Determining an Optimal Set of Flesh Points on Tongue, Lips, and Jaw for Continuous Silent Speech Recognition

Jun Wang[1,2], *Seongjun Hahm*[1], *Ted Mau*[3]

[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
{wangjun, seongjun.hahm}@utdallas.edu; ted.mau@utsouthwestern.edu

Abstract

Articulatory data have gained increasing interest in speech recognition with or without acoustic data. Electromagnetic articulograph (EMA) is one of the affordable, currently used techniques for tracking the movement of flesh points on articulators (e.g., tongue) during speech. Determining an optimal set of sensors is important for optimizing the clinical applications of EMA data, due to the inconvenience of attaching sensors on tongue and other intraoral articulators, particularly for patients with neurological diseases. A recent study found an optimal set (tongue tip and body back, upper and lower lips) on tongue and lips for isolated phoneme, word, or short phrase classification from articulatory movement data. This four-sensor set, however, has not been verified in continuous silent speech recognition. In this paper, we investigated the use of data from sensor combinations in continuous speech recognition to verify the finding using a publicly available data set MOCHA-TIMIT. The long-standing speech recognition approach Gaussian mixture model (GMM)-hidden Markov model (HMM) and a recently available approach deep neural network (DNN)-HMM were used as the recognizers. Experimental results confirmed that the four-sensor set is optimal out of the full set of sensors on tongue, lips, and jaw. Adding upper incisor and/or velum data further improved the recognition performance slightly.

Index Terms: silent speech recognition, deep neural network, hidden Markov model, electromagnetic articulograph, articulation, dysarthria

1. Introduction

With the availability of affordable devices for tongue movement data collection, articulatory data have obtained interest not only in speech science [1, 2, 3, 4] but also in speech technology (i.e., automatic speech recognition) [5, 6]. First, articulatory data have been successfully used to improve the speech recognition accuracy [5]. Articulatory data are particularly useful when speech signals are noisy or low quality [7] for recognizing dysarthric speech [8, 9]. Second, when acoustic data is not available, a silent speech interface (SSI) based on articulatory data has potential clinical applications [10, 11]. An SSI recognizes speech from articulatory data only (without using audio data) [12, 13] and then drives a text-to-speech synthesizer for sound playback [14, 15]. For example, SSIs can be used to assist the oral communication for patients with severe voice disorders or without the ability to produce speech sounds (e.g., due to laryngectomy, a surgical removal of larynx due to treatment of laryngeal cancer) [16]. There are currently limited options to assist speech communication for those individuals (e.g., esophageal speech, tracheo-esophageal speech or tracheo-esophageal puncture (TEP) speech, and electrolarynx). These approaches, however, produce an abnormal sounding voice [17, 18], which impacts the quality of life of laryngectomees. Current text-to-speech technologies have been able to produce speech with natural sounding voice for SSIs [19]. One of the current challenges of SSI development is silent speech recognition algorithms (without using audio data) [10, 20] or mapping articulatory information to speech [21, 22, 23].

Electromagnetic motion tracking is one of the affordable, currently used technologies for tracking tongue movement during speech [19, 24, 25]. There are currently two commercially available devices, EMA AG series (by Carstens) and Wave system (by NDI, Inc.) [26]. Tongue tracking using electromagnetic devices is accomplished through attaching small sensors on the surface of tongue and other articulators. In prior work, the number of tongue sensors and their locations have been justified based on long-standing assumptions about tongue movement patterns in classic phonetics [27], or the specific purpose of the study. Other techniques that have been used to record non-audio articulatory information include ultrasound [28, 29], and surface electromyography (EMG) [30, 31].

Determining an optimal set of tongue sensors for speech production is significant for both science and technology. Scientifically, determining an optimal set of sensors can improve the understanding of the coordination of articulators for speech production [32]. Technologically, it can be helpful for clinical applications including (1) silent speech interfaces, (2) speech recognition with articulatory information [5, 33], and (3) speech training using real-time visual feedback of tongue movements [34, 35]. In literature, three or four EMA sensors on the tongue have been commonly used (e.g., [1, 3, 4, 5, 36, 37]). The use of more sensors than necessary comes at a cost for both researchers and subjects; the procedure for attaching sensors to the tongue is time intensive and can cause discomfort and therefore may limit the scope of EMA for practical use, particularly for persons with neurological diseases (e.g., Parkinson's disease [38] and amyotrophic lateral sclerosis [39]).

Here, *optimal* set means a sensor set that contains the least number of sensors that performs no worse than other sets with more sensors. There may be more than one optimal set with the same number of sensors.

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 79–85,
Dresden, Germany, 11 September, 2015.

Until recently, a study found two tongue sensors (Tongue Tip and Tongue Body Back) and two lip sensors (Upper Lip and Lower Lip) are optimal for isolated phoneme (vowels and consonants), word, and short phrase classification [32, 40]. The classification results based on data using the optimal set were not significantly different from these based on data from the full set with four tongue sensors (Tongue Tip, Tongue Blade, Tongue Body Front, and Tongue Body Back) plus the two lip sensors [32]. However, this set has not been verified in continuous silent speech recognition or speech recognition from both acoustic and articulatory data. If the two-tongue-sensor set can be confirmed for continuous speech recognition, it would be beneficial for future collection of a larger articulatory data set. Other studies compared the whole tongue and lips (e.g., [41] using ultrasound and optical data), but not on flesh points.

In this paper, we investigated the optimal set of tongue sensors for speaker-dependent continuous silent speech recognition (using articulatory data only) and speech recognition (using combined acoustic and articulatory data). The goals were (1) to confirm if more than two tongue sensors are unnecessary for continuous silent speech recognition and speech recognition using both acoustic and articulatory data when only tongue and lips are used, and (2) to provide a reference for choosing the number of sensors and their locations on the tongue, lips, jaw and other articulators for future studies. However, due to the space limitation, this paper did not verify if the hypothesized optimal four-sensor set is unique. The articulatory and acoustic data in the MOCHA-TIMIT data set [42] were used in this experiment. The MOCHA-TIMIT data set is appropriate for this study because it contains data collected from sensors attached on multiple articulators, including three sensors on the tongue, two on the lips, two on the incisors, and one on the velum. In addition, both MOCHA-TIMIT and the data set in [32] have tongue tip and body back (or dorsum). Thus the first goal of this paper became to verify if the tongue blade sensor is unnecessary in addition to the hypothesized optimal set [32, 40]. The traditional speech recognition approach Gaussian mixture model (GMM)-hidden Markov model (HMM) [5] and a recently available and promising approach deep neural network (DNN)-HMM [43, 44] were used.

2. Method

2.1. Data set

MOCHA (Multi-CHannel Articulatory)-TIMIT data set consists of simultaneous recordings of speech, articulatory movement, and other forms of data collected from 2 British English speakers (1 male - MSAK0 and 1 female - FSEW0) [42]. There are 920 sentences (extracted from TIMIT database) in total. Individual phonemes and silences within each sentence have been labeled.

The articulatory and acoustic data in MOCHA-TIMIT were collected using an Electromagnetic Articulograph (EMA, Carstens Medizinelektronik GmbH, Germany) by attaching sensors to upper lip (UL), lower lip (LL), upper incisor (UI), lower incisor (LI), tongue tip (TT), tongue blade (TB), tongue dorsum (TD), and velum (V) with 500 Hz sampling rate. Each sensor had x (front-back) and y (vertical) trajectories. Therefore, the acoustic data and the 16-dimensional x and y motion data obtained from UI, LI, V, UL, LL, TT, TB, and TD were used.

TT was 5-10 mm to the tongue apex; TB was 2-3 cm from TT; TD was 2-3 cm from TB [42]. This roughly matched with the tongue tip sensor in [32, 40], which was also 5-10 mm to tongue apex, and the tongue body back in [32, 40], which was about 40 mm from tongue tip. Thus, as mentioned earlier, the goal (1) in this paper became to verify if the middle tongue sensor (TB) was unnecessary.

2.2. Recognizers

A long-standing approach GMM-HMM and a promising approach DNN-HMM were used as the recognizers in this experiment.

2.2.1. Gaussian Mixture Model-Hidden Markov Model

GMM-HMM has been used in speech recognition for decades [45]. The core idea of GMM is compact representation of distribution using means and variances. GMM is a generative model and trained to represent as closely as possible the distribution (e.g., using means and variances) of training data. In many applications, the number of mixtures for GMMs is adjusted to avoid overfitting.

2.2.2. Deep Neural Network-Hidden Markov Model

DNN-HMM recently attracted the interests of speech recognition researchers because it showed a significant performance improvement compared with GMM-HMM when replacing GMM to DNN in (acoustic) speech recognition [44, 46]. We adopted the DNN training approach based on restricted Boltzmann machines (RBMs) [47].

The DNN (stacked RBMs) were subsequently fine-tuned using the backpropagation algorithm. A detailed explanation and discussion of the DNN can be found in [47, 48].

2.3. Experimental setup

Data from individual sensors or combinations of sensors were used in speech recognition experiments (from articulatory data only or from combined acoustic and articulatory data). The recognition performances obtained from individual sensors or their combinations were compared to determine (1) if Tongue Blade was unnecessary in addition to the other two tongue sensors and lips (Tongue Tip, Tongue Dorsum, Upper Lip, and Lower Lip), and (2) if the performance improved when more sensor's data (e.g., upper incisor and velum) were added.

In each experiment, a 5-fold cross validation strategy with a jackknife procedure was performed to set training and test sets in the experiment [42, 49]. In each of the five executions, a group of 92 sentences were selected for test with the remaining 368 sentences for training. Due to the high degree of variation in the articulation across speakers and there were only two speakers in MOCHA-TIMIT, speaker-dependent recognition was conducted. The average training data length for each cross validation became 21.3 mins (368 sentences) for the female speaker and 20.6 mins (368 sentences) for the male speaker. The average test data length along 5 cross validations was 5.3 mins (92 sentences) for the female speaker and 5.2 mins (92 sentences) for the male speaker, respectively.

Articulatory features were extracted from the corpus using EMAtools [50]. The original articulatory features and their first and second derivatives were concatenated to build various dimensional feature vectors for each set of sensors. The "breath" segments were merged with "silence" for both training and testing [49]. The input features in DNN were a concatenation of articulatory feature vectors (number of sensors $\times$ 2-dimension articulatory movement data + Δ + $\Delta\Delta$) with 9

Table 1: *Experimental setup.*

Articulatory Feature	
Low pass filtering	40 Hz cutoff, 5th order Butterworth
Sampling rate	100 Hz (downsampled from 500 Hz)
Feature vector	articulatory movement vector + Δ + $\Delta\Delta$ (e.g., 6 dim. for 1 sensor, 48 dim. for 8 sensors)
Acoustic Feature	
Sampling rate	16 kHz
Feature vector	MFCC vector (13 dim.) + Δ + $\Delta\Delta$ (39 dim.)
Frame size	25 ms
Common	
Frame rate	10 ms
Mean normalization	Applied
GMM-HMM topology	
Monophone	context-independent 137 states (44 phones $\times$ 3 states, 5 states for silence), $\approx$ 14 mixtures 3-state left to right HMM
Training method	Maximum likelihood estimation
DNN-HMM topology	
Monophone	context-independent input layer dimension varies based on the set of sensors (e.g., 54 for 1 sensor, 432 for 8 sensors) 137 output layer dimension (including 5 outputs for silence) 1,024 nodes for each hidden layer 1 to 6-depth hidden layers
Training method	RBM pre-training, back-propagation
Language model	bi-gram phoneme language model

frames (4 preceding, current, and 4 succeeding frames). As it concatenates multiple feature vectors in the time domain, DNN has time-dependent context information which HMM takes using multiple states [43, 51]. Mel-frequency cepstral coefficients (MFCCs) were extracted from the acoustic data and used as the acoustic features in the recognition experiments.

The GMM-HMM system was trained using maximum likelihood estimation (MLE) without using segment information provided in MOCHA-TIMIT corpus (flat initialization). The DNN-HMM system was pre-trained using contrastive-divergence algorithm on RBMs and fine-tuned using back-propagation algorithm. A bi-gram phoneme language model was trained using all 44 phonemes provided in label files of the corpus.

Table 1 lists the details of the experimental setup and major parameters in GMM-HMM and DNN-HMM. The training and decoding were performed using the Kaldi speech recognition toolkit [52].

A phoneme error rate (PER) was used as a performance measure, which is the ratio of the sum of the number of errors over the total number of phonemes. The PER is represented by

$$\text{PER} = (S + D + I)/N \quad (1)$$

where S represents the number of substitution errors, D is the number of deletion errors, I stands for the number of insertion errors, and N is the total number of phonemes in the test set. For DNN, we conducted experiments using 1 to 6 hidden layers and the best performance was reported. Finally, the PERs from each test group in the 5-fold cross validation were averaged as the overall PER.

3. Results and Discussion

Experimental results are shown in Figures 1 to 4 and discussed below. Figures 1 and 2 show the silent speech recognition performance on individual or combinations of sensors for both speakers using GMM-HMM or DNN-HMM, respectively. Figures 3 and 4 give the speech recognition from MFCCs plus individual or combinations of sensors' data using GMM-HMM and DNN-HMM, respectively.

3.1. General observations

First, the recognition performances obtained from individual sensor's data had consistently lower performance (higher PERs) than from the combinations of sensors (Figures 1 to 4). Although it seems intuitive, to our knowledge, this is the first time the individual EMA sensor's performance were examined in continuous silent speech recognition or speech recognition from combined acoustic and articulatory data.

Second, when the performances obtained using data from individual sensors were compared, upper incisor (UI) and velum (V) had the worst performance; the three individual tongue sensors had a similar performance and were the best among all sensors; lip sensors were between the tongue sensors (TT, TB, TD) and UI and velum (V). This finding is highly consistent with the descriptive knowledge in classic phonetics that tongue is the primary articulator [27].

3.2. {TT, TD, UL, LL} and other combinations

Silent speech recognition performance substantially degraded if any of the sensor in previously found optimal four-sensor set (i.e., TT, TD, UL, and LL, marked bold in Figures 1 and 2) was not used [32]. The optimal set of sensors using GMM-HMM and articulatory data yielded a PER of 42.0% and 40.9% for the female and male speakers, respectively. DNN-HMM with this optimal set yielded a PER of 38.2% and 36.5% for the female and male speakers, respectively.

As TB, UI, LI (jaw), or all of the three sensors' data were added on top of the four-sensor set, there was no improvement using GMM-HMM, but a slight improvement using DNN-HMM. When using all sensors' (including V) data together, a substantial improvement was obtained using either GMM-HMM or DNN-HMM.

These results suggest the four-sensor set ({TT, TD, UL, LL}) was an optimal set for silent speech recognition out of the full set of sensors on the tongue, lips, and jaw. However, adding extra data source (e.g., UI and V) could still improve the performance.

Speech recognition from combined acoustic and articulatory data (Figures 3 and 4) also substantially degraded if any of the sensor in {TT, TD, UL, and LL} was missing, for recognizers. However, GMM-HMM and DNN-HMM results showed different patterns when adding more sensors data to {TT, TD, UL, LL}. GMM-HMM showed no improvement to the optimal set (23.0% for female and 22.6% for male) when adding more sensor's data (22.7% for female and 22.8% for male); while DNN-HMM (19.7% for female and 19.5% for male) showed significant error reduction compared to the optimal set (20.4% for female and 20.3% for male). This observation suggests DNN has more potential than GMM to incorporate more data sources to further improve the recognition performance.

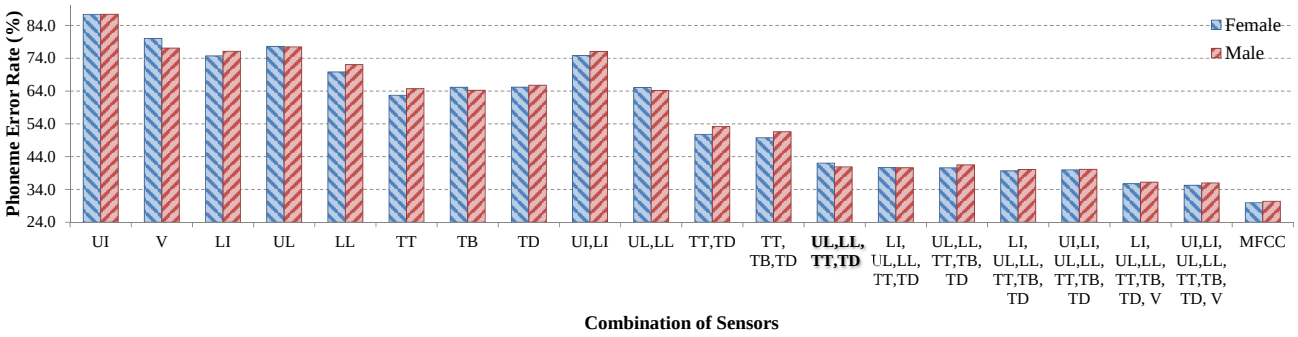

Figure 1: *Phoneme Error Rates (PER; %) obtained using GMM-HMM and articulatory features.*

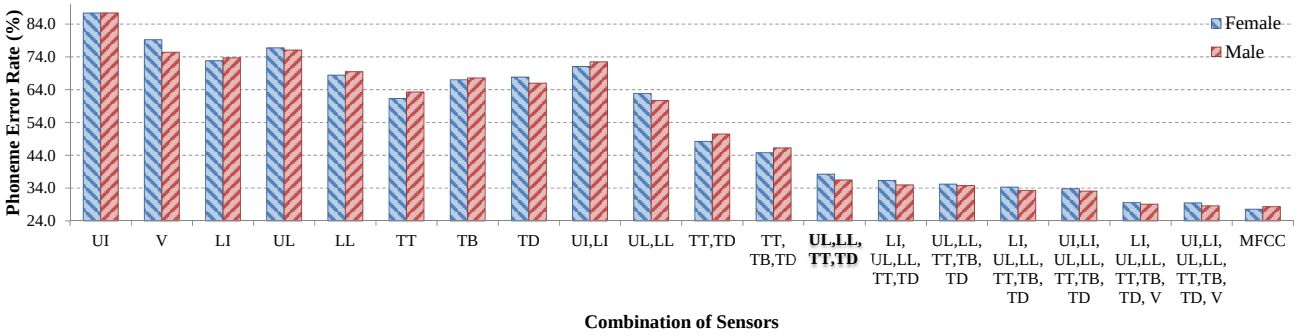

Figure 2: *Phoneme Error Rates (PER; %) obtained using DNN-HMM and articulatory features.*

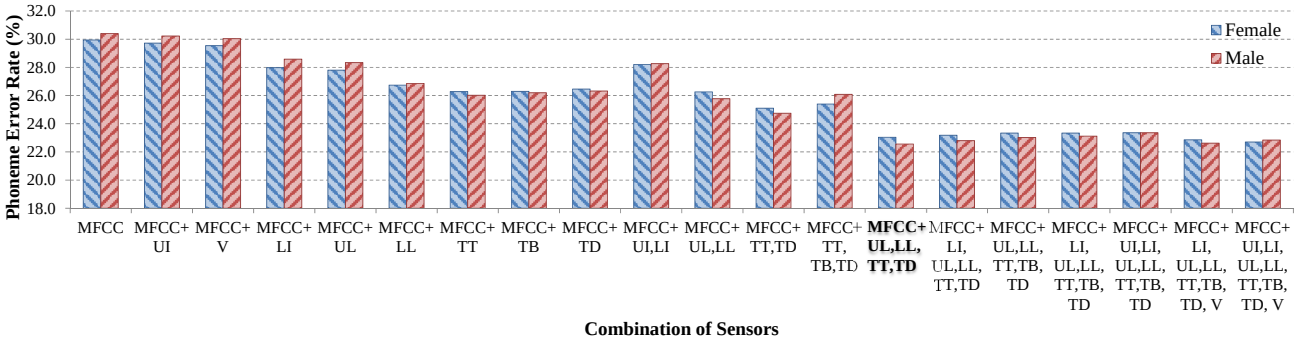

Figure 3: *Phoneme Error Rates (PER; %) obtained using GMM-HMM and combined articulatory and acoustic features.*

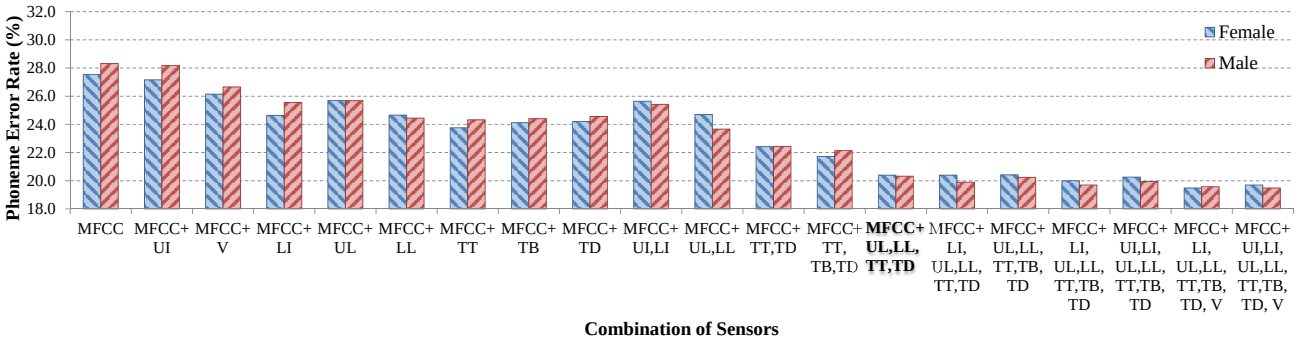

Figure 4: *Phoneme Error Rates (PER; %) obtained using DNN-HMM and combined articulatory and acoustic features.*

The most important conclusion from the results above may be, for future studies in which data are collected only from tongue, lips, or jaw (i.e. not from velum), {TT, TD, UL, LL} is an optimal set for silent speech recognition or speech recognition from combined acoustic and articulatory data. However, adding upper incisor and/or velum data can still further improve the performance slightly.

3.3. {TT, TD, UL, LL} vs {TT, TB, TD, UL, LL}

Table 2 lists the results obtained from {TT, TD, UL, LL} and {TT, TB, TD, UL, LL} to provide a close-up performance comparison of the two sets, which further confirms adding TB

Table 2: *Phoneme Error Rates (PER; %) obtained from sensor combination {TT, TD, UL, LL} and {TT, TB, TD, UL, LL}.*

Speaker	Model	Feature	Combination of Sensors		Performance
			TT,TD,UL,LL	TT,TB,TD,UL,LL	Difference
Female	GMM-HMM	EMA	42.04	40.60	+1.44
		MFCC + EMA	23.04	23.34	-0.30
	DNN-HMM	EMA	38.24	35.20	+3.04
		MFCC + EMA	20.40	20.42	-0.02
Male	GMM-HMM	EMA	40.88	41.48	-0.60
		MFCC + EMA	22.56	23.02	-0.46
	DNN-HMM	EMA	36.46	34.74	+1.72
		MFCC + EMA	20.32	20.24	+0.08
Average					+0.61

(Tongue Blade) did not significantly improve the speech recognition performance in addition to {TT, TD, UL, LL}. The right-most column of Table 2 lists the performance difference between {TT, TD, UL, LL} and {TT, TB, TD, UL, LL} (positive means a better performance with TB; negative means worse performance). The average performance difference of the two sensor sets in all eight configurations (female vs male speaker, GMM vs DNN, with or without MFCC) was +0.61, which means adding TB reduced only 0.61% of PER.

3.4. {TT, TD, UL, LL} may not be the only four-sensor optimal set

The four-sensor set ({TT, TD, UL, LL}) may be just one of the possible optimal four-sensor sets, because of the high coupling of adjacent parts [3]. Figures 1 to 4 also show the three tongue sensors, TT (Tongue Tip), TD (Tongue Dorsum) and TB (Tongue Blade) have no significant differences in performance when used individually, which may suggest they are interchangeable. In other words, any two tongue sensors may achieve no significant difference in recognition performance with {TT, TD}. A further analysis using data from all tongue sensor pairs is needed to test this hypothesis.

Nevertheless, we still suggest {TT, TD} as the optimal tongue sensor pair, since TT and TD are anatomically farther apart from each other than other tongue sensor pairs, thus TT and TD may be more independent and have less redundant information. In addition, from the user's (subject) perspective, the sensor location on the tongue may not matter, as long as they are in the comfortable zone (from tongue tip to tongue body back).

3.5. Velum sensor

Adding velum (V) data in addition to other sensors always improved the speech recognition performance, although velum in isolation achieved the worse performance. Velum is the primary articulator for controlling nasal sounds in English (e.g., /m/ and /n/). Velum provides unique information that other articulators do not. However, we still do not think attaching sensors on the velum is suitable for practical use of EMA, considering the trade-off of the discomfort of attaching velum sensor on subjects and the slight improvement of recognition performance.

3.6. DNN-HMM outperformed GMM-HMM

DNN-HMM outperformed GMM-HMM in all experimental configurations (Figures 1 to 4). Although the focus of this paper was not comparing GMM-HMM and DNN-HMM, the results indicate the DNN-HMM outperformed GMM-HMM in both silent speech recognition and speech recognition from combined acoustic and articulatory data. This finding is consistent with the recent literature in silent speech recognition [53], acoustic speech recognition [44, 48], and speech recognition from combined acoustic and articulatory data [46, 54]. We expect DNN-HMM has potential to further improve the recognition performance from articulatory data or from combined acoustic and articulatory data with a better structure or when combined with other approaches (e.g., speaker adaptation [55]).

4. Conclusions and Future Work

In this paper, we have confirmed a previously found optimal set of sensors on the tongue and lips (Tongue Tip, Tongue Dorsum, Upper Lip and Lower Lip) [32] through experiments with continuous silent speech recognition and speech recognition from combined acoustic and articulatory data, when only tongue, lips, upper incisor, and lower incisor data are available (i.e., no velum data). Although velum data can further (slightly) improve the recognition performance on top of the four-sensor set, it is not recommended for practical use because it causes discomfort for subjects. In addition, the four-sensor set may not be unique, since the individual tongue sensors have no significant accuracy difference. Finally, DNN-HMM outperformed GMM-HMM in both silent speech recognition and speech recognition from combined acoustic and articulatory data.

These findings provide a reference for future relevant studies on choosing the number of sensors and their locations on the tongue. However, as mentioned earlier, determining an appropriate set of sensors may depend on the specific purpose of the study. For example, a sensor on the side of the tongue may be used in studies that focus on lateral tongue curvature during speech production [56, 57].

Future work includes (1) verifying if TT, TB, and TD are interchangeable, or determining if {TT, TD, UL, LL} is the unique four-sensor optimal set, and (2) sensor combinations in speaker-independent silent speech recognition experiments [58, 59, 54].

5. Acknowledgment

This work was supported by the National Institutes of Health (NIH) through grants R03 DC013990 and R01 DC013547. We would like to thank Dr. Jordan R. Green, Dr. Ashok Samal, and the support from the Communication Technology Center, University of Texas at Dallas.

6. References

[1] J. S. Perkell and W. L. Nelson, "Variability in production of the vowels /i/ and /a/," *The Journal of the Acoustical Society of America*, vol. 77, no. 4, pp. 1889–1895, 1985.

[2] J. Westbury, "X-ray microbeam speech production database users handbook," *University of Wisconsin*, 1994.

[3] 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.

[4] 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.

[5] 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.

[6] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magami-Doss, and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 jhu summer workshop," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, 2007, pp. IV–621–IV–624.

[7] K. Kirchhoff, G. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319, 2012.

[8] F. Rudzicz, "Using articulatory likelihoods in the recognition of dysarthric speech," *Speech Communication*, vol. 54, no. 3, pp. 430 – 444, 2012.

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

[10] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[11] J. Wang, A. Samal, J. Green, and T. Carrell, "Vowel recognition from articulatory position time-series data," in *Proc. of International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2009, pp. 1–6.

[12] J. Wang, A. Balasubramanian, L. Mojica de la Vega, J. Green, A. Samal, and B. Prabhakaran, "Word recognition from continuous articulatory movement time-series data using symbolic representations," in *Proc. of Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Grenoble, France, 2013, pp. 119–127.

[13] 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.

[14] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe, "Recent improvements on microsoft's trainable text-to-speech system-whistler," in *Proc. of ICASSP*, vol. 2, Munich, Germany, 1997, pp. 959–962.

[15] S. Manitsaris, B. Denby, F. Xavier, J. Cai, M. Stone, P. Roussel, and G.Dreyfus, "An open source speech synthesis module for a visual-speech recognition system," in *Proc. of Acoustics*, Nates, France, 2012, pp. 3937–3941.

[16] B. Bailey, J. Johnson, and S. Newlands, *Head & neck surgery–otolaryngology*. Lippincott Williams & Wilkins, 2006, vol. 1.

[17] H. Liu and M. Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332, 2007.

[18] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham, "Reconstructing the Voice of an Individual Following Laryngectomy," *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–66, 2011.

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

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

[21] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech and Language*, 2015.

[22] M. W. Marlene Zahner, Matthias Janke and T. Schultz, "Conversion from facial myoelectric signals to speech: A unit selection approach," in *Proc. of INTERSPEECH*, 2013, pp. 1331–1335.

[23] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech and Language*, 2015.

[24] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical engineering & physics*, vol. 30, no. 4, pp. 419–425, 2008.

[25] R. Hofe, J. Bai, L. A. Cheah, S. R. Ell, J. M. Gilbert, R. K. Moore, and P. D. Green, "Performance of the MVOCA silent speech interface across multiple speakers," in *Proc. of INTERSPEECH*, 2013, pp. 1140–1143.

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

[27] P. Ladefoged and K. Johnson, *A Course in Phonetics, 6th Edition*. Wadsworth Cengage Learning, 2011.

[28] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[29] B. Denby, J. Cai, P. Roussel, G. Dreyfus, L. Crevier-Buchman, C. Pillot-Loiseau, T. Hueber, G. Chollet *et al.*, "Tests of an interactive, phrasebook-style, post-laryngectomy voice-replacement system," in *Proc. of ICPhS XVII*, Hong Kong, 2011, pp. 572–575.

[30] C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Communication*, vol. 52, no. 4, pp. 354–366, 2010.

[31] Y. Deng, J. Heaton, and G. Meltzner, "Towards a practical silent speech recognition system," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 1164–1168.

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

[33] F. Rudzicz, "Correcting errors in speech recognition with articulatory dynamics," in *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 60–68.

[34] W. Katz, T. F. Campbell, J. Wang, E. Farrar, C. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "Opti-speech: A real-time, 3d visual feedback system for speech training," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 1174–1178.

[35] P. Badin, A. B. Youssef, G. Bailly, F. Elisei, and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," in *Proc. of SLATE workshop*, 2010.

[36] 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.

[37] F. H. Guenther, C. Y. Espy-wilson, S. E. Boyce, M. L. Matthies, M. Zandipour, J. S. Perkell, P. Frank, and H. Guenther, "Articulatory tradeoffs reduce acoustic variability during american english /r/ production," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2854–65, 1999.

[38] S. Hahm and J. Wang, "Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data," in *Proc. of INTERSPEECH*, 2015.

[39] 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.

[40] 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.

[41] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Phone recognition from ultrasound and optical video sequences for a silent speech interface," in *Proc. of INTERSPEECH*, 2008, pp. 2032–2035.

[42] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. of ICSLP*, Beijing China, 2000, pp. 145–148.

[43] 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.

[44] 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.

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

[46] 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.

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

[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] E. Uraga and T. Hain, "Automatic speech recognition experiments with articulatory data," in *Proc. of INTERSPEECH*, Pittsburgh, USA, 2006, pp. 353–356.

[50] N. Nguyen, "A MATLAB toolbox for the analysis of articulatory data in the production of speech," *Behavior Research Methods, Instruments, & Computers*, vol. 32, no. 3, pp. 464–467, 2000.

[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.

[52] 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.

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

[54] S. Hahm, D. Heitzman, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015.

[55] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7947–7951.

[56] J. Wang, W. Katz, and T. F. Campbell, "Contribution of tongue lateral to consonant production," in *Proc. of INTERSPEECH*, Singapore, 2014, pp. 174–178.

[57] A. Ji, J. Berry, and M. Johnson, "The electromagnetic articulography mandarin accented english (ema-mae) corpus of acoustic and 3d articulatory kinematic data," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 7719–7723.

[58] 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.

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