SILENT SPEECH RECOGNITION FROM ARTICULATORY MOVEMENTS USING DEEP NEURAL NETWORK

Seongjun Hahm[1], Jun Wang[1,2,3]

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

ABSTRACT

Laryngectomee patients lose their ability to produce speech sounds and suffer in their daily communication. There are currently limited communication options for these patients. Silent speech interfaces (SSIs), which recognize speech from articulatory information (i.e., without using audio information), have potential to assist the oral communication of persons with laryngectomy or other speech or voice disorders. One of the challenging problems in SSI development is to accurately recognize speech from articulatory data. Deep neural network (DNN)-hidden Markov model (HMM) has recently been successfully used in (acoustic) speech recognition, which shows significant improvements over the long-standing approach Gaussian mixture model (GMM)-HMM. DNN-HMM, however, has rarely been used in silent speech recognition. This paper investigated the use of DNN-HMM in recognizing speech from articulatory movement data. The articulatory data in the MOCHA-TIMIT data set was used in the experiment. Results indicated the performance improvement of DNN-HMM over GMM-HMM in silent speech recognition.

Keywords: silent speech recognition, articulatory movements, deep neural network, hidden Markov model

1. INTRODUCTION

People with impaired speech rely on assistive devices for their daily communication. For example, persons after laryngectomy (a surgical removal of larynx due to the treatment of cancer) or with neurological speech disorders lose their ability to speak and suffer in their daily communication [2]. Although there are currently several options to assist the speech communication for laryngectomees (i.e., esophageal speech, tracheo-esophageal speech, and electrolarynx), these approaches frequently produce an abnormal sounding voice with a pitch that is aberrantly low and limited in range [1,20]. An alternative technology with natural output voice is highly needed.

Silent speech interfaces (SSIs) have potential to provide an alternative way to assist those patients to produce speech with natural sounding voice [6]. A number of techniques have been used to record non-audio articulatory information such as ultrasound [5, 16], surface electromyography (EMG) [8, 17], and electromagnetic articulograph (EMA) [9, 32]. To accomplish the mapping from recorded articulatory movements to speech sounds, an SSI combines two technologies: silent speech recognition [28,33] and text-to-speech synthesis [14,22]. Silent speech recognition recognizes words or sentences from articulatory movements; text-to-speech synthesis then plays synthesized sounds based on the recognized text, which is ready for this application [32]. SSIs have even potential to use the patient's own voice (recorded pre-surgery) to drive the speech output [1]. Thus, the current research focuses on the development of accurate silent speech recognition algorithms.

In the past decades, various machine learning techniques have been successfully used for speech recognition from acoustic data, articulatory data, or combined, including the traditional Gaussian mixture model (GMM)-hidden Markov model (HMM) [18, 27, 34, 35]. Other approaches include multi-stream HMM [11, 15], support vector machine [28, 33], neural network [25], dynamic Bayesian network [26], and subspace GMM (SGMM) [8].

Deep neural network (DNN)-HMM has recently been applied in (acoustic) speech recognition [7, 21], which shows significant improvements over the long standing approach GMM-HMM. DNN-HMM has been adopted in commercial speech recognition systems (e.g., Google voice, Apple Siri) [12]. The promise of DNN-HMM to improve the acoustic speech recognition accuracy motivated the application of DNN in silent speech recognition.

DNN-HMM, however, has rarely been used in silent speech recognition (i.e., without acoustic information). Canevari and colleagues used DNN-HMM for speech recognition with combined acoustic and articulatory features, which showed performance improvement over GMM-HMM [3, 4]. Their experiments did not include a comparison of DNN-HMM and GMM-HMM for silent speech recognition from articulatory features only. Thus, it remains unknown whether DNN-HMM outperforms GMM-HMM in silent speech recognition.

This paper investigated the use of DNN-HMM in silent speech recognition from articulatory movements data (i.e., without using acoustic data). The performance of DNN-HMM was compared with the traditional approach, GMM-HMM. The articulatory movement data in the MOCHA-TIMIT data set [34] was used in the experiment. The MOCHA-TIMIT contains 2-dimensional

(vertical and anterior-posterior) movement of sensors attached to the tongue, lips, and other articulators of two speakers. DNNs with different number of hidden layers were also tested and the results were reported.

2. DEEP NEURAL NETWORK

Speech pattern recognition problem is of finding (deciding) appropriate word sequences based on speech (or articulatory) data. This procedure can be represented using Bayes theorem as follows:

$$\begin{aligned}\hat{\boldsymbol{W}} &= \underset{\boldsymbol{W}}{\operatorname{argmax}} p(\boldsymbol{W}|\boldsymbol{O}) \\ &= \underset{\boldsymbol{W}}{\operatorname{argmax}} \frac{p(\boldsymbol{O}|\boldsymbol{W})p(\boldsymbol{W})}{p(\boldsymbol{O})} \\ &\propto \underset{\boldsymbol{W}}{\operatorname{argmax}} p(\boldsymbol{O}|\boldsymbol{W})p(\boldsymbol{W})\end{aligned} \tag{1}$$

where $\boldsymbol{W}$ is word sequences, $\boldsymbol{O}$ is observation vectors (can be either speech waveforms or articulatory movements; ideally can be any kind of data representing speech characteristics), $\hat{\boldsymbol{W}}$ is the optimal word sequences, and $p(\boldsymbol{W})$ is a prior probability obtained from a language model. In Eq. 1, $p(\boldsymbol{O}|\boldsymbol{W})$ is a posterior probability that can be obtained from the model trained in advance. This model could be either GMM-HMM or DNN-HMM. In many tasks, DNN-HMM showed the significant performance improvement compared with GMM-HMM as replacing GMM to DNN [12,21]. We adopt the DNN training approach based on restricted Boltzmann machines (RBMs) [13].

A probability to visible and hidden vector pair is represented by following energy function.

$$p(\boldsymbol{v},\boldsymbol{h}) = \frac{\exp\{-E(\boldsymbol{v},\boldsymbol{h})\}}{\sum_{\boldsymbol{v}}\exp\{-E(\boldsymbol{v},\boldsymbol{h})\}} \tag{2}$$

where $\boldsymbol{v}$ and $\boldsymbol{h}$ are the binary state vectors of visible and hidden unit, respectively. An energy function for joint vector pair, $(\boldsymbol{v},\boldsymbol{h})$, of the visible and hidden units are given by

$$E(\boldsymbol{v},\boldsymbol{h}) = -\boldsymbol{a}'\boldsymbol{v} - \boldsymbol{b}'\boldsymbol{h} - \boldsymbol{v}'\boldsymbol{C}\boldsymbol{h} \tag{3}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the bias vectors and $\boldsymbol{C}$ is the weight matrix between $\boldsymbol{a}$ and $\boldsymbol{b}$. RBM pre-training of each layer is usually performed by contrastive divergence based on maximum likelihood criterion [12,13,21]

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \prod_{t}\sum_{\boldsymbol{h}} p(\boldsymbol{o}_t,\boldsymbol{h}|\boldsymbol{\Theta}) \tag{4}$$

where $\boldsymbol{\Theta} \triangleq \{\boldsymbol{C},\boldsymbol{a},\boldsymbol{b}\}$ and $\boldsymbol{o}_t$ is the observation vector at frame (time) t.

The stacked RBMs (DNN) are trained in iterative manner. The trained DNN are subsequently fine-tuned using backpropagation algorithm. A detailed explanation and further discussion of the DNN-HMM can be found in [12,13,21].

3. EXPERIMENTAL DESIGN

3.1. Data set

MOCHA (Multi-CHannel Articulatory)-TIMIT Data set was used in this project [34]. MOCHA-TIMIT data set has 920 sentences (extracted from TIMIT database) from 2 British English speakers (1 male - MSAK0 and 1 female - FSEW0). This data set consisted of simultaneous recordings of speech, articulatory movement data, and other forms of data.

Only the articulatory movement data was used in the experiment. The articulatory data was 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 with 500 Hz sampling rate. Each sensor has x (front-back) and y (vertical) trajectories. At this stage, we used only the data from tongue and lips, the primary articulators. Data from upper incisor, lower incisor, and velum will be used in future experiments. Therefore, 10-dimensional x and y motion data obtained from UL, LL, TT, TB, and TD were used.

3.2. Experimental Setup

A 5-fold cross validation strategy with a jackknife procedure was performed to set training and test sets in the experiment [27, 34]. In each of the five executions, a group of 92 sentences was selected for test with the remained 368 sentences for training. Due to the high degree of variation in the articulation across speakers, speaker-dependent recognition was conducted. The average training data length for each cross validation becomes 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 the five cross validations is 5.3 mins (92 sentences) for the female speaker and 5.2 mins (92 sentences) for the male speaker, respectively.

Articulatory features from the corpus were extracted using EMAtools [23]. The original articulatory features and their first and second derivatives were concatenated to build 30-dimensional feature vectors. The "breath" segments were merged with "silence" for both training and testing [27]. For DNN-HMM input features, the original features were concatenated to create 270-dimensional feature vector (9×30 articulatory movements vector) with 4 preceding, current, and 4 succeeding frames. As concatenating multiple feature vectors in time domain, DNN-HMM has time-dependent context information which GMM-HMM takes using multiple states [21]. The GMM-HMM system was trained using maximum likelihood estimation (MLE). The DNN-HMM systems were 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 and used to construct the final weighted finite state transducer (WFST). Table 1 shows the detailed experimental setup. The training and decoding were per-

Table 1: Experimental setup.

Feature	
Low pass filtering [30]	40 Hz cutoff 5th order Butterworth
Sampling rate	100 Hz (down sampled from 500 Hz)
Feature vector	articulatory movement vector + Δ + ΔΔ (30 dimensions)
Frame Length	25 ms
Frame rate	10 ms
Mean normalization	Applied
GMM-HMM topology	
Monophone	context-independent 137 states (44 phones × 3 states, 5 states for silence), ≈ 14 mixtures 3-state left to right HMM
Training method	Maximum likelihood estimation (MLE)
DNN-HMM topology	
Monophone	context-independent 270 input layer dimension 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

formed using the Kaldi speech recognition toolkit [24].

Phoneme error rate (PER) was used as a performance measure, which is the ratio of the sum of the number of substitution, deletion, and insertion errors over the total number of phonemes. PER is given by

$$\text{PER} = \frac{S+D+I}{N} \quad (5)$$

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. Finally, PERs for each test group in the five cross validations were averaged as the overall PER.

4. RESULTS AND DISCUSSION

The experimental results are shown in Fig. 1. DNN-HMM with different number of hidden layers outperformed GMM-HMM for both speakers, which is encouraging. A two-sample t-test was used to check the significance between the results obtained using GMM-HMM and DNN-HMM. The significances were marked in Fig. 1. For the male speaker (MSAK0), 5-layer DNN-HMM had the best performance (lowest PER), 35.5% (5.4% absolute PER reduction), whereas GMM-HMM had 40.9% PER. For the female speaker (FSEW0), 5-layer DNN-HMM had the best performance, 37.1% PER (4.5% absolute PER reduction), whereas GMM-HMM had 41.5% PER.

In [4], when only audio data was used, PER was reduced from 38.0% (using GMM-HMM) to 32.2% (5.8% absolute PER reduction) using DNN-HMM. When using both acoustic and articulatory data, DNN-HMM improved performance from 32.2% (acoustic features only) to 23.7% (combined features; 8.5% absolute PER reduction). Although higher PERs were expected in our experiments, where only articulatory features were used, a similar level of improvement of DNN-HMM over GMM-HMM was also observed.

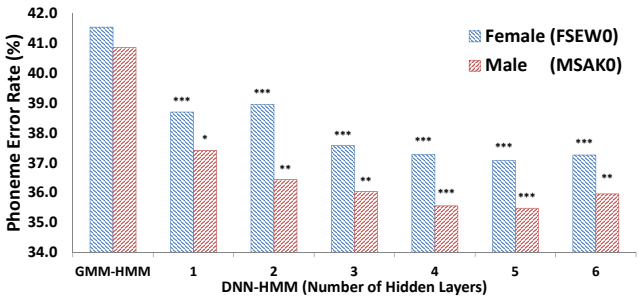

Figure 1: Phoneme Error Rate (PER; %) of GMM-HMM and DNN-HMM experiment. Significances between the results obtained using GMM-HMM and DNN-HMM with different number of hidden layers are marked: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The performance of DNN-HMM was consistently improved as the number of hidden layers was increased. As shown in Fig. 1, the performance reached the best when the number of hidden layers is 5 for both female and male speakers. Even though MOCHA-TIMIT is a relatively small data set, the results showed similar trends with the previous acoustic speech recognition on the TIMIT task using DNN-HMM [21]. However, further investigations (e.g., fine-tuning of the model structure including the number of nodes for each layer) are needed to understand how the number of hidden layers in DNN-HMM impacts the silent speech recognition performance.

Plosives, fricatives, and affricates are expected to be challenging to recognize from articulatory movements due to their articulation characteristics. For example, some voiced and voiceless plosive pairs (e.g., /p/ and /b/) have the same place of articulation. Thus it is interesting to see if DNN-HMM can achieve better results than GMM-HMM on distinguishing them. Tables 2 to 5 show the confusion matrices for classifying plosives (i.e., /p/, /b/, /t/, /d/, /k/, and /g/), fricatives (i.e., /f/, /v/, /s/, and /z/), and affricates (i.e., /tʃ/ and /dʒ/) using GMM-HMM and DNN-HMM for the male and the female speaker, respectively. There are only a few samples for fricative /ʒ/ in MOCHA-TIMIT database. Thus, /ʒ/ was excluded in the tables. The numbers in the tables are sum of raw numbers of samples along the five cross-validations. The row numbers are the actual samples and the column numbers are the predicted samples. The diagonal numbers are the correctly predicted number of samples.

DNN-HMM showed better recognition of plosives, fricatives, and affricates than GMM-HMM for both speakers. The diagonal numbers in Tables 3 and 5 (DNN-HMM) are all greater than those in Tables 2 and 4 (GMM-HMM) except for /b/ and /k/ for the female speaker. DNN-HMM achieved less numbers of deletions and insertions than GMM-HMM. However, DNN-HMM showed more misclassifications for the voiced-voiceless plosive pairs with the same place of articulation, for example, labial plosives /p/ - /b/, alveolar plosives /t/ - /d/, and velar plosives /k/ - /g/. The results indicated DNN-HMM did not improve the results for distinguishing those plosive pairs with the same place of articulation but different voice feature (voiced vs voiceless), although DNN-

Table 2: Confusion Matrix for selected *plosives*, *fricatives*, and *affricates* using GMM-HMM for the *female* speaker.

		Plosives						Fricatives				Affricates		
		/p/	/b/	/t/	/d/	/k/	/g/	/f/	/v/	/s/	/z/	/tʃ/	/dʒ/	Del
Plosives	/p/	189	62		1			2	1		2			60
	/b/	60	136		2			1	1	5				53
	/t/	6	4	390	72	5	2	1	3	12	21	2	3	261
	/d/	1	1	100	164	5	1		2	9	4	2	2	175
	/k/	1			1	386	51		1		2			51
	/g/			1		79	57					1	1	26
Fricatives	/f/	2	1	3	1		2	162	29	1		1		40
	/v/	1		1	2			45	115	1	1			39
	/s/	1	1	14	5		3	1		444	108		2	98
	/z/	1	1	14	4	7		1		115	232		1	89
Affricates	/tʃ/			1	3				1			37	25	13
	/dʒ/	1	1	2	2					1		14	57	32
	Ins	8	8	27	19	32	5	8	9	22	17	10	6	

Table 3: Confusion Matrix for selected *plosives*, *fricatives*, and *affricates* using DNN-HMM for the *female* speaker.

		Plosives						Fricatives				Affricates		
		/p/	/b/	/t/	/d/	/k/	/g/	/f/	/v/	/s/	/z/	/tʃ/	/dʒ/	Del
Plosives	/p/	209	68							1	1			49
	/b/	86	131			1				3			1	47
	/t/	4	2	409	101	3		1	2	4	19	3	2	238
	/d/	3		109	188				2	11	6	3	1	142
	/k/				2	371	66		1		1			48
	/g/					71	73	1		1				26
Fricatives	/f/	4		1	6		1	185	30	1	1		1	22
	/v/		1		2			41	141	1				27
	/s/	2	1	3	2	1		1		470	104	2	1	95
	/z/			7	5	1				111	265			74
Affricates	/tʃ/			4					1			45	18	17
	/dʒ/			6	2							21	64	20
	Ins	5	11	20	18	14	3	5	4	19	14	4	4	

Table 4: Confusion Matrix for selected *plosives*, *fricatives*, and *affricates* using GMM-HMM for the *male* speaker.

		Plosives						Fricatives				Affricates		
		/p/	/b/	/t/	/d/	/k/	/g/	/f/	/v/	/s/	/z/	/tʃ/	/dʒ/	Del
Plosives	/p/	170	64	1	1					1	2			66
	/b/	70	112		2			1		3	2	1		43
	/t/	2	4	405	67	2	2	2	1	6	29		6	260
	/d/	2	2	75	151	2		1	2	15	6	3	1	180
	/k/		1			340	55		2		2	1		75
	/g/	1		2	1	73	53					1		40
Fricatives	/f/	1		2	4	1	1	151	31	2	2	1		51
	/v/			1	1	2		34	114	2	1			49
	/s/		2	21	8	1	2	1		416	88			119
	/z/			29	7	4		1	2	94	232	1	2	97
Affricates	/tʃ/			2							1	33	26	14
	/dʒ/				1			1		2		15	57	24
	Ins	9	8	27	11	19	7	2	8	9	11	6	7	

Table 5: Confusion Matrix for selected *plosives*, *fricatives*, and *affricates* using DNN-HMM for the *male* speaker.

		Plosives						Fricatives				Affricates		
		/p/	/b/	/t/	/d/	/k/	/g/	/f/	/v/	/s/	/z/	/tʃ/	/dʒ/	Del
Plosives	/p/	203	71					1			1			50
	/b/	87	122							3			1	36
	/t/	2	1	433	74	1	6	4	1	5	23	2	5	242
	/d/	3		59	215		1	1	1	13	2	2		167
	/k/		1	1		347	66			1				68
	/g/					84	79							15
Fricatives	/f/	2		2			2	178	45			1		22
	/v/			3	1			36	151		2			22
	/s/	1		15	3		4			446	103			103
	/z/			20	9			1	1	113	268	2		58
Affricates	/tʃ/			1					1		1	45	25	10
	/dʒ/			4	3					2		27	70	18
	Ins	3	11	21	12	9	12	4	2	15	12	2	6	

HMM is generally better than GMM-HMM in classifying all the plosives. A similar pattern was observed for fricatives. DNN-HMM had more misclassifications than GMM-HMM in distinguishing those fricative pairs with the same place of articulation but different voice feature (voiced vs voiceless), for example, labio-dental fricatives /f/ and /v/ and alveolar fricatives /s/ and /z/. In a word, it seems that the overall error rates were reduced for classifying plosives and fricatives using DNN-HMM, inherent difficulty for voiced and voiceless consonant classification still remains when acoustic information is not available. Uraga and Hain showed acoustic features (mel-frequency cepstral coefficients; MFCCs) contains more information than articulatory features for classifying those voiced and voiceless plosives [27].

Although the experimental results have illustrated the significant performance improvement of DNN-HMM over GMM-HMM, we expect DNN-HMM still has potential to further improve silent speech recognition accuracy when combined with other approaches used in acoustic speech recognition (e.g., speaker/environment adaptation [19]).

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have investigated the use of DNN-HMM in silent speech recognition from articulatory movements (i.e., without using acoustic data). The performance of DNN-HMM and GMM-HMM was compared. Experimental results illustrated the significant performance improvement of DNN-HMM over GMM-HMM.

Future directions include (1) the use of time-series data processing techniques for variation reduction (e.g., symbolic aggregation representation [28]), (2) speaker-independent silent speech recognition [31] using DNN-HMM (extension to context-dependent triphone system), (3) investigating adaptation schemes [10], and (4) removing redundant information between sensors for silent speech recognition [4, 29].

6. ACKNOWLEDGMENT

This work was in part supported by the National Institutes of Health through a grant (R01 DC013547). We would like to thank the support from the Communication Technology Center, University of Texas at Dallas.

7. REFERENCES

[1] Ahmad Khan, Z., Green, P., Creer, S., Cunningham, S. 2011. Reconstructing the voice of an individual following laryngectomy. *Augmentative and Alternative Communication* 27(1), 61–66.

[2] Bailey, B., Johnson, J., Newlands, S. 2006. *Head & neck*

surgery–otolaryngology volume 1. Lippincott Williams & Wilkins.

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

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

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

[6] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., Brumberg, J. 2010. Silent speech interfaces. *Speech Communication* 52(4), 270–287.

[7] Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., Acero, A. 2013. Recent advances in deep learning for speech research at Microsoft. *Proc. of ICASSP* Vancouver, Canada. 8604–8608.

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

[9] Fagan, M., Ell, S., Gilbert, J., Sarrazin, E., Chapman, P. 2008. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical engineering & physics* 30(4), 419–425.

[10] Hahm, S., Ogawa, A., Delcroix, M., Fujimoto, M., Hori, T., Nakamura, A. 2013. Feature space variational Bayesian linear regression and its combination with model space VBLR. *Proc. of ICASSP* Vancouver, Canada. 7898–7902.

[11] Heracleous, P., Hagita, N. 2011. Automatic recognition of speech without any audio information. *Proc. of ICASSP* Prague, Czech Republic. 2392–2395.

[12] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29(6), 82–97.

[13] Hinton, G. E. 2012. A practical guide to training restricted boltzmann machines. In: *Neural Networks: Tricks of the Trade*. Springer 599–619.

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

[15] Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., Stone, M. 2009. Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface. *Proc. of INTERSPEECH* Brighton, UK. 640–643.

[16] Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., Stone, M. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication* 52(4), 288–300.

[17] Jorgensen, C., Dusan, S. 2010. Speech interfaces based upon surface electromyography. *Speech Communication* 52(4), 354–366.

[18] King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M. 2007. Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America* 121(2), 723–742.

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

[20] Liu, H., Ng, M. 2007. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx* 34(3), 327–332.

[21] Mohamed, A.-R., Dahl, G., Hinton, G. 2012. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing* 20(1), 14–22.

[22] Nakano, Y., Tachibana, M., Yamagishi, J., Kobayashi, T. 2006. Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis. *Proc. of INTERSPEECH* Pittsburgh, USA. 2286–2289.

[23] Nguyen, N. 2000. A MATLAB toolbox for the analysis of articulatory data in the production of speech. *Behavior Research Methods, Instruments, & Computers* 32(3), 464–467.

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

[25] Richmond, K. 2009. Preliminary inversion mapping results with a new EMA corpus. *Proc. of INTERSPEECH* Brighton, UK. 2835–2838.

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

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

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

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

[30] Wang, J., Green, J., Samal, A., Yunusova, Y. 2013. Articulatory distinctiveness of vowels and consonants: A data-driven approach. *Journal of Speech, Language, and Hearing Research* 56(5), 1539–1551.

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

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

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

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

[35] Zen, H., Nankaku, Y., Tokuda, K. 2011. Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Transactions on Audio, Speech, and Language Processing* 19(2), 417–430.