findings also indicate that the patients (after a limited time of adaptation to the laryngeal removal surgery) might keep the similar articulatory patterns. To our knowledge, this is the first project with phoneme/word-level SSR evaluation using laryngectomees. However, further analyses with more laryngectomy patients are needed to verify these findings.

E. Discussion

The experimental results show promising results (small variation across speakers and improved recognition performance), but the number of participants in this study was relatively small. It is unknown if the performance can be generalized to a larger number of participants. Fig. 9 shows PERs on healthy speakers in our database according to the number of training speakers. Our default number of training speakers in Section IV-B was 10 for each cross validation. It is observed that the performance is improved as the number of training speakers increases for both the SI baseline and normalized systems. Specifically, as the number of training speakers increases, normalization approaches on DNN and BLSTM are more effective than baseline DNN. Although this analysis is based on our database with twelve healthy speakers, we found the normalization approaches may be promising when articulatory data from a larger number of speakers are available. We have plans to actively collect additional data including healthy speakers as well as laryngectomy patients. Therefore, future work will include more analysis on the dataset with a larger number of participants including the publicly available XRMB dataset [52].

Although our algorithms for speaker-independent SSR are promising, EMA is still cumbersome and cannot be used in daily life. However, significant advances have been made in the hardware technology. We anticipate the machine will be smaller, portable, and even wireless in the next decades [51]. The focus of this paper is to develop software algorithms for SSR. These algorithms can be directly applied to flesh point articulatory data and adapted for other articulatory data modalities that are generated by next generation of tongue tracking devices.

In our experiments, LSTM/BLSTM always produced better performance than DNN. This may be due to the ability of modeling temporal context in LSTM/BLSTM although DNN can also model the temporal context using a context window of several adjacent frames. It is an open question how much language structure the LSTM units really learn on our silent speech database. Future work also includes the analysis on a larger dataset with more diverse language structure.

We used LSTM/BLSTM with HMM in this work. However, many researchers have recently tried to drop the HMM using connectionist temporal classification (CTC) training criterion [53] in an end-to-end acoustic speech recognition system. Our articulatory normalization approaches are basically data variation reduction approaches, and therefore, our methods may be still effective on CTC criterion-based neural network models. However, a further study is required to verify this issue.

We have included word error rates (WER) on our database. Although our database has a relatively small number of phrases and a small vocabulary, it is designed using most widely used phrases in daily life for AAC users. Also, our final research goal is to make an SSR algorithm for clinical applications such as an assistive communication device for laryngectomy patients in the future. The small phrases may not be enough in normal speech recognition setting, but it may be still useful in clinical applications.

We included a case study on two laryngectomy patients as an extra analysis (Section IV-D) to demonstrate the gap of the speaker independent SSR performance on healthy speakers and the target patients. The results from two patients show a worse performance than healthy speakers[6]. Nevertheless, we verified the feasibility of our normalization approaches compared to speaker-dependent models. Future direction to improve the performance on laryngectomy patients could use a dataset with a larger number of laryngectomy patients and use part of the patient's data for training. The articulatory pattern of healthy control and laryngectomy patients remains unknown, but we anticipate this is more effective for SSR on the patients. A better understanding of the articulatory pattern of healthy control and laryngectomy patients would also be helpful for model tuning.

V. Conclusions

In this paper, an effective method to automatically recognize speech information from flesh point articulatory movements (without using audio information) was proposed. The method relies on two important parts: 1) articulatory normalization method using a combination of Procrustes matching, fMLLR, and i-vector for addressing the articulatory variation resulting from speaker's anatomy and articulation patterns, and 2) BLSTM-based deep neural network for modeling articulatory movements in temporal context. A series of experiments were performed on twelve healthy participants to evaluate the effectiveness of the proposed method in terms of PER and WER. Also, we additionally performed experiments on two laryngectomy patients. The experimental results showed the effectiveness in the aspects of 1) the performance through comparison with other SSR systems with articulatory normalization methods, achieving significant performance improvements, 2) adaptability for online recognition, achieving utterance-based articulatory normalization, and 3) the generality of the proposed approach, showing better SSR results for laryngectomy patients as well. Thus, our framework presents a possibility of speaker-independent silent speech recognition for healthy as well as laryngectomy speakers. Our method may also be successfully applied to speech recognition in noisy environment as an alternative technology.

Acknowledgments

This paper was received on Dec. 15, 2016. This work was supported by the National Institutes of Health (No. R03DC013990) and by the American Speech-Language-Hearing Foundation through a New Century Scholar Grant.

[6] The performance of laryngectomy patients is notably worse on average, but is similar to the performance on some of the healthy subjects (e.g., SPK4 or SPK12). Perhaps more laryngectomy patients may produce better results on average. However, a further study is needed to verify it using a larger dataset.

References

1. Bailey BJ, Johnson JT, Newlands SD. Head and Neck Surgery – Otolaryngology. 4. Lippincot, Willians & Wilkins; Philadelphia, PA, USA: 2006.
2. Mau T, Muhlestein J, Callahan S, Chan RW. Modulating phonation through alteration of vocal fold medial surface contour. The Laryngoscope. 2012; 122(9):2005–2014. [PubMed: 22865592]
3. Mau T. Diagnostic evaluation and management of hoarseness. Medical Clinics of North America. 2010; 94(5):945–960. [PubMed: 20736105]
4. Liu H, Ng ML. Electrolarynx in voice rehabilitation. Auris Nasus Larynx. Sep.2007 34(3):327–332. [PubMed: 17239553]
5. Khan JA, Green P, Creer S, Cunningham S. Reconstructing the voice of an individual following laryngectomy. Augment. Altern. Commun. Mar.2011 27(1):61–66. [PubMed: 21284563]
6. Denby B, Schultz T, Honda K, Hueber T, Gilbert JM, Brumberg JS. Silent speech interfaces. Speech Commun. 2010; 52:270–287.
7. Wang J, Samal A, Green JR, Rudzicz F. Whole-word recognition from articulatory movements for silent speech interfaces; Proc. Annual Conference of the International Speech Communication Association; Sep.. 2012 1327–1330.
8. Fan Y, Qian Y, Xie F, Soong FK. TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks; Proc. Annual Conference of the International Speech Communication Association; Sep.. 2014 1964–1968.
9. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. 2016
10. 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 Commun. 2010; 52:288–300.
11. Denby B, Cai J, Roussel P, Dreyfus G, Crevier-Buchman L, Pillot-Loiseau C, Hueber T, Chollet G. Proc. Int. Cong. Phonet. Sci. Hong Kong, China: 2011. Tests of an interactive, phrasebook-style post-laryngectomy voice-replacement system; 572–575.
12. Jorgensen C, Dusan S. Speech interfaces based upon surface electromyography. Speech Commun. 2010; 52:354–366.
13. Deng Y, Heaton J, Meltzner G. Proc. Annual Conference of the International Speech Communication Association. Singapore: 2014. Towards a practical silent speech recognition system; 1164–1168.
14. Fagan MJ, Ell SR, Gilbert JM, Sarrazin E, Chapman PM. Development of a (silent) speech recognition system for patients following laryngectomy. Med. Eng. Phys. 2008; 30(4):419–425. [PubMed: 17600751]
15. Wang J, Samal A, Green JR. Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph; Proc. ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies; 2014.
16. Kent RD, Adams SG, Tuner GS. Models of Speech Production. In: Lass NG, editorPrinciples of Experimental Phonetics. Mosby: 1996. 3–45.
17. Wang J, Green J, Samal A, Yunusova Y. Articulatory distinctiveness of vowels and consonants: A data-driven approach. J Speech. Language, and Hearing Research. 2013; 56(5):1539–1551.
18. Johnson K, Ladefoged P, Lindau M. Individual differences in vowel production. J Acoust. Soc. Am. Aug.1993 94(2):701–714. [PubMed: 8370875]
19. Hashi M, Westbury JR, Honda K. Vowel posture normalization. J Acoust. Soc. Am. Oct.1998 104(4):2426–2437. [PubMed: 10491704]
20. Simpson AP. Gender-specific differences in the articulatory and acoustic realization of interword vowel sequences in American English. 5th Seminar on Speech Production: Models and Data. Kloster Seeon. 2000:209–212.
21. Westbury JR, Hashi M, Lindstrom MJ. Differences among speakers in lingual articulation for American English /ɹ/. Speech Commun. 1998; 26(3):203–226.

22. Li S, Wang L. Cross linguistic comparison of Mandarin and English EMA articulatory data; Proc. Annual Conference of the International Speech Communication Association; Sep.. 2012 903–906.

23. Felps D, Gutierrez-Osuna R. Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis. Department of Computer Science and Engineering, Texas A&M University, Tech. Rep.; 2010. TAMU-CS-TR-2010-5-3

24. Dryden IL, Mardia KV. Statistical Shape Analysis. John Wiley & Sons; New York: 1998.

25. Wang J, Samal A, Green JR. Across-speaker articulatory normalization for speaker-independent silent speech recognition; Proc. Annual Conference of the International Speech Communication Association; Sep.. 2014 1179–1183.

26. Gales M. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 1998; 12(2):75–98.

27. Senior A, Lopez-Moreno I. Improving DNN speaker independence with i-vector inputs; Proc. IEEE Int. Conf. Acoust. Speech, and Signal Process; Mar.. 2014 225–229.

28. Rouvier M, Favre B. Speaker adaptation of DNN-based ASR with i-vectors: Does it actually adapt models to speakers?; Proc. Annual Conference of the International Speech Communication Association; Sep.. 2014 14–18.

29. Mohamed A, Dahl GE, Hinton G. Acoustic modeling using deep belief networks. IEEE Trans. Audio, Speech, Lang. Process. Jan.2012 20(1):14–22.

30. Dahl GE, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech Lang. Process. Jan.2012 20(1)

31. Hahm S, Wang J. Silent speech recognition from articulatory movements using deep neural network. Proc. Int. Congress Phonetic Sci. 2015:1–5.

32. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997; 9(8):1735–1780. [PubMed: 9377276]

33. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks. Jun; 2005 18(5–6):602–610. [PubMed: 16112549]

34. Wollmer M, Schuller B, Eyben F, Rigoll G. Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. J Selected Topics in Signal Processing. Oct.2010 4:867–881.

35. Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM; Proc. IEEE Workshop on Automatic Speech Recognition and Understanding; 2013. 273–278.

36. Fan Y, Qian Y, Xie F, Soong FK. TTS synthesis with bidirectional LSTM based recurrent neural networks; Proc. Annual Conference of the International Speech Communication Association; 2014. 1964–1968.

37. Wang J, Hahm S. Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training; Proc. Annual Conference of the International Speech Communication Association; Sep.. 2015 2415–2419.

38. Green J, Wang J, Wilson DL. Smash: A tool for articulatory data processing and analysis. Proc. Interspeech. Sep.2013 :1331–1335.

39. Berry J. Accuracy of the NDI wave speech research system. J Speech. Language, and Hearing Research. 2011; 54(5):1295–1301.

40. Wang J, Samal A, Rong P, Green JR. An optimal set of flesh points on tongue and lips for speech-movement classification. J Speech. Language, and Hearing Research. 2016; 59(1):15–26.

41. Wang L, Ning H, Hu W, Tan T. Gait recognition based on Procrustes shape analysis; Proc. Int. Conf. Image Process; Sep.. 2002 III-433–436.

42. Jin N, Mokhtarian F. Human motion recognition based on statistical shape analysis; Proc. Int. Conf. Adv. Video and Signal Based Surveillance; 2005. 4–9.

43. Povey D, Yao K. A basis representation of constrained MLLR transforms for robust adaptation. Comput. Speech Lang. 2012; 26(1):35–51.

44. Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech, Lang., Process. May; 2011 19(4):788–798.

45. Hinton G, Osindero S, The YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006; 18:1527–1554. [PubMed: 16764513]

46. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely K. The Kaldi speech recognition toolkit; Proc. IEEE Workshop on Automatic Speech Recognition and Understanding; 2011. 1–4.

47. Hofe R, Ell SR, Fagan MJ, Gilbert JM, Green PD, Moore RK, Rybchenko SI. Small-vocabulrary speech recognition using a silent speech interface based on magnetic sensing. Speech Commun. 2013; 55:22–32.

48. Wang J, Green JR, Samal A. Individual articulator's contribution to phoneme production; Proc. IEEE Int. Conf. Acoust. Speech, and Signal Process; Mar.. 2013 7785–7789.

49. Wang J, Hahm S, Mau T. Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition; Proc. ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies; 2015. 79–85.

50. Wand M, Schultz T. Analysis of phone confusion in EMG-based speech recognition; Proc. IEEE Int. Conf. Acoust. Speech, and Signal Process; Mar.. 2011 757–760.

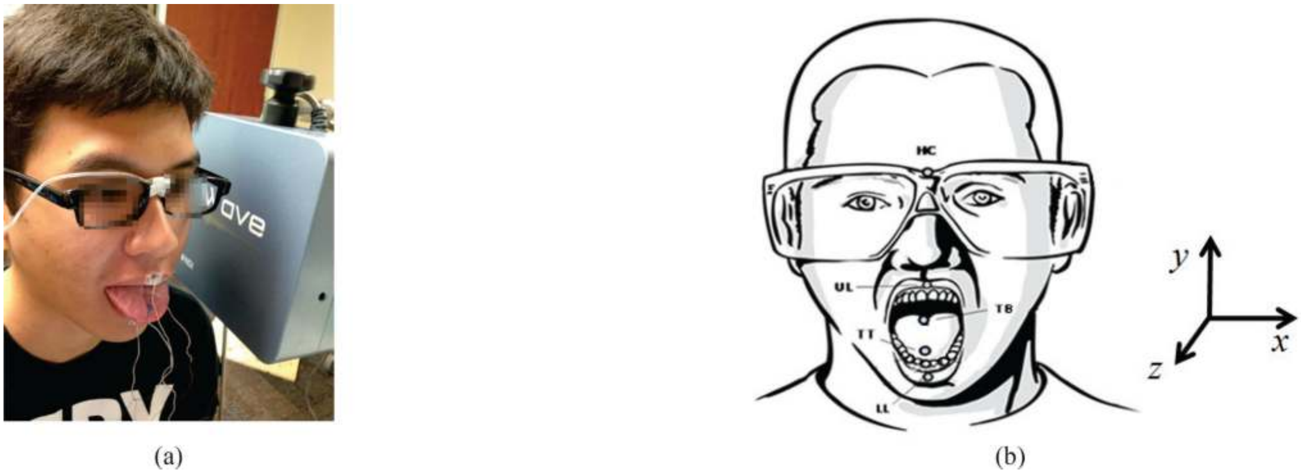
51. Huo X, Ghovanloo M. Tongue drive: A wireless tongue-operated means for people with severe disabilities to communicate their intentions. IEEE Communications Magazine. Oct.2012 50(10): 128–135.

52. Westbury JR. X-Ray Microbeam Speech Production Database User's Handbook Version 1.0. Waisman Center on Mental Retardation & Human Development, University of Wisconsin; Madison, WI: Jun.. 1994

53. Graves A, Fernandez S, Gomez F, Schmidhuber J. Proc. the 23rd International Conference on Machine Learning. Pittsburgh, PA: 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks.

54. Kim MJ, Kim Y, Kim H. Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model. IEEE/ACM Trans. Audio, Speech Lang. Process. Apr.2015 23(4):694–704.

55. Cao B, Kim M, van Santen J, Mau T, Wang J. Integrating articulatory information in deep learning-based text-to-speech synthesis; Proceedings of Annual Conference of the International Speech Communication Association; Aug.. 2017



(a)

(b)

Fig. 1.

Data collection setup: (a) the Wave system with a subject and (b) sensor labels and locations. HC: Head Center, TT: Tongue Tip, TB: Tongue Body Back, UL: Upper Lip, and LL: Lower Lip. Direction *x* is left-right, *y* is vertical, and *z* is front-back.

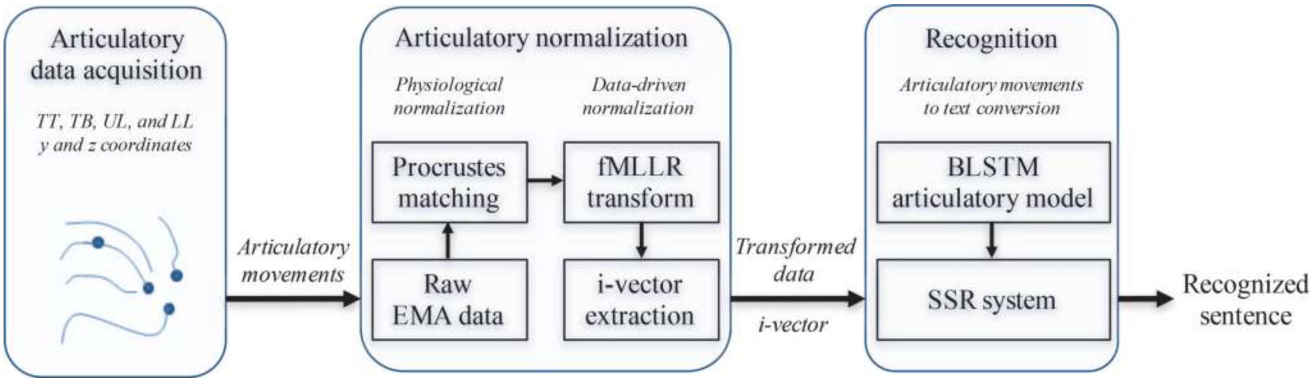

Articulatory data acquisition
TT, TB, UL, and LL y and z coordinates
Articulatory movements
Articulatory normalization
Physiological normalization
Data-driven normalization
Procrustes matching
fMLLR transform
Raw EMA data
i-vector extraction
Transformed data
i-vector
Recognition
Articulatory movements to text conversion
BLSTM articulatory model
SSR system
Recognized sentence

Fig. 2.

Proposed speaker-independent silent speech recognition framework with articulatory normalization and bidirectional long short term memory (BLSTM)-based articulatory model.

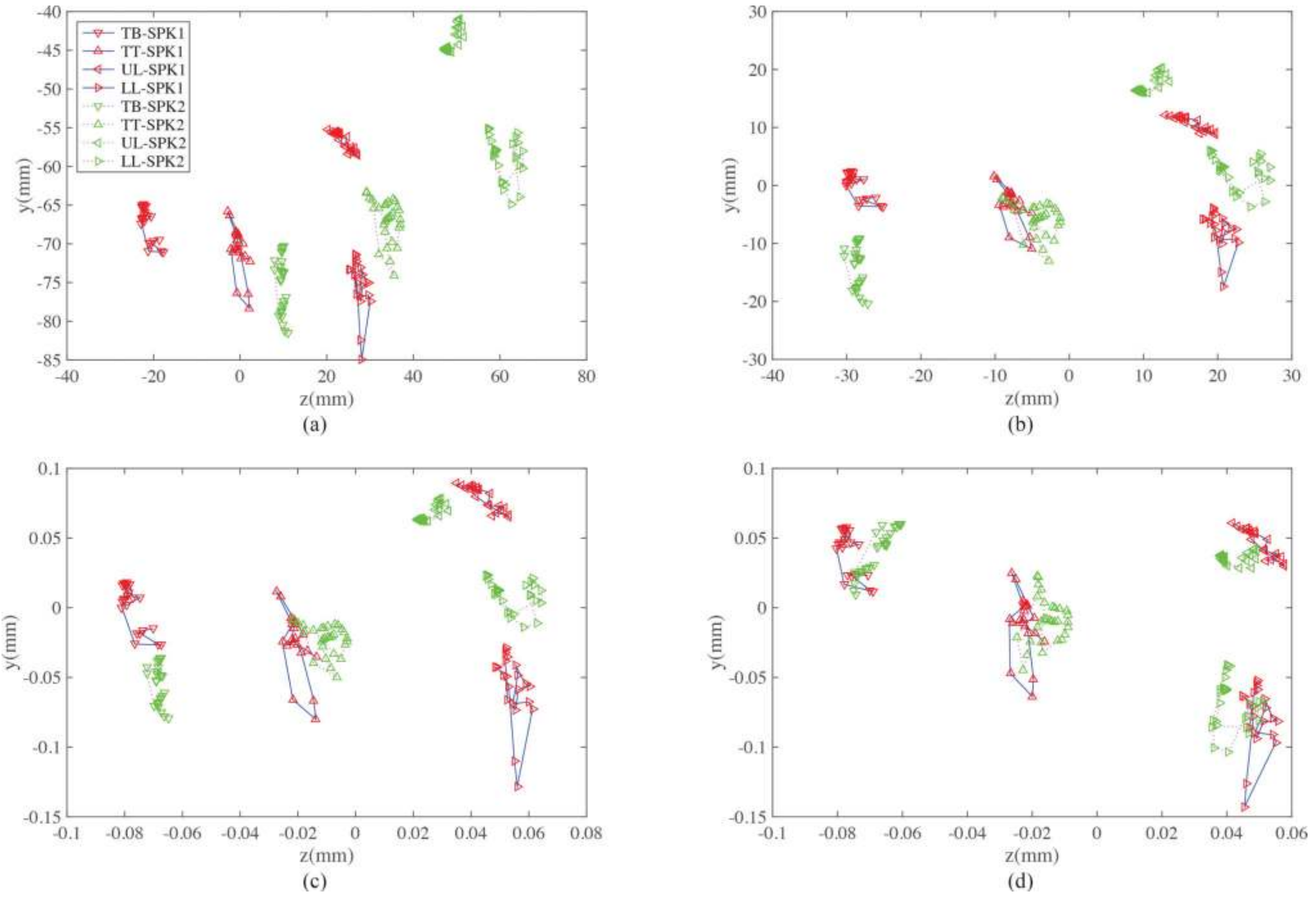


Fig. 3.

Example of Procrustes matching of motion path of four articulators (TT, TB, UL, and LL) for producing "*How are you doing*?" from two healthy speakers (a) original data, (b) translation, (c) translation+scaling, and (d) translation+scaling+rotation. Note that each curve is down-sampled to 22 points for visualization.

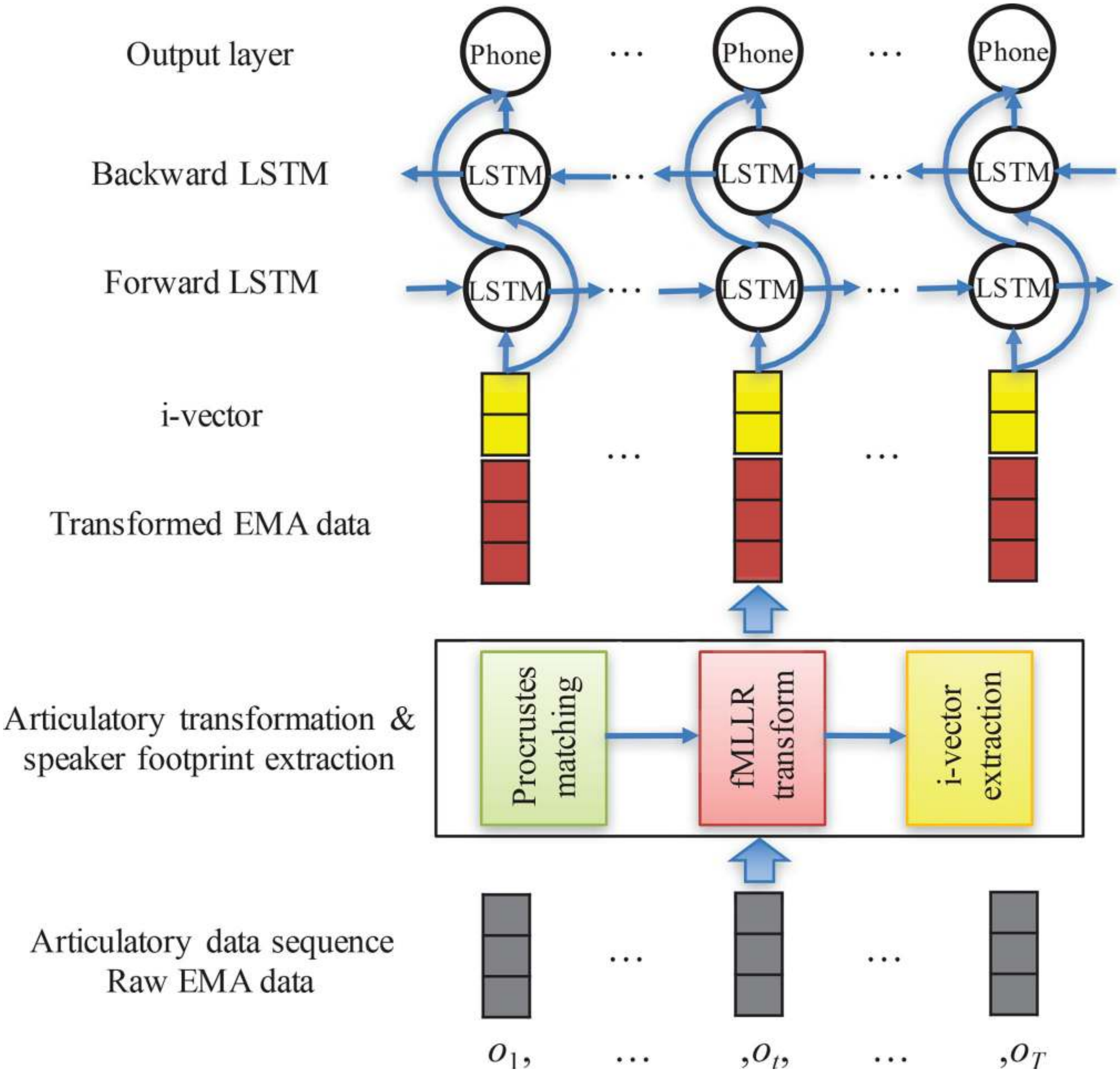


Fig. 4.
BLSTM-based SSR system with articulatory normalization.

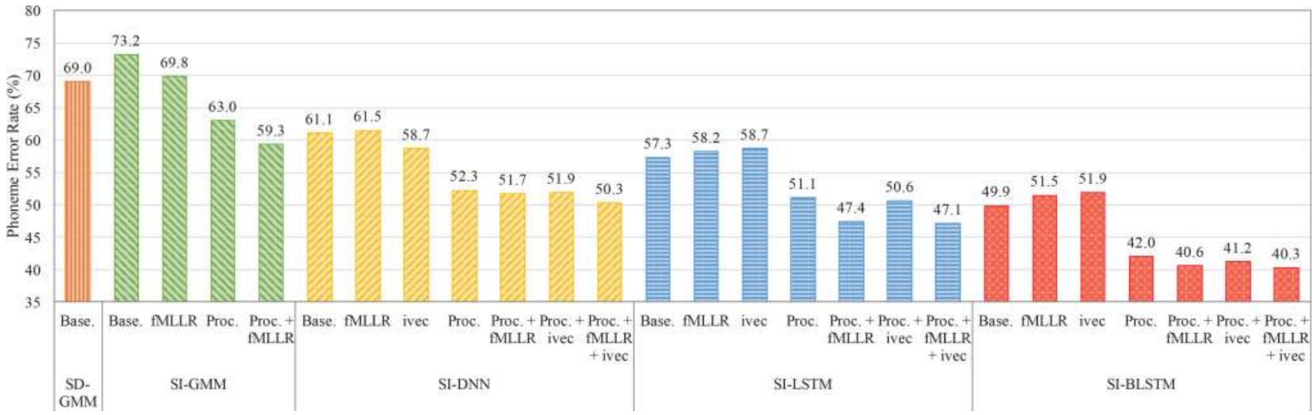
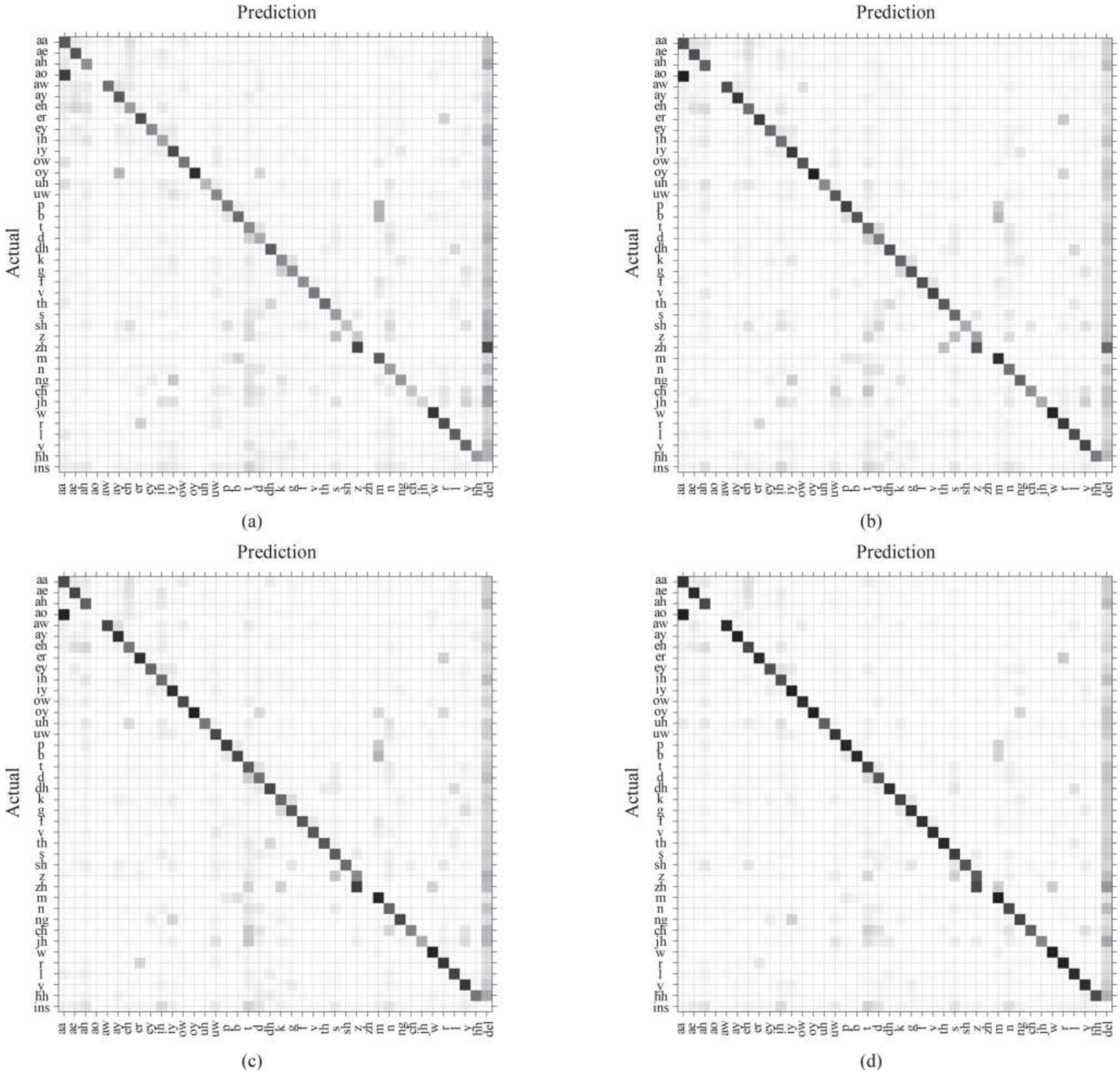


Fig. 5.

PERs of articulatory normalization methods on speaker-independent (SI) GMM, DNN, LSTM, and BLSTM based silent speech recognition systems. “Base.” means baseline, “Proc.” indicates Procrustes matching, and “ivec” is the i-vector approach.



(a)

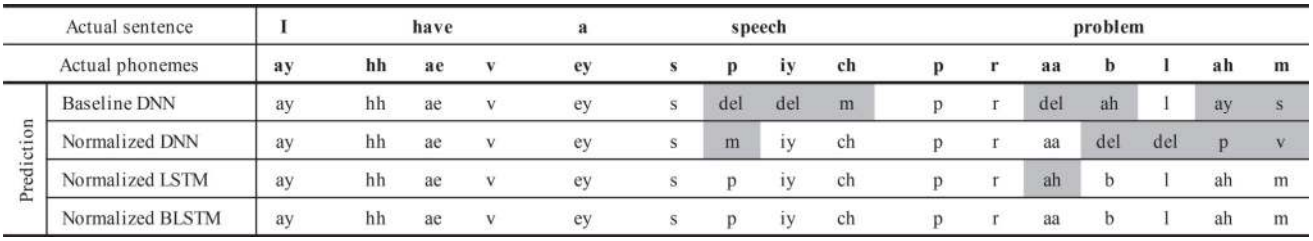
(b)

(c)

(d)

Fig. 6.

Phoneme confusion matrices from (a) baseline DNN, (b) normalized DNN, (c) normalized LSTM, and (d) normalized BLSTM. The bottom-most row and the right-most column indicate insertions and deletions, respectively.



	Actual sentence	I	have			a	speech				problem						
	Actual phonemes	**ay**	**hh**	**ae**	**v**	**ey**	**s**	**p**	**iy**	**ch**	**p**	**r**	**aa**	**b**	**l**	**ah**	**m**
Prediction	Baseline DNN	ay	hh	ae	v	ey	s	del	del	m	p	r	del	ah	l	ay	s
	Normalized DNN	ay	hh	ae	v	ey	s	m	iy	ch	p	r	aa	del	del	p	v
	Normalized LSTM	ay	hh	ae	v	ey	s	p	iy	ch	p	r	ah	b	l	ah	m
	Normalized BLSTM	ay	hh	ae	v	ey	s	p	iy	ch	p	r	aa	b	l	ah	m

Fig. 7.

Alignment example between actual and predicted phoneme sequences from baseline DNN, normalized DNN, normalized LSTM, and normalized BLSTM-based SSR systems when producing "*I have a speech problem*" from a healthy speaker. Phoneme errors are shaded.

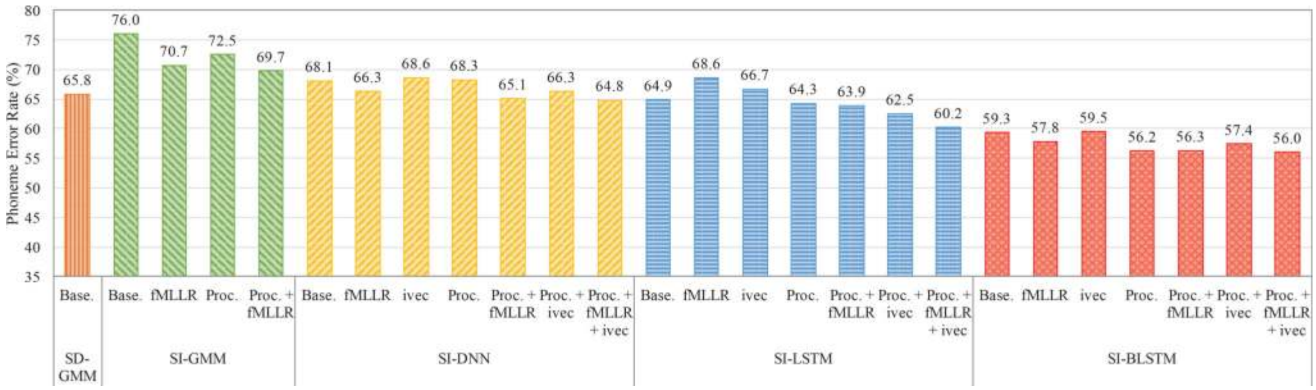


Fig. 8.

PERs of articulatory normalization methods on speaker-independent (SI) GMM, DNN, LSTM, and BLSTM based silent speech recognition systems for laryngectomy patients.

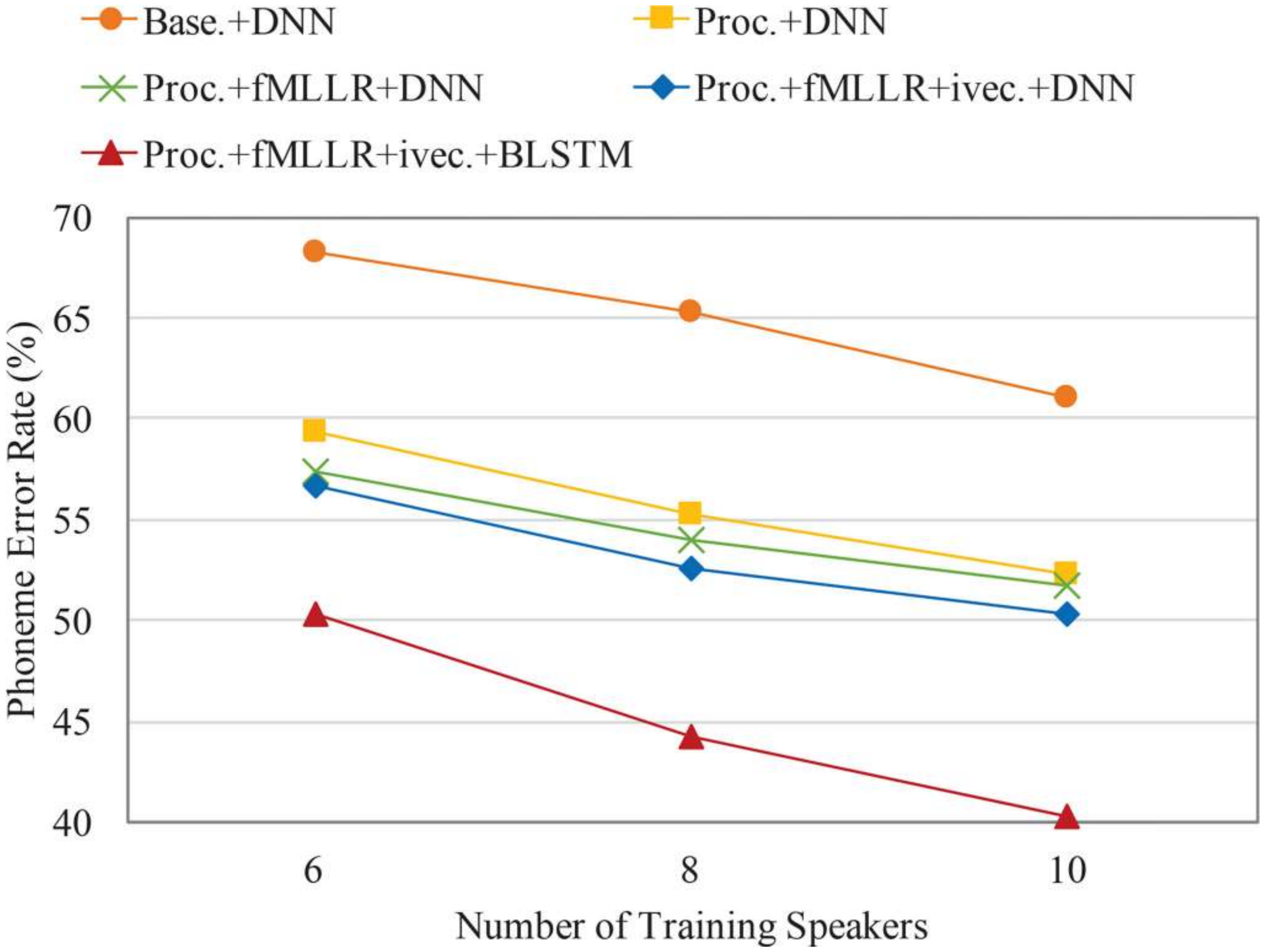


Fig. 9.

PERs according to the number of training speakers using healthy speakers in our database.

TABLE I

Experimental Setup

Articulatory data	
Feature vector	Articulatory movement data (8 dim.) + $\Delta + \Delta\Delta$ (24 dim.)
Low pass filtering	20 Hz cutoff 5th order Butterworth
Sampling rate	100 Hz
Mean normalization	Applied
GMM-HMM topology	
Monophone	122 states (39 phones × 3 states + 5 states for silence), total 1000 Gaussians
Triphone	773 states, total 7000 Gaussians
HMM	3 states left-to-right HMM
Training method	Maximum likelihood estimation
DNN-HMM topology	
Input layer dim.	216 (24 dim. × 9 context windows) + 20 dim. i-vector
Output layer dim.	773 for triphone
No. of hidden units	512 units for each hidden layer
Depth	3 hidden layers
Training method	RBM pre-training, back propagation
(B)LSTM-HMM topology	
Input layer dim.	24 dim. + 20 dim. i-vector
Output layer dim.	773 for triphone
No. of LSTM cell units for each hidden layer	640 units for LSTM 320 forward & 320 backward units for BLSTM
Depth	2 forward hidden layers for LSTM 2 forward & 2 backward layers for BLSTM
Training method	Back propagation through time (BPTT)
Language model	Bigram phoneme language model Unigram word language model Bigram word language model

TABLE II

PERs (%) of Procrustes Matching on SI-GMM for Healthy Speakers

Procrustes matching	Monophone	Triphone
Baseline	76.6	73.2
Translation	92.3	90.5
Rotation	76.4	71.9
Scaling	89.8	89.5
Translation+Rotation	**69.9**	**63.0**
Translation+Scaling	88.0	87.3
Rotation+Scaling	90.4	88.9
Translation+Scaling+Rotation	85.5	84.9

TABLE III

PERs (%) of Baseline DNN, Articulatory Normalized DNN, LSTM, and BLSTM for Each Individual. “Norm.” Means a Combination of Procrustes Matching, fMLLR, and i-vector Methods

Speaker ID	**SD**	**SI**			
	Base.+ GMM	**Base.+ DNN**	**Norm.+ DNN**	**Norm.+ LSTM**	**Norm.+ BLSTM**
SPK1	69.1	57.5	44.8	42.2	**36.1**
SPK2	64.4	78.6	51.4	45.5	**42.9**
SPK3	69.7	43.8	32.7	31.8	**29.1**
SPK4	73.6	69	68.2	64.7	**59.2**
SPK5	66.9	55.7	45.6	42.6	**32.5**
SPK6	74.5	61.6	43.9	41.7	**32.2**
SPK7	73.4	65.9	68.8	67.8	**59.7**
SPK8	67.3	54	45.1	41.6	**33.6**
SPK9	73.1	64.7	43.3	42.8	**30.2**
SPK10	63.6	55.1	47.8	45.4	**34.9**
SPK11	63.7	57.2	48.7	46.1	**38.4**
SPK12	68.6	69.5	61.3	58.9	**54.3**
Mean	69.0	61.0	50.1	47.5	40.2
STD	3.9	9.1	10.7	10.6	11.2

TABLE IV

WERs (%) of SI Baseline DNN, Articulatory Normalized DNN, LSTM, and BLSTM for Unigram and Bigram Language Models

Method	**Unigram**	**Bigram**
Baseline+DNN	75.5	52.4
Normalization (Proc.)+DNN	60.5	35.5
Normalization (Proc.+fMLLR)+DNN	59.1	33.1
Normalization (Proc.+fMLLR+ivec.)+DNN	56.7	32.4
Normalization (Proc.+fMLLR+ivec.)+LSTM	56.5	31.7
Normalization (Proc.+fMLLR+ivec.)+BLSTM	**42.7**	**23.6**

TABLE V

WERs (%) of SI Baseline DNN, Articulatory Normalized DNN, LSTM, and BLSTM Using Unigram and Bigram Language Models for Laryngectomy Patients

Method	Unigram	Bigram
Baseline+DNN	90.6	77.3
Normalization (Proc.)+DNN	91.6	80.1
Normalization (Proc.+fMLLR)+DNN	90.4	75.0
Normalization (Proc.+fMLLR+ivec.)+DNN	90.1	74.4
Normalization (Proc.+fMLLR+ivec.)+LSTM	77.5	64.9
Normalization (Proc.+fMLLR+ivec.)+BLSTM	**70.5**	**55.2**