Recognizing Dysarthric Speech due to Amyotrophic Lateral Sclerosis with Across-Speaker Articulatory Normalization

Seongjun Hahm[1], *Daragh Heitzman*[3], *Jun Wang*[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]MDA/ALS Center, Texas Neurology, Dallas, Texas, United States
{seongjun.hahm, wangjun}@utdallas.edu; dheitzman@texasneurology.com

Abstract

Recent dysarthric speech recognition studies using mixed data from a collection of neurological diseases suggested articulatory data can help to improve the speech recognition performance. This project was specifically designed for the speaker-independent recognition of dysarthric speech due to amyotrophic lateral sclerosis (ALS) using articulatory data. In this paper, we investigated three across-speaker normalization approaches in acoustic, articulatory, and both spaces: Procrustes matching (a physiological approach in articulatory space), vocal tract length normalization (a data-driven approach in acoustic space), and feature space maximum likelihood linear regression (a model-based approach for both spaces), to address the issue of high degree of variation of articulation across different speakers. A preliminary ALS data set was collected and used to evaluate the approaches. Two recognizers, Gaussian mixture model (GMM) - hidden Markov model (HMM) and deep neural network (DNN) - HMM, were used. Experimental results showed adding articulatory data significantly reduced the phoneme error rates (PERs) using any or combined normalization approaches. DNN-HMM outperformed GMM-HMM in all configurations. The best performance (30.7% PER) was obtained by triphone DNN-HMM + acoustic and articulatory data + all three normalization approaches, a 15.3% absolute PER reduction from the baseline using triphone GMM-HMM + acoustic data.

Index Terms: Dysarthric speech recognition, Procrustes matching, vocal track length normalization, fMLLR, hidden Markov models, deep neural network

1. Introduction

Although automatic speech recognition (ASR) technologies have been commercially available for healthy talkers, these technologies did not perform satisfactorily well when directly used for talkers with dysarthria, a motor speech disorder due to neurological or other injury [1]. Dysarthric speech is always with degraded speech intelligibility due to impaired voice and articulation functions [1–3]. For example, Parkinson's disease and amyotrophic lateral sclerosis (ALS) impact the patient's motor functions and therefore impair their speech. Only a few studies have been focused on dysarthric speech recognition [4–6]. Recent studies using mixed data from a variety of neurological diseases indicated articulatory data can improve the speech recognition performance [7, 8]. However, dysarthric speech recognition particularly for ALS has rarely been studied.

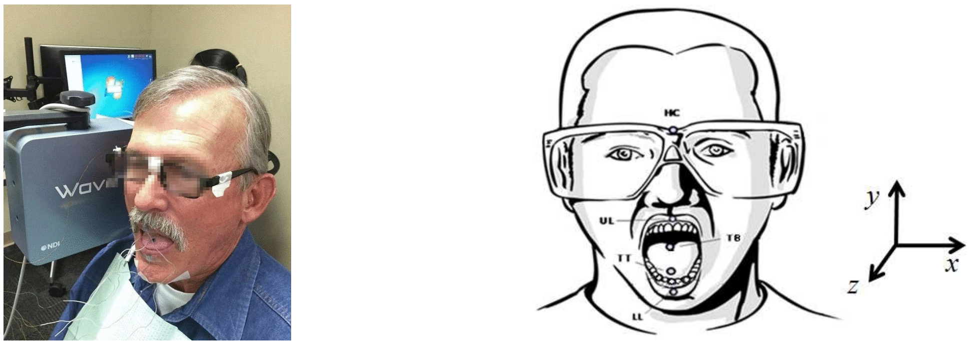
ALS, also known as Lou Gehrig's disease, is the most common motor neuron disease that causes the death of both upper and lower motor neurons [9]. The cause of the disease is unknown for most of the patients and only a small portion (5-10%) of patients is inherited [10]. As the disease progresses, the patient's speech intelligibility declines [11, 12]. Eventually all patients have degraded speech and need an assistive device for communication [13]. Normal speech recognition technology (typically trained on healthy talkers' data) does not work satisfactorily well for the patients. Therefore, ALS patients' ability to use modern speech technology (e.g., smart home environment control driven by speech recognition) is limited. This project, to our best knowledge, is the first one specifically designed to improve speech recognition performance for ALS using articulatory data.

Based on the recent literature on speech recognition with articulatory data (e.g., [7, 14–20]), we hypothesized the followings for dysarthric speech recognition for ALS: 1) adding articulatory data (collected from ALS patients) would improve the speech recognition performance, 2) feature normalization in articulatory, acoustic, and both spaces is critical and necessary for speaker-independent dysarthric speech recognition with articulatory data, and 3) recent state-of-the-art approach, deep neural network (DNN)-hidden Markov model (HMM) would outperform the long-standing approach, Gaussian mixture model (GMM)-HMM.

The high degree of variation in articulatory patterns across speakers has been a barrier for speaker-independent speech recognition with articulatory data. Multiple sources contributed to the inter-talker variation including gender, dialect, individual vocal tract anatomy, and different co-articulation patterns [21]. However, speaker-independent approaches are important for reducing the amount of training data required from each user. Only limited articulatory data samples are often available from individuals with ALS (even with healthy talkers) due to the logistic difficulty of articulatory data collection [22]. For example, in data collection using electromagnetic articulograph (EMA), small sensors have to be attached on the tongue using dental glue [23]. The procedure requires the patient to hold his/her tongue to a position for a while so that the glue can take effect.

To reduce speaker-specific difference, researchers have tried different approaches to normalize the articulatory movements including data-driven approaches (e.g., principal component analysis [7]) or physiological approaches including aligning the tongue position when producing vowels [24–26], consonants [27, 28], and pseudo-words [29] to a reference (e.g., palate [24, 25], or a general tongue shape [27]).

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 47–54,
Dresden, Germany, 11 September, 2015.



(a) *Wave System*

(b) *Sensor Locations. Labels are described in text.*

Figure 1: *Data collection setup.*

Procrustes matching, a bidimensional shape analysis technique [30], has been used to minimize the translational, scaling, and rotational effects of articulatory data across speakers [28, 29, 31]. Recent studies indicated Procrustes matching was effective for speaker-independent silent speech recognition (i.e., recognizing speech from articulatory data only) [18, 19]. Procrustes matching, however, has rarely been used in dysarthric speech recognition with articulatory data.

In addition, we adopted two other representative approaches for across-speaker data normalization. Vocal tract length normalization (VTLN) which has been widely used in acoustic speech recognition [32–36], a data-driven approach in acoustic space, was used to extract normalized acoustic features. The third approach, feature space maximum likelihood linear regression (fMLLR), a model-based adaptation, was used for both acoustic and articulatory data.

In this paper, we investigated the use of 1) articulatory data as additional information source for speech, 2) Procrustes matching, VTLN, and fMLLR as feature normalization approaches individually or combined, 3) two machine learning classifiers, GMM-HMM and DNN-HMM. The effectiveness of these speaker-independent dysarthric speech recognition approaches were evaluated with a preliminary data collected from multiple early diagnosed ALS patients.

2. Data Collection

The dysarthric speech and articulatory data used in this experiment were part of an ongoing project that targets to assess the motor speech decline due to ALS [12, 37].

2.1. Participants and stimuli

Five patients with ALS (3 females and 2 males), American English talkers, participated in the data collection (Table 1). They are all early diagnosed (within half to one year). Severity of these participants with ALS was mild with average speech intelligibility of 94.54% (SD=3.40), with SPK2 not measured. The average age of the patients was 59.80 (SD=7.73). During each session, each subject produced up to 2 or 4 repetitions of 20 unique sentences at their normal speaking rate and loudness. These sentences are used in daily conversations (e.g., *How are you?*) or related to patients (e.g., *This is an emergency*, *I need to see a doctor.*). Some of the sentences were selected from [18, 38].

2.2. Tongue motion tracking device - Wave

The Wave system (NDI Inc., Waterloo, Canada) was used to register the 3-dimensional (x, y, and z; lateral, vertical, and anterior-posterior axes) movements of the tongue and lips during speech production (Figure 1a). Our previous studies [39–41] found four articulators, tongue tip, tongue body back, upper lip, and lower lip, are optimal for this application. Therefore, we used the optimal four sensors for data collection. One sensor was attached on the subject's head and the data were used to calculate the movements of other articulators independent of the head [42]. 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 [22, 23, 38]. The spatial precision of motion tracking using Wave is approximately 0.5 mm [43]. The sampling rate for recording was 100 Hz.

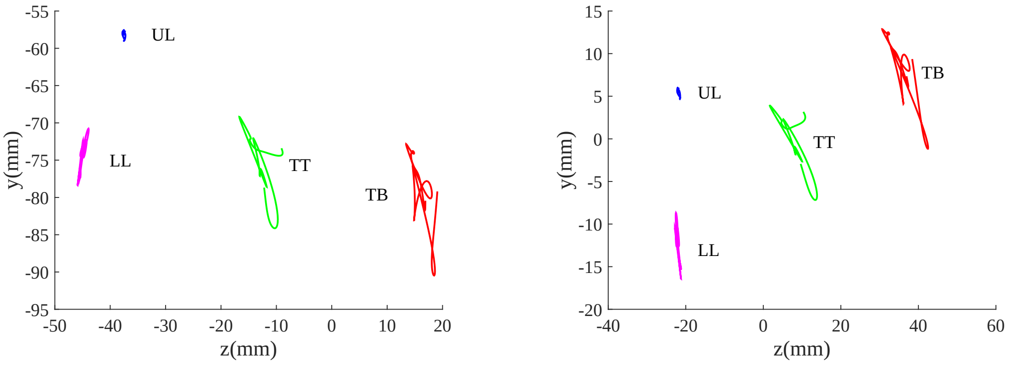
2.3. Procedure

Participants were seated with their head within a calibrated magnetic field (right next to the textbook-sized magnetic field generator). 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-minute training session helped the participants to adapt to the wired sensors before the formal data collection.

Figure 1b 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) and TB (Tongue Body Back) were attached at the mid-line of the tongue [22]. TT was about approximately 10 mm from the tongue apex. TB was as far back as possible and about 30 to 40 mm from TT [22]. 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

Table 1: *ALS participants and data size information.*

	Gender	Age	# Phrases	# Frames
SPK1	Female	53	39	5776
SPK2	Female	71	39	5219
SPK3	Male	61	79	9463
SPK4	Female	52	80	13625
SPK5	Male	62	79	9520
Total			316	43603



(a) *Original Data* (b) *After Procrustes Matching*

Figure 2: *Example of a shape (motion path of four articulators; TT, TB, UL, and LL of SPK5) for producing "Call me back when you can". In this coordinate system, y is vertical and z is anterior-posterior.*

for analysis.

2.4. Data processing

Data processing was applied on the raw sensor position data prior to analysis. 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 1b, 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 [22, 23].

In total, 316 sentence samples (for unique twenty phrases) were obtained from the five participants and were used for analysis. It could be expected ALS patients have different lateral movement patterns with healthy subjects (x in Figure 1b) [22], however for this study only y and z coordinates of the tongue and lip sensors were used for analysis.

3. Method

3.1. Procrustes matching: A physiological approach for articulatory data

Procrustes matching (or Procrustes analysis [30]) is a robust statistical bidimensional shape analysis technique, where 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 [22, 29, 31].

In this project, Procrustes matching was used to match the physiological inter-talker difference (tongue and lip orientation). The downsampled time-series multi-sensor and multi-dimensional articulatory data form articulatory shapes. An example is shown in Figure 2 [18]. This shape contains trajectories of the continuous motion paths of four sensors attached on tongue and lips, 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 [19, 22, 31].

Let S be a set of landmarks as shown below.

$$S = \{(y_i, z_i)\}, \;\; i = 1, \ldots, n \tag{1}$$

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, where y is vertical and z is front-back. The transformation in Procrustes matching is described using parameters $\{(c_y, c_z), (\beta_y, \beta_z), \theta\}$:

$$\begin{bmatrix} \bar{y}_i \\ \bar{z}_i \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \beta_y \\ \beta_z \end{bmatrix} \begin{bmatrix} y_i - c_y \\ z_i - c_z \end{bmatrix} \tag{2}$$

where (c_y, c_z) are the translation factors (centroids of the two shapes); Scaling factor β is the square root of the sum of the squares of all data points along the dimension; θ is the angle to rotate [30].

Each participant's articulatory shape 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 slightly worse performance in speaker-independent dysarthric speech recognition.

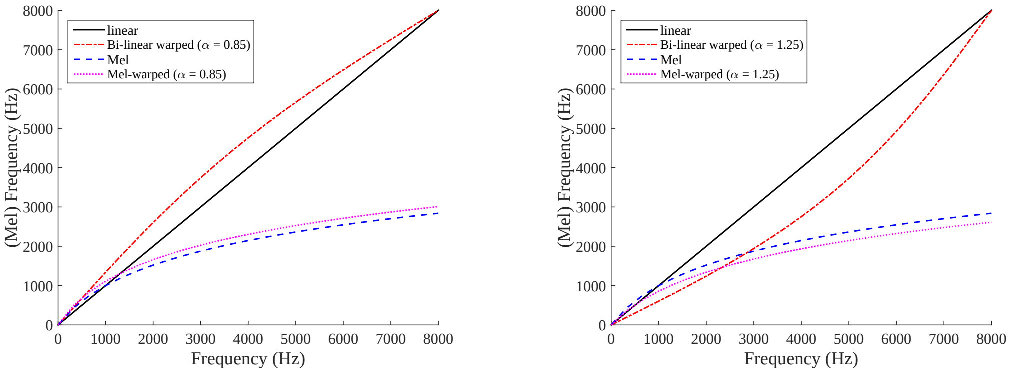
The normalization procedure was done in two steps. First, all articulatory data (e.g., a shape in Figure 2) of each speaker were translated to the centroid (average position of all data points in the shape). This step removed the locational effects between speakers. Second, all shapes of speakers were 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, (c_y, c_z) are the centroid of shape S; Scaling factor (β_y, β_z) is set to $[1\ 1]'$; θ is the angle of the S to the reference shape in which upper and lower lips form a vertical line. Figure 2 shows an example, original data (Figure 2a) and the shape after Procrustes matching (Figure 2b).

3.2. Vocal tract length normalization: A data-driven approach for acoustic data

Vocal tract length normalization is a representative approach to normalize speaker-dependent characteristics for speech recognition systems [32–36]. This approach is to normalize vocal tract length indirectly from acoustic data, because vocal tract length is highly relevant with pitch and formants [34]. Warping factor α is applied in linear frequency space by Bilinear rule,

$$\hat{F} = F + 2\tan^{-1}\left(\frac{(1-\alpha)\sin(F)}{1-(1-\alpha)\cos(F)}\right) \tag{3}$$

where F is normalized frequency (i.e., divided by sampling frequency, F_s) and α is the warping factor and $F = w/(2\pi F_s)$. Warped Mel-frequency is calculated by applying warping factor



(a) *Warping factor = 0.85 (minumum)*

(b) *Warping factor = 1.25 (maximum)*

Figure 3: *Example of (Mel) warped frequency scale (sampling rate: 16 kHz).*

α in Mel-frequency space,

$$M_\alpha(w) = 2595 \log_{10}\left(1 + \frac{w}{\alpha_0 \alpha}\right) \tag{4}$$

where α_0 is 1400π [34] and $w = 2\pi f$ (f: raw frequency). Figure 3 shows an example of (Mel) warped frequency scale between 0.85 and 1.25, the range obtained through empirical studies [34, 44].

In this work, we used linear transformation-based VTLN approach in cepstral space (MFCCs) [35, 36, 44], which was proved equivalent to the above approach [32, 34, 45].

3.3. fMLLR: A model-based approach for both articulatory and acoustic data

fMLLR (also called CMLLR; constrained maximum likelihood linear regression) is one of the representative approaches for across-speaker feature space normalization.

For each speaker, a transformation matrix $\boldsymbol{A}$ and a bias vector $\boldsymbol{b}$ are estimated and used for feature vector transformation:

$$\hat{\boldsymbol{o}}(t) = \boldsymbol{A}\boldsymbol{o}(t) + \boldsymbol{b} \tag{5}$$

where $\boldsymbol{o}(t)$ is the input feature vector at frame t and is transformed to $\hat{\boldsymbol{o}}(t)$. This transformed $\hat{\boldsymbol{o}}(t)$ is used for training GMM-HMM or DNN-HMM and also for decoding. A more detailed explanation of fMLLR can be found in [46].

3.4. Combination of normalization approaches

Besides the individual use of each normalization approach above, we also investigated combinations of these approaches. In this paper, speaker adaptive training (SAT) [46, 47] was conducted using 1) Procrustes matching, VTLN, or fMLLR individually, and 2) combinations with these approaches. We assume the speaker labels for observation are known for training stage. In testing stage, input feature vectors were also transformed using normalization approach(es) as we used in training before they were fed into GMM-HMM or DNN-HMM.

3.5. Recognizer and experimental setup

The long-standing GMM-HMM and recently available DNN-HMM were used as the recognizers [16, 20, 44, 48–50]. In this experiment, window size was 25 ms for acoustic features and frame rate was 10 ms for both acoustic and articulatory features. For each frame, static features plus derivative and acceleration form 39-dimensional mel-frequency cepstral coefficient (MFCC) vectors for acoustic features and 24-dimensional vectors for articulatory features, and these were fed into GMM-HMM or DNN-HMM. HMM is left-to-right 3-state with a monophone or a triphone context model. Maximum likelihood estimation (MLE) training approach (with or without SAT) was used for training GMM-HMM. The input layer of DNN has 216 (24 × 9 frames – 4 previous plus current plus 4 succeeding frames) dimensions for articulatory features and 351 (39 × 9 frames) dimensions for acoustic features. The output layer has 113 dimensions (36 phonemes × 3 states + 1 silence × 5 states) and approximately 200 dimensions (varies for each configuration in triphone model) for monophone and triphone models, respectively. We used 1 to 6 hidden layers and each layer had 512 nodes. The best performance obtained using 1 to 6 layers was

Table 2: *Experimental setup.*

Acoustic Feature	
Feature vector	MFCC (13-dim. vectors) + Δ + $\Delta\Delta$ (39 dim.)
Sampling rate	16 kHz
Windows length	25 ms
Articulatory Feature	
Feature vector	articulatory movement vector (8 dim.) + Δ + $\Delta\Delta$ (24 dim.)
Low pass filtering	20 Hz cutoff 5th order Butterworth
Sampling rate	100 Hz
Concatenated Feature	
Feature vector	MFCC + articulatory movement vector (21 dim.) + Δ + $\Delta\Delta$ (63 dim.)
Common	
Frame rate	10 ms
Mean normalization	Applied
GMM-HMM topology	
Monophone	113 states (36 phones × 3 states, 5 states for silence), total ≈ 1000 mixtures
Triphone	≈ 200 states, total ≈ 1750 mixtures 3-state left to right HMM
Training method	Maximum likelihood estimation (MLE) with and without SAT
DNN-HMM topology	
Input layer dim.	216 (articulatory) 351 (acoustic) 567 (concatenated)
Output layer dim.	113 (monophone) ≈ 200 (triphone)
No. of nodes	512 nodes for each hidden layer
Depth	1 to 6-depth hidden layers
Training method	RBM pre-training, back-propagation
Language model	bi-gram phoneme language model

Table 3: *Angles (in degrees) and centroids (C_y and C_z) in Procrustes matching for each patient.*

	SPK1	SPK2	SPK3	SPK4	SPK5
Angle	34.20°	32.70°	22.11°	22.85°	25.41°
C_y	-62.26	-63.31	-73.29	-71.89	-71.95
C_z	-33.51	-40.89	-26.32	-30.61	-19.60

Note: The degree indicates a counterclockwise rotation. Radians converted from degrees were actually used in the rotation.

Table 4: *Warping factor (α) for each speaker in testing or training stages.*

	CV1	CV2	CV3	CV4	CV5
SPK1	**0.94**	0.95	0.96	0.94	0.99
SPK2	0.93	**0.95**	0.94	0.92	0.98
SPK3	1.01	1.01	**0.99**	0.99	1.04
SPK4	0.95	0.95	0.97	**0.94**	1.00
SPK5	1.05	1.05	1.07	1.05	**1.06**

Note: Diagonal values are for testing and off-diagonal values are for training in each cross-validation (CV). Speakers 1, 2, and 4 are female; speakers 3 and 5 are male.

reported. Table 2 shows the detailed experimental setup. The training and decoding were performed using the Kaldi speech recognition toolkit [44].

Phoneme error rate (PER) was used as the measure of dysarthric 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.

4. Results & Discussion

Table 3 shows detailed parameters (angles and centroids) for Procrustes matching, which varies for different speakers. Table 4 and Figure 4 show the warping factors for each speaker and their histogram. The histogram of ALS patients follows general trend of warping factor distribution for females (typically < 1.0) and males (typically > 1.0).

Figures 5, 6, 7, and 8 give the PERs of speaker-independent dysarthric (due to ALS) speech recognition results using different context models and recognizers, respectively: (1) monophone GMM-HMM, (2) triphone GMM-HMM, (3) monophone DNN-HMM, and (4) triphone DNN-HMM with individual or combinations of VTLN, Procrustes matching, and fMLLR. These results suggest that VTLN, Procrustes matching, and fMLLR were all effective for speaker-independent dysarthric speech recognition from acoustic data, articulatory data, or combined. When comparing the three normalization approaches individually (if applies), no approach was universally better than others in all experimental configurations. A better performance was always obtained when the normalization approaches were combined. Baseline results were obtained without using any normalization approach.

Adding articulatory data to acoustic data always showed performance improvement in all configurations (monophone/triphone or GMM-HMM/DNN-HMM), which is consistent with the literature [7]. The overall best performance was obtained when the three normalization approaches, VTLN (acoustic space), Procrustes matching (articulatory space), and fMLLR (both acoustic and articulatory space), were used together with triphone DNN-HMM model (30.7%).

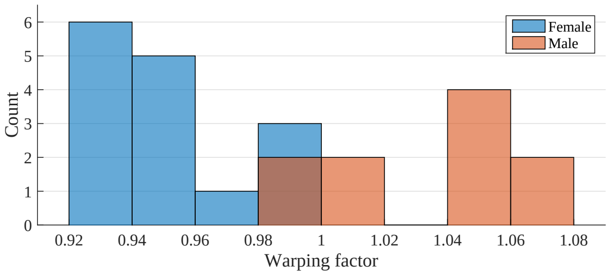


Figure 4: *Histogram of warping factors (step size = 0.02).*

Surprisingly, speaker-independent silent speech recognition (using articulatory data only) with DNN-HMM obtained even better results than the recognition results from acoustic (MFCC) features (see left half of Figures 7 and 8). This finding shows the potential of articulatory data when the patient's speech is significantly impaired as the disease progresses. However, since the data set is small, a further study with a larger data set is required to verify this finding.

Moreover, DNN-HMM outperformed GMM-HMM in all configurations (monophone/triphone, VTLN/Procrustes matching/fMLLR). This finding is consistent with the acoustic [20,51] and silent speech recognition literature [17, 19].

In the current approach, fMLLR was not separately applied to acoustic and articulatory data (i.e., full transformation matrix), because the two types of data are concatenated before applying fMLLR. Due to the different nature of acoustic (in frequency domain) and articulatory data (in spatial domain), in the future, we consider to make $\boldsymbol{A}$ in Eq. 5 a block-diagonal transformation matrix. The block-diagonal matrix will separate the processing for acoustic and articulatory data.

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 ALS patients. Further studies with a larger vocabulary from more ALS patients are necessary to explore the limits of the current approaches.

5. Conclusions & Future Work

This paper investigated speaker-independent dysarthric speech recognition using the data from patients with ALS and also with three across-speaker normalization approaches: a physiological approach, Procrustes matching, a data-driven approach, VTLN, and a model-based approach, fMLLR. GMM-HMM and DNN-HMM were used as the machine learning classifiers. Experimental results showed the effectiveness of feature normalization approaches. The best performance was obtained when the three approaches were used together with triphone DNN-HMM.

Future work includes test of the normalization approaches using a larger data set collected from more ALS subjects (e.g, by combining our data set with the ALS data in TORGO [8]).

6. Acknowledgments

This work was supported by the National Institutes of Health through grants R01 DC013547 and R03 DC013990. We would like to thank Dr. Jordan R. Green, Dr. Thomas F. Campbell, Dr. Anusha Thomas, Jana Mueller, Victoria Juarez, Saara Raja, Soujanya Koduri, Kumail Haider, the volunteering partic-

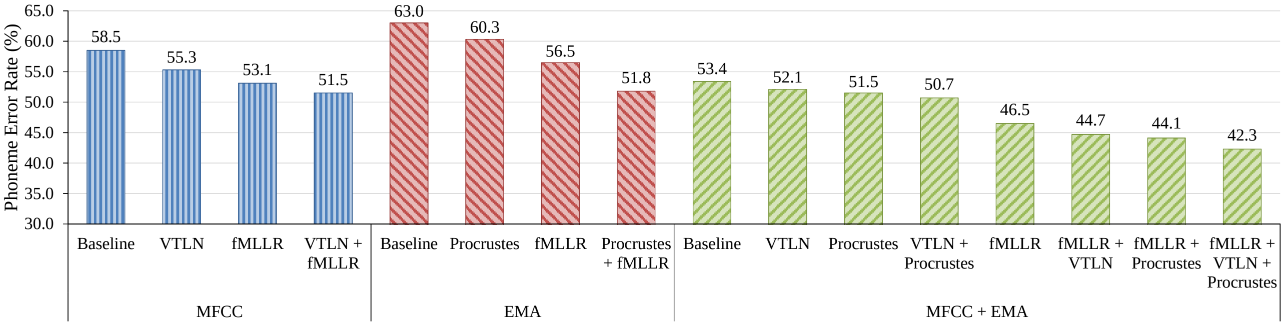


Figure 5: *Phoneme Error Rates (PERs; %) of speaker-independent recognition using* ***monophone GMM-HMM*** *with fMLLR, VTLN, and/or Procrustes matching.*

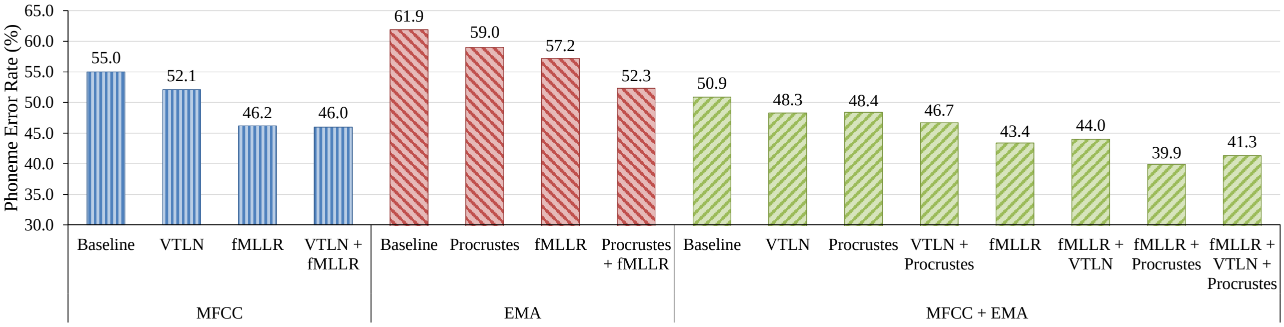


Figure 6: *Phoneme Error Rates (PERs; %) of speaker-independent recognition using* ***triphone GMM-HMM*** *with fMLLR, VTLN, and/or Procrustes matching.*

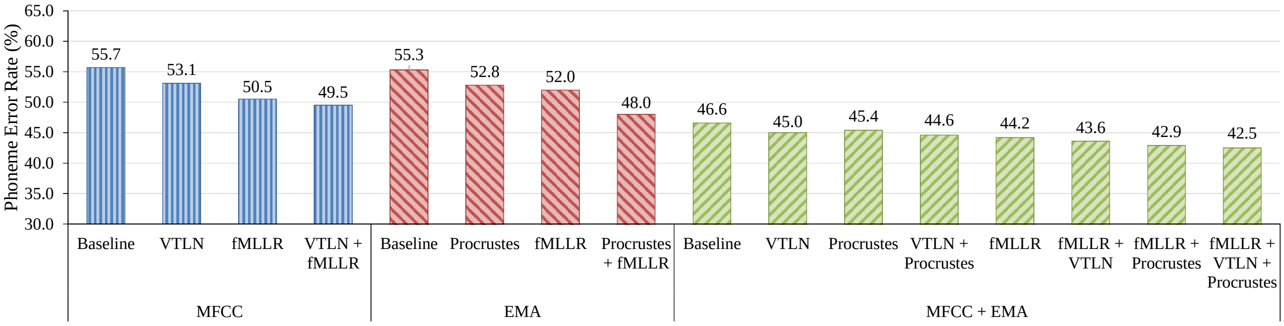


Figure 7: *Phoneme Error Rates (PERs; %) of speaker-independent recognition using* ***monophone DNN-HMM*** *with fMLLR, VTLN, and/or Procrustes matching.*

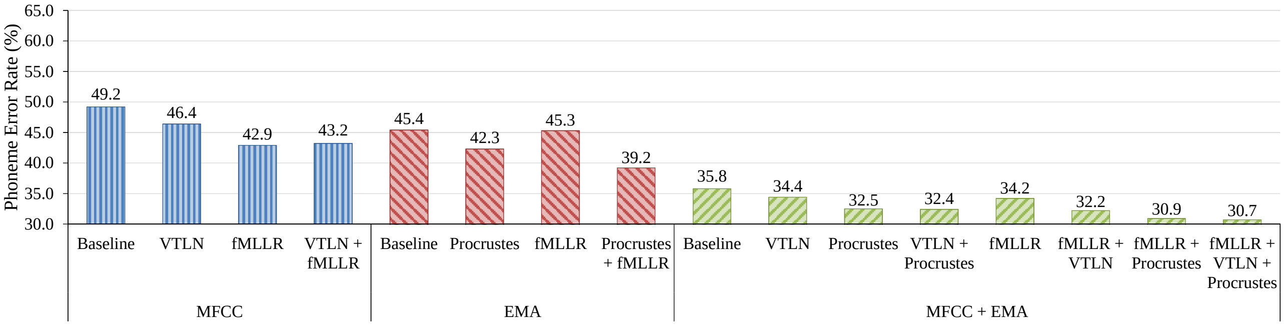


Figure 8: *Phoneme Error Rates (PERs; %) of speaker-independent recognition using* ***triphone DNN-HMM*** *with fMLLR, VTLN, and/or Procrustes matching.*

ipants, and the Communication Technology Center, University of Texas at Dallas.

7. References

[1] J. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Mosby; 3rd edition, 2005.

[2] E. Hanson, K. M. Yorkston, and D. Britton, "Dysarthria in amyotrophic lateral sclerosis: A systematic review of characteristics, speech treatment and AAC options," *Journal of Medical Speech - Language Pathology*, vol. 19, no. 3, pp. 12–30, 2011.

[3] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel contrast and speech intelligibility in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4, pp. 187–194, 2011.

[4] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech and Language*, vol. 27, no. 6, pp. 1147–1162, 2013.

[5] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP J. Adv. Signal Process*, vol. 2009, pp. 2:1–2:14, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/308340

[6] K. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4924–4927.

[7] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2011.

[8] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria." *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[9] M. Kiernan, S. Vucic, B. Cheah, M. Turner, A. Eisen, O. Hardiman, J. Burrell, and M. Zoing, "Amyotrophic lateral sclerosis," *Lancet*, vol. 377, no. 9769, pp. 942–955, 2011.

[10] J. Sreedharan, I. P. Blair, V. B. Tripathi, X. Hu, C. Vance, B. Rogelj, S. Ackerley, J. C. Durnall, K. L. Williams, E. Buratti, F. Baralle, J. de Belleroche, J. D. Mitchell, P. N. Leigh, A. Al-Chalabi, C. C. Miller, G. Nicholson, and C. E. Shaw, "TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis," *Science*, vol. 319, no. 5870, pp. 1668–1672, 2008.

[11] B. R. Brooks, R. Miller, M. Swash, and T. Munsat, "El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotroph Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 1, no. 5, pp. 293–299, 2000.

[12] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, pp. 494–500, 2013.

[13] D. Beukelman, S. Fager, and A. Nordness, "Communication support for people with ALS," *Neurology Research International*, no. 714693, p. 6 pages, 2011.

[14] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[15] J. Wang, "Silent speech recognition from articulatory motion," Ph.D. dissertation, The University of Nebraska-Lincoln, 2011.

[16] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Cross-corpus and cross-linguistic evaluation of a speaker-dependent DNN-HMM ASR system using EMA data," in *Proc. of Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, 2013.

[17] S. Hahm and J. Wang, "Silent speech recognition from articulatory movements using deep neural network," in *Proc. of the 18th Intl. Congress of Phonetic Sciences*, 2015.

[18] J. Wang, A. Samal, and J. Green, "Across-speaker articulatory normalization for speaker-independent silent speech recognition," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 1179–1183.

[19] J. Wang and S. Hahm, "Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training," in *Proc. of INTERSPEECH*, 2015.

[20] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping." in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 1297–1301.

[21] R. Kent, S. Adams, and G. Tuner, "Models of speech production," in *Principles of Experimental Phonetics, (Lass, N.J., ed.)*, 1996, pp. 3–45, Mosby.

[22] J. Wang, J. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[23] J. Green, J. Wang, and D. L. Wilson, "Smash: A tool for articulatory data processing and analysis," in *Proc. of INTERSPEECH*, 2013, pp. 1331–1335.

[24] K. Johnson, P. Ladefoged, and M. Lindau, "Individual differences in vowel production," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 701–714, 1993.

[25] M. Hashi, J. R. Westbury, and K. Honda, "Vowel posture normalization," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2426–2437, 1998.

[26] A. P. Simpson, "Gender-specific differences in the articulatory and acoustic realization of interword vowel sequences in american english," in *5th Seminar on Speech Production: Models and Data. Kloster Seeon*, 2000, pp. 209–212.

[27] J. R. Westbury, M. Hashi, and M. J Lindstrom, "Differences among speakers in lingual articulation for American English /ɹ/," *Speech Communication*, vol. 26, no. 3, pp. 203–226, 1998.

[28] S. Li and L. Wang, "Cross linguistic comparison of Mandarin and English EMA articulatory data," in *Proc. of INTERSPEECH*, 2012, pp. 903–906.

[29] D. Felps, S. Aryal, and R. Gutierrez-Osuna, "Normalization of articulatory data through procrustes transformations and analysis-by-synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3027–3031.

[30] I. L. Dryden and K. V. Mardia, *Statistical shape analysis*. John Wiley & Sons New York, 1998, vol. 4.

[31] J. Wang, J. R. Green, A. Samal, and D. B. Marx, "Quantifying articulatory distinctiveness of vowels," in *Proc. of INTERSPEECH*, 2011, pp. 277–280.

[32] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. of ICASSP*, vol. 1, 1996, pp. 346–348.

[33] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Proc. of ICASSP*, vol. 2, 1997, pp. 1039–1042.

[34] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," *CMU-CS-97-148, Carnegie Mellon University, Pittsburgh, PA*, 1997.

[35] T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," in *Proc. of Eurospeech*, 2001, pp. 1649–1652.

[36] D. Kim, S. Umesh, M. Gales, T. Hain, and P. Woodland, "Using VTLN for broadcast news transcription," in *Proc. of INTERSPEECH*, 2004, pp. 1953–1956.

[37] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, "Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behavioral Neurology*, no. 183027, pp. 1–11, 2015.

[38] J. Wang, A. Samal, J. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 4985–4988.

[39] J. Wang, J. Green, and A. Samal, "Individual articulator's contribution to phoneme production," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7785–7789.

[40] J. Wang, S. Hahm, and T. Mau, "Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition," in *ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015.

[41] J. Wang, A. Samal, P. Rong, and J. R. Green, "An optimal set of flesh points on tongue and lips for speech movement classification," *Journal of Speech, Language, and Hearing Research*, In press.

[42] J. R. Green and Y.-T. Wang, "Tongue-surface movement patterns during speech and swallowing," *The Journal of the Acoustical Society of America*, vol. 113, no. 5, pp. 2820–2833, 2003.

[43] J. Berry, "Accuracy of the NDI wave speech research system." *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1295–301, 2011.

[44] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and V. K., "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, Waikoloa, USA, 2011, pp. 1–4.

[45] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[46] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and language*, vol. 12, no. 2, pp. 75–98, 1998.

[47] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Proc. of ICASSP*, vol. 2, 1997, pp. 1043–1046.

[48] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[49] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[50] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at microsoft," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 8604–8608.

[51] A.-R. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.