SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies
13 September 2016, San Francisco, USA

Recognizing Whispered Speech Produced by an Individual with Surgically Reconstructed Larynx Using Articulatory Movement Data

Beiming Cao[1], *Myungjong Kim*[1], *Ted Mau*[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]Department of Otolaryngology - Head and Neck Surgery
University of Texas Southwestern Medical Center, Dallas, Texas, United States
{beiming.cao, myungjong.kim, wangjun}@utdallas.edu; ted.mau@utsouthwestern.edu

Abstract

Individuals with larynx (vocal folds) impaired have problems in controlling their glottal vibration, producing whispered speech with extreme hoarseness. Standard automatic speech recognition using only acoustic cues is typically ineffective for whispered speech because the corresponding spectral characteristics are distorted. Articulatory cues such as the tongue and lip motion may help in recognizing whispered speech since articulatory motion patterns are generally not affected. In this paper, we investigated whispered speech recognition for patients with reconstructed larynx using articulatory movement data. A data set with both acoustic and articulatory motion data was collected from a patient with surgically reconstructed larynx using an electromagnetic articulograph. Two speech recognition systems, Gaussian mixture model-hidden Markov model (GMM-HMM) and deep neural network-HMM (DNN-HMM), were used in the experiments. Experimental results showed adding either tongue or lip motion data to acoustic features such as mel-frequency cepstral coefficient (MFCC) significantly reduced the phone error rates on both speech recognition systems. Adding both tongue and lip data achieved the best performance.

Index Terms: whispered speech recognition, larynx reconstruction, speech articulation, deep neural network, hidden Markov model

1. Introduction

Larynx is one of the most important articulators for speech and sound production. Vocal fold vibration produces the sounding source for speech. People who have their larynx (vocal fold) impaired due to physical impairment or treatment of laryngeal cancer suffer during their daily life. A surgery can help these patients reconstruct or repair their larynx, but their phonation can hardly be completely recovered [1]. Patients with surgically reconstructed larynx generally have problems in controlling laryngeal functions, thus producing whispered speech with extreme hoarseness [2]. Therefore, assistive automatic speech recognition (ASR) technology is necessary so that they can interact with computers or smart phones in their daily life like normal people do. A standard ASR system that focuses on recognizing normal speech does not work well for these patients, because their speech mostly contains an unvoiced mode of phonation. Thus, ASR systems that are specialized for whispered speech are needed [3].

Whispered speech produced by patients with reconstructed larynx can be treated as a kind of disordered speech, which degrades the performance of conventional speech recognition systems [4, 5]. Whispered speech misses glottal excitation, leading to abnormal energy distribution between phone classes, variations of spectral tilt, and formant shifts due to abnormal configurations of the vocal tract [3, 6], which are the main causes of performance degradation of a standard ASR system. To improve the performance of whispered speech recognition, most of the conventional studies used whispered speech data that are collected from normal talkers and focused on reducing the mismatch between normal and whispered speech in acoustic domain through acoustic model adaptation and feature transformation [7–11].

Articulatory information has been proven effective in the applications of normal speech recognition [12–16] and dysarthric speech recognition [17, 18]. Compared to acoustic features, articulatory features are expected to be less affected for these patients who produce whispered speech [19]. There are a few studies applying articulatory or non-audio information in whispered speech recognition [10, 19, 20]. For example in [19], the authors applied articulatory features (also known as phonological attributes) of whispered speech. Most of the existing work using articulatory information focused on descriptive or derived articulatory features in acoustic domain. Articulatory movement data such as tongue and lip motion have rarely been applied in this application.

In this paper, we investigated whispered speech recognition for a larynx reconstructed patient using tongue and lip motion data. To our knowledge, this is the first study for whispered speech recognition with articulatory data. Tongue and lip motion data were collected using an electromagnetic articulograph. Two speech recognizers were used: Gaussian mixture model-hidden Markov model (GMM-HMM) and deep neural network-hidden Markov model (DNN-HMM). In the experiments, we examined several settings on both speech recognition systems to verify the effectiveness of adding articulatory movement data: mel-frequency cepstral coefficient (MFCC)-based acoustic features, lip and tongue movement-based articulatory data, and MFCC with articulatory data. The remaining of the paper is organized as follows: Section 2 describes our acoustic and articulatory data collected from a patient with surgically reconstructed larynx. In Section 3, we present our experimental design including speech recognition systems and experimental setup. Section 4 shows experimental results and discussion. Conclusions are summarized in Section 5.

10.21437/SLPAT.2016-14

(a) Wave System (b) Sensor Locations

Figure 1: *Articulatory (tongue and lip) motion data collection setup.*

2. Data Collection

2.1. Participants and stimuli

The patient is a male of 23 years old. He had his larynx damaged in an accident and then took a larynx reconstruction surgery in 2014. His speech showed extreme hoarseness. He is not using assistive device on a daily basis. He participated in the data collection, where he produced a sequence of 132 phrases at his habitual speaking rate. The phrases were selected from the phrases that are frequently spoken by persons who use augmentative and alternative communication (AAC) devices [21, 22].

2.2. Tongue motion tracking device and procedure

An electromagnetic articulograph (Wave system, Northern Digital Inc., Waterloo, Canada) was used for articulatory data collection (Figure 1a). Four small sensors were attached to the surface of patient's articulators, two of them were attached to tongue tip (TT, 5-10mm to tongue apex) and tongue back (TB, 20-30mm back from TT) using dental glue (PeriAcry1 90, GluStitch). The other two were attached to upper lip (UL) and lower lip (LL) using normal double-sided tape. In addition, another sensor was attached to the middle of forehead for head correction. Our prior work indicated that the four-sensor set consisting of tongue tip, tongue back, upper lip, and lower lip are an optimal set for this application [23–25]. The positions of all five sensors are shown in Figure 1b. With this approach, three-dimension movement data of articulators were tracked and recorded. The sampling rate in Wave recording in this project was 100Hz, The spatial precision of movement tracking is about 0.5mm [26].

The patient was seated next to the magnetic field generator, which is the blue box in Figure 1a, and read the 132 phrases. A three-minute training session helped the patient to adapt to speak with tongue sensors before the formal data collection session.

Before data analysis, the translation and rotation of the head sensor were subtracted from the motion data of tongue and lip sensors to obtain head-independent articulatory data. Figure 1b illustrates the derived 3D Cartesian coordinates system, in which x is left-right direction; y is vertical; and z is front-back direction. We assume the tongue and lip motion patterns of the patient remain the same as normal talkers, where the movement in x direction is not significant in speech production. Therefore, only y and z coordinates were used for analysis in this study [27].

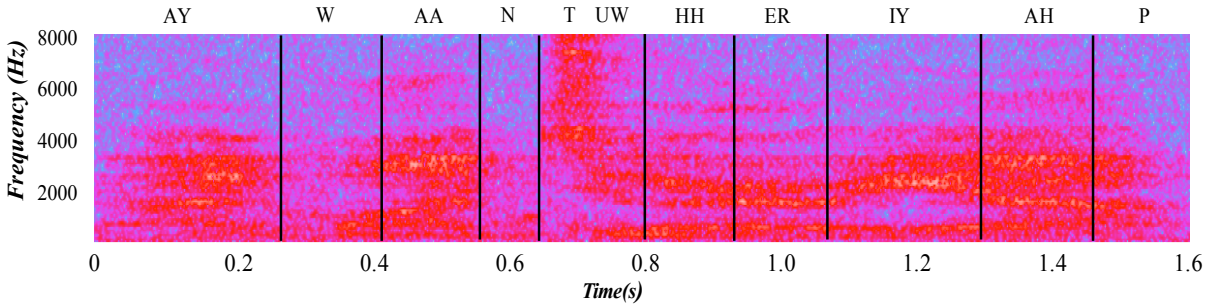
Acoustic data were collected synchronously with the articulatory movement data by built-in microphone in the Wave system. In total, the data set contains 2,292 phone samples of 39 unique phones.

2.3. Acoustic data

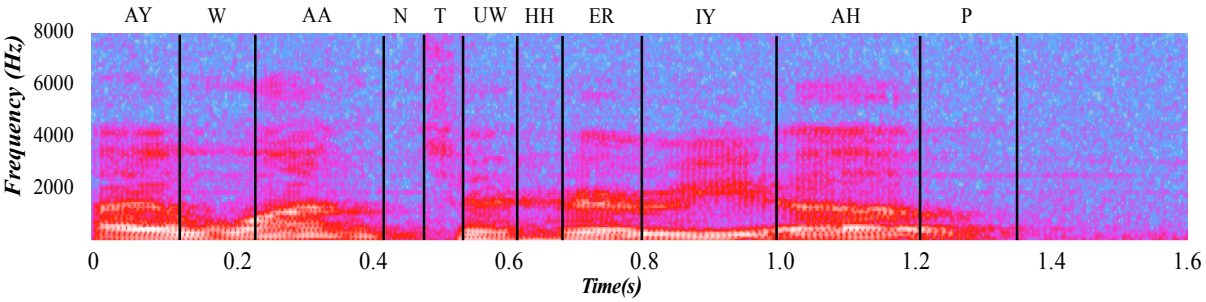
Figure 2 shows the spectrograms of whispered speech and normal (vocalized) speech examples producing the same phrase *I want to hurry up*. Figure 2a is an example spectrogram of whispered speech produced by the patient in this study. Figure 2b is an example of normal speech produced by a healthy speaker. The healthy speaker's data example was just used to illustrate the difference between the spectrograms of whispered and normal speech, and therefore it was not used in analysis. In the figure, brighter color (and reddish) denotes higher energy. As illustrated in Figure 2, for normal speech, the phone boundaries are relatively clear based on spectral energy shape and formant frequencies, and it is easy to distinguish. For whispered speech, however, most of phones have very similar spectral pattern without fundamental frequency, which makes it hard to find the boundary between the phones using acoustic information only. For example, the phone pairs like 'AH' and 'P' in word 'up', can hardly be distinguished in whispered speech, also the 'HH' and 'ER' in word 'hurry' are not easy to classify. On the other hand, those two phone pairs can be clearly distinguished in normal speech, showing that vowels have higher energy and distinct formant frequencies. The ambiguity of phone boundaries contributed to lower performance in whispered speech recognition using standard ASR techniques.

2.4. Articulatory data

Figure 3a and 3b give examples of articulatory movement data, which are obtained from the motion tracking of sensors when uttering same phrase (*I want to hurry up*) in Figure 2, respectively. As mentioned previously, four sensors were attached to articulators (upper lip, lower lip, tongue tip, and tongue back). As illustrated in Figure 3, the articulatory movement pattern of whispered speech somewhat resembles the articulatory movement pattern of normal speech, although the motion range of whispered speech by the patient was larger than that by the healthy talker. Therefore, we expected that articulatory motion data may improve the performance of whispered speech recognition. The larger motion range of tongue and lips of the patient may be because he uses more force than normal talkers during his speech production. For illustration purpose, the two articulatory shapes (tongue and lip sensor motion curves) in Figure



(a) Whispered (hoarse) speech produced by the patient



(b) Normal Speech

Figure 2: *Spectrogram examples of whispered and normal speech producing "I want to hurry up".*

3 were rotated slightly so that the UL and the LL sensors were aligned vertically.

3. Method

Standard speech recognition systems are typically based on hidden Markov models (HMMs), which are an effective framework for modeling time-varying spectral feature vector sequence [28]. A number of techniques for adapting or describing input features are typically used together with HMM. In this study, we used two speech recognizers: the long-standing Gaussian mixture model (GMM)-HMM and the recently available deep neural network (DNN)-HMM. Major parameter configuration of the two recognizers was shown in Table 1.

3.1. GMM-HMM

The Gaussian mixture model (GMM) is a statistical model for describing speech features in a conventional speech recognition system. Given enough Gaussian components, GMMs can model the relationship between acoustic features and phone classes as a mixture of Gaussian probabilistic density functions. More detail explanation of GMM can be found in [29]. GMM-HMM is a model that is "hanging" GMMs to states of HMM, in which GMMs are used for characterizing speech features and HMM is responsible for characterizing temporal properties.

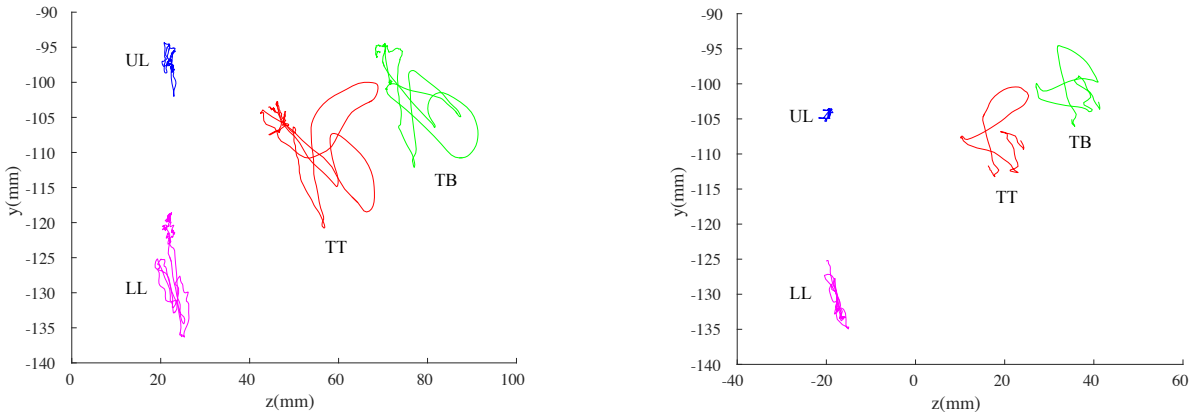
GMM-HMM have been widely used in modeling speech features and as an acoustic model for speech recognition for decades until DNN attracted more interests in the literature recently. However, GMM is still promising when using a small data set. In addition, because of its rapid implementation and execution, we included GMM as a baseline approach. Table 1 gives the major parameters for GMM-HMM.

3.2. DNN-HMM

Deep neural networks (DNNs) with multiple hidden layers have been shown to outperform GMMs on a variety of speech recognition benchmarks [30] including recent works that involved articulatory data [17, 31]. DNN-HMM takes multiple frames of speech features as input and produces posterior probabilities over HMM states as output. The DNN training is based on restricted Boltzmann machines (RBMs). The weights between nodes in neighboring layers at iteration $t+1$ are updated based on iteration t using stochastic gradient descent described by the following equation:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \tag{1}$$

in which w_{ij} is the weight between nodes i and j of two layers next to each other, η is the learning rate, and C is the cost function. The output posterior probabilities of DNN are used for decoding. More detailed description of DNN can be found in [30, 32–34]. The similar structure and setup of DNN in [31] were used in this experiment which has 5 hidden layers, each of hidden layers has 512 nodes. We tested all layers from 1 to 6 in each experimental configuration, and the best result was obtained when using 5 hidden layers. The one-subject data set has a relatively small size, thus we used only 512 nodes. The input layer would take 9 frames at a time (4 previous plus current plus 4 succeeding frames), therefore the dimension of input layer changed given different types of data. For example, for the experiments using both MFCC and articulatory data, the dimension of each frame is 13-dimensional MFCC plus 8-dimensional movement plus their delta and delta of delta formed a 63-dimensional vectors that were fed into the DNN. But for the experiments using only MFCC, the frame dimension is 39. The output layer has 122 dimensions (39 phones $\times$ 3 states each phone plus 5 states for silence).



(a) Whispered speech produced by the patient

(b) Normal speech produced by a healthy talker

Figure 3: *Examples of motion path of four articulators TT, TB, UL, and LL of whispered and normal speech for producing "I want to hurry up". In this coordinate system, y is vertical and z is anterior-posterior.*

3.3. Experimental Setup

In this project, frame rate was 10 ms (equivalent to sampling rate of articulatory data recording: 100 Hz). Two dimensional (vertical and anterior-posterior) EMA data of four sensors (tongue tip, tongue body back, upper lip, and lower lip) were used for the experiments. As mentioned previously, for each frame, all either acoustic features, i.e., MFCCs or articulatory movement data plus delta and delta of delta form vectors that were fed into a recognizer. HMM has left-to-right 3-states with a context-independent monophone models. Triphone models were not considered due to the small size of our data set in this work. A bi-gram phone-level language model was used. The training and decoding were performed using the Kaldi speech recognition toolkit [35].

Phone error rate (PER) was used as a whispered speech recognition performance measure, which is the summation of deletion, insertion, and substitution phone errors divided by the number of all phones. For each of two recognizers, whispered speech recognition experiments were conducted using different combinations of features, including MFCC only, MFCC concatenated with lip motion data, MFCC with tongue data, and MFCC with both of lip and tongue data.

Three-fold cross validation was used in the experiments. The average performance of the executions was calculated as the overall performance.

Table 1: *Experimental setup.*

Acoustic Feature	
Feature vector	MFCC (13-dim. vectors) + Δ + $\Delta\Delta$ (39 dim.)
Sampling rate	22050 kHz
Windows length	25 ms
Articulatory Feature (both tongue and lips)	
Feature vector	articulatory movement vector (8 sensors) + Δ + $\Delta\Delta$ (24 dim.)
Concatenated Feature	
Feature vector	MFCC + articulatory movement vector (21-dim vector) + Δ + $\Delta\Delta$ (63 dim.)
Common	
Frame rate	10 ms
GMM-HMM topology	
Monophone	122 states (39 phones $\times$ 3 states, 5 states for silence), total $\approx$ 1000 Gaussians (each state $\approx$ 8 Gaussians) 3-state left to right HMM
Training method	maximum likelihood estimation (MLE)
DNN-HMM topology	
Input	9 frames at a time (4 previous plus current plus 4 succeeding frames)
Input layer dim.	216 (9 $\times$ 24 for articulatory) 351 (9 $\times$ 39 for acoustic) 567 (concatenated)
Output layer dim.	122 (monophone)
No. of nodes	512 nodes for each hidden layer
Depth	5-depth hidden layers
Training method	RBM pre-training, back-propagation
Language model	bi-gram phone language model

4. Results & Discussion

Figure 4 shows the average PERs of speaker-dependent whispered speech recognition for the patient. The baseline results (67.2% for GMM-HMM and 66.1% for DNN-HMM) were obtained using only acoustic (MFCC) features.

The PERs were reduced by adding either tongue motion data (63.6% for GMM-HMM and 63.3% for DNN-HMM) or lip motion data (65.6% for GMM-HMM and 65.6% for DNN-HMM) to MFCC features although the PERs of independent lip motion data or tongue motion data are higher than that obtained with MFCC features only. Particularly, using tongue motion data was more effective than with lip motion data, producing better phone recognition performance. This result is consistent with our previous speech recognition tasks with articulatory data [24], because tongue motion contains more information than lip motion during speech production [25].

The best performance was achieved when both lip and tongue data were applied with acoustic data, 63.0% for GMM-HMM and 62.1% for DNN-HMM. These results indicate that MFCC, lip motion data, and tongue motion data have complementary information in distinguishing phones.

A two-tailed t-test was performed to measure if there were statistical significance between the performances of the configuration with MFCC only and other configurations. As indicated in Figure 4, most data configurations of MFCC+articulatory features showed a statistical significance with the MFCC configuration. The results suggested that adding tongue data or both lip and tongue data to MFCC features significantly im-

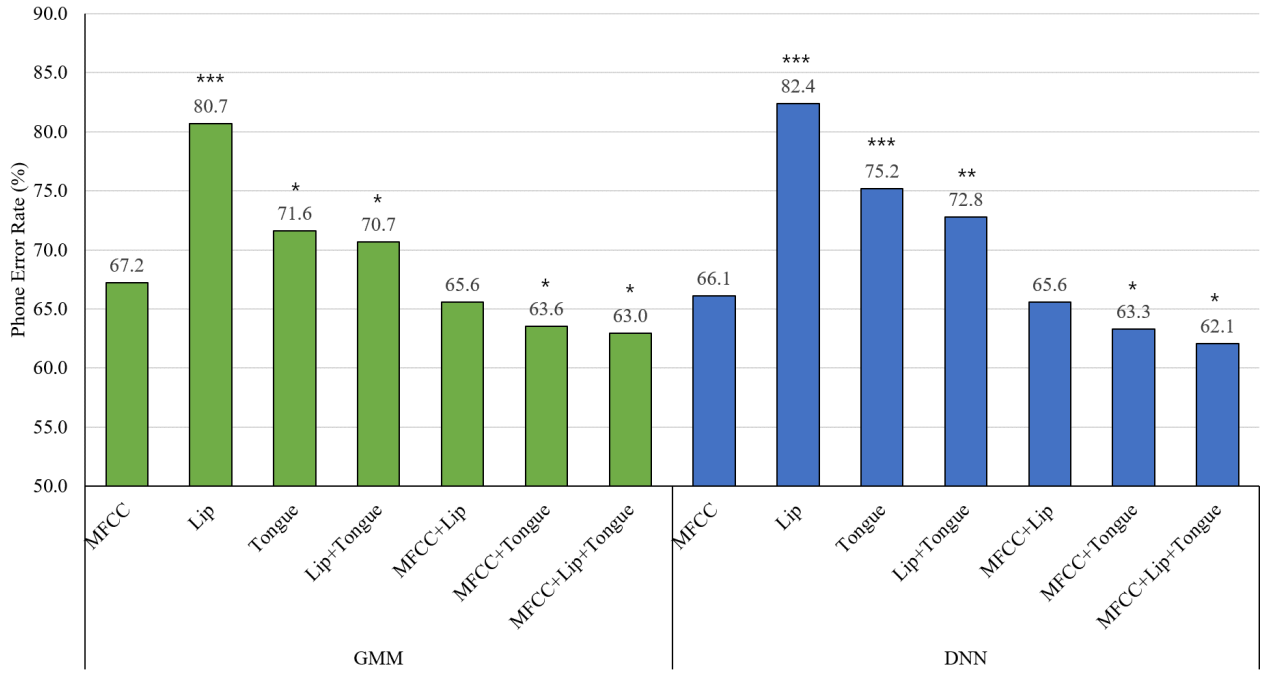


Figure 4: *Phone error rates of whispered speech recognition obtained in the experiments using monophone GMM-HMM and DNN-HMM with various types of data (MFCC, Lip, Tongue, Lip+Tongue, MFCC+Lip, MFCC+Tongue, and MFCC+Lip+Tongue). Statistical significances between the results obtained using MFCC and other data types on each ASR model are marked: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.*

proved the performance. Adding lip movement data, however, did not show a significance, although a slight improvement was observed. The observation may be because of the small data size. A further study with a data set of larger size is needed to verify these findings.

To understand the contribution by adding articulatory movement data in whispered speech recognition, we also tested the recognition performance from articulatory data only (i.e., without using acoustic features, or silent speech recognition). Figure 4 gives the silent speech recognition results of using GMM-HMM and DNN-HMM, respectively. For both GMM-HMM and DNN-HMM, the least performances were obtained when using lip data only; the best performances were obtained when using both tongue and lip data. The results on individual articulatory data configurations (lip, tongue, lip + tongue) were positively correlated by the contribution of adding the data. In other words, using tongue data only obtained a better recognition performance than using lip data only, which explained why adding tongue information better improved the whispered speech recognition than adding lip information. These findings are consistent with our previous work for silent speech recognition using data from combinations of articulators (sensors) [24, 25]. Using only articulatory data always obtained less performance than using acoustic data only, which is also consistent with our prior finding [24].

Table 2: *Numbers of deletion, insertion, and substitution errors in the experiment of whispered speech recognition with articulatory data.*

Model	Feature	Del	Ins	Sub
GMM	MFCC	723	78	740
	MFCC+Lip	640	110	752
	MFCC+Tongue	657	109	691
	MFCC+Lip+Tongue	652	117	674
DNN	MFCC	696	71	752
	MFCC+Lip	782	62	656
	MFCC+Tongue	761	59	632
	MFCC+Lip+Tongue	783	56	610

In addition, Table 2 gives a summary of deletion, insertion, and substitution in the phone recognition errors in these experiments. Table 2 provides more details that different articulatory data decrease the PER in different ways. For GMM-HMM, adding lip data would decrease the number of deletion by 83 but increased the numbers of insertion and substitution. However, adding tongue data decreased the number of substitution and deletion, but increased the number of insertion. For DNN-HMM, adding either tongue or lip data would considerably decrease insertion and substitution errors, although it increased deletion errors. As discussed earlier, we think adding articulatory motion data will help the recognizer to find the boundaries between phones. However, how tongue or lips affect the number of deletion, insertion, and substitution needs to be verified with a larger data set.

DNN typically outperformed GMM in ASR using acoustic data only [30] or using both acoustic and articulatory data [17, 33]. In this work, DNN performance was slightly better than that of GMM as well. Although our data set is small and DNN typically requires a larger data set, DNN still can model the complex structure of whispered speech in this project. This result indicates that DNN will be promising for whispered

speech recognition with articulatory data for a larger, multiple-speaker data set.

In summary, the experimental results demonstrated the effectiveness of applying articulatory movement data to whispered (hoarse) speech recognition. In addition, the results indicated that adding tongue motion data will improve the performance more than that by adding lip motion data in whispered speech recognition. The best performance was obtained when both tongue and lip motion data were used.

Limitation. Although the results are promising, the method (adding articulatory data on top of acoustic data) has been evaluated with only one subject (patient) with whispered speech. A further study with a multiple-speaker data set is needed to verify these findings.

5. Conclusions & Future Work

The effectiveness of articulatory (tongue and lips) movement data in whispered speech recognition has been tested with a data set that was collected from an individual with a surgically reconstructed larynx. The experimental results suggested that adding articulatory movement data decreased the PER of whispered speech recognition for widely used ASR models: GMM-HMM and DNN-HMM. The best performance was obtained when acoustic, tongue, and lip movement data were used together.

Future work includes verifying the findings using a larger data set and using other latest ASR models such as deep recurrent neural networks [36].

6. Acknowledgements

This work was supported by the National Institutes of Health (NIH) under award number R03 DC013990 and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We thank Dr. Seongjun Hahm, Joanna Brown, Betsy Ruiz, Janis Deane, Laura Toles, Christina Duran, Amy Hamilton, and the volunteering participants.

7. References

[1] T. Mau, J. Muhlestein, S. Callahan, and R. W. Chan, "Modulating phonation through alteration of vocal fold medial surface contour," *The Laryngoscope*, vol. 122, no. 9, pp. 2005–2014, 2012.

[2] T. Mau, "Diagnostic evaluation and management of hoarseness," *Medical Clinics of North America*, vol. 94, no. 5, pp. 945 – 960, 2010.

[3] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[4] M. Kim, J. Wang, and H. Kim, "Dysarthric speech recognition using kullback-leibler divergence-based hidden markov model," in *Interspeech*, 2016.

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

[6] S. Ghaffarzadegan, H. Bořil, and J. H. L. Hansen, "Model and feature based compensation for whispered speech recognition," in *Interspeech 2014*, Singapore, Sept 2014, pp. 2420–2424.

[7] ——, "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2014*, Florence, Italy, May 2014, pp. 2563–2567.

[8] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2011.

[9] A. Mathur, S. M. Reddy, and R. M. Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–20, 2012.

[10] C.-Y. Yang, G. Brown, L. Lu, J. Yamagishi, and S. King, "Noise-robust whispered speech recognition using a non-audible-murmur microphone with vts compensation," in *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*. IEEE, 2012, pp. 220–223.

[11] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted." in *Proc. INTERSPEECH*, 2007, pp. 2289–2292.

[12] 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.

[13] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. International Conference on Spoken Language Processing*, Beijing China, 2000, pp. 145–148.

[14] P. K. Ghosh and S. S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL251–El257, Sep. 2011.

[15] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3, pp. 303–319, 2002.

[16] S.-C. S. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Interspeech*, 2006.

[17] S. Hahm, H. Daragh, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *Proc. the ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 47–54.

[18] 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, May 2011.

[19] S.-C. S. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005, pp. 1009–1012.

[20] ——, "Adaptation for soft whisper recognition using a throat microphone," in *Interspeech*, 2004.

[21] D. R. Beukelman, K. M., Yorkston, M. Poblete, and C. Naranjo, "Analysis of communication samples produced by adult communication aid users," *Journal of Speech and Hearing Disorders*, vol. 49, pp. 360–367, 1984.

[22] J. Wang, A. Samal, and J. Green, "Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph," in *Proc. ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, Baltimore, USA, 2014, pp. 38–45.

[23] J. Wang, J. Green, and A. Samal, "Individual articulator's contribution to phoneme production," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013, pp. 7785–7789.

[24] 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 *Proc. ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, pp. 79–85.

[25] 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*, vol. 59, pp. 15–26, 2016.

[26] J. Berry, "Accuracy of the ndi wave speech research sysetm," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 295–301, 2011.

[27] 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.

[28] M. Gales and S. Young, "The application of hidden markov models in speech recognition," *Foundations and trends in signal processing*, vol. 1, no. 3, pp. 195–304, 2008.

[29] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[30] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "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.

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

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

[33] 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. Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, 2013.

[34] ——, "Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping." in *Interspeech*, Lyon, France, 2013, pp. 1297–1301.

[35] 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. IEEE 2011 workshop on automatic speech recognition and understanding*, Waikoloa, USA, 2011, pp. 1–4.

[36] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 6645–6649.