INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Multiview Representation Learning via Deep CCA for Silent Speech Recognition

Myungjong Kim[1], *Beiming Cao*[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, United States
[3]Department of Otolaryngology-Head and Neck Surgery, University of Texas Southwestern Medical Center, United States

{myungjong.kim, beiming.cao, wangjun}@utdallas.edu; ted.mau@utsouthwestern.edu

Abstract

Silent speech recognition (SSR) converts non-audio information such as articulatory (tongue and lip) movements to text. Articulatory movements generally have less information than acoustic features for speech recognition, and therefore, the performance of SSR may be limited. Multiview representation learning, which can learn better representations by analyzing multiple information sources simultaneously, has been recently successfully used in speech processing and acoustic speech recognition. However, it has rarely been used in SSR. In this paper, we investigate SSR based on multiview representation learning via canonical correlation analysis (CCA). When both acoustic and articulatory data are available during training, it is possible to effectively learn a representation of articulatory movements from the multiview data with CCA. To further represent the complex structure of the multiview data, we apply deep CCA, where the functional form of the feature mapping is a deep neural network. This approach was evaluated in a speaker-independent SSR task using a data set collected from seven English speakers using an electromagnetic articulograph (EMA). Experimental results showed the effectiveness of the multiview representation learning via deep CCA over the CCA-based multiview approach as well as baseline articulatory movement data on Gaussian mixture model and deep neural network-based SSR systems.

Index Terms: articulatory movements, deep canonical correlation analysis, multiview representation learning, silent speech recognition

1. Introduction

Laryngectomy is a surgical removal of the larynx for the treatment of laryngeal or other oral cavity cancers [1]; therefore, persons after laryngectomy lose their ability to produce speech sounds and suffer in their daily communication [2, 3]. Although there are several options to communicate for those patients such as esophageal speech, trachea-esophageal speech, and electrolarynx, these approaches frequently produce an abnormal sounding voice that is hard to understand by normal listeners [4, 5].

Silent speech interfaces (SSIs) [6] have the potential to provide an alternative way to assist those patients to produce speech with natural sounding voice from the movements of their articulators such as the tongue and lips. SSIs typically include an articulatory movement recorder, a silent speech recognizer that converts articulatory movements to text [7], and a text-to-speech synthesizer [8]. A variety of techniques have been used to record articulatory movements including ulstrasound [9, 10], surface electromyography [11, 12], and electromagnetic articulograph (EMA) [13, 14] (including portable magnetic sensing [15]). EMA is a direct and attractive approach to record the articulatory movements since it captures the 3D motion of several sensors adhered to the tongue and lips [14]. We used EMA sensors in this work to track the precise Cartesian coordinates of articulators. Text-to-speech synthesis with a standard voice has been well studied and is ready for this application [16]. Researchers on TTS are currently exploring how to restore the laryngectomee′s own voice [4, 17] with limited training data. Therefore, the core problem in current SSI research is developing effective algorithms that map articulatory movements to text information.

Related work on silent speech recognition (SSR) has been mostly focused on articulatory modeling to characterize articulatory movements. Researchers have developed SSR approaches based on dynamic time warping (DTW) [14, 18], Gaussian mixture model (GMM)-hidden Markov model (HMM) [10, 11], and support vector machine (SVM) [7, 12]. Hahm and Wang [19] recently reported that deep neural network (DNN)-HMM based articulatory models outperformed conventional GMM-HMM based models in speaker-dependent SSR with EMA data.

A more recent focus in SSR research was to reduce the inter-speaker difference of articulatory movements, so that speaker-independent SSR can be feasible [20]. Articulatory movement data are directly affected by the speaker′s anatomy (e.g., tongue shape and size) and the speaker′s articulatory patterns. To reduce the physiological difference, researchers have tried to normalize the articulatory movements by aligning the tongue position when producing vowels [21, 22] and consonants [23] to a reference (e.g., palate [21,22] or a general tongue shape [23]). Procrustes matching, a robust shape analysis technique [24], was used to reduce the translational, rotational, and scaling effects of articulatory data across speakers [25]. Also, feature space maximum likelihood linear regression (fMLLR), which is a data-driven speaker normalization approach, has been successfully applied to SSR with EMA data [20].

Despite the advances in SSR research, conventional approaches have largely relied on using articulatory movement information only. Articulatory movements generally have less information than acoustic features for recognizing phonemes [26], and therefore, the performance of SSR may be limited. Multiview representation learning can capture better information by simultaneously analyzing multiple information sources, such as articulatory and acoustic data. The multiview representation learning approach is generally based on learning a feature transformation of the primary view (available for both train-

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-952

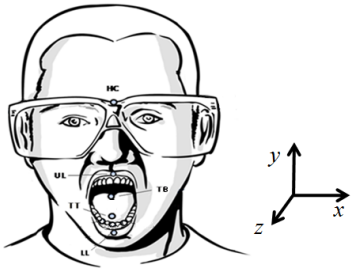

Figure 1: *Sensor labels and locations.*

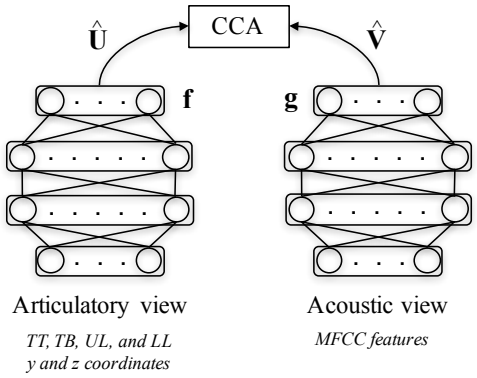

Figure 2: *Schematic diagram of deep CCA.*

ing and testing) that captures useful information from the second view (accessible only for training) using a paired two-view training set. Speaker-independent SSR models make the use of multiview representation learning possible, because acoustic data could be collected from training speakers. Of course, potential target test speakers (e.g., laryngecomees) will be unable to produce acoustic speech. Multiview representation learning has been recently successfully applied in speech processing [27] and acoustic speech recognition [28]. Multiview representation learning, however, has rarely been studied in SSR.

In this paper, we investigated the effectiveness of multiview representation learning via canonical correlation analysis (CCA), which learns features in multiviews that are maximally correlated, for SSR. To further represent the complex structure of the multiview data, we apply deep CCA, where the functional form of the feature mapping is a deep neural network. We used articulatory movement data as the primary view and acoustic features such as mel-frequency cepstral coefficients (MFCCs) as the second view. Therefore, it is expected that the feature transformation from deep CCA can capture useful information from the acoustic view, which makes it possible to improve phonetic recognition accuracy.

2. Data Collection

2.1. Participants and speech tasks

Seven American English speakers (4 females and 3 males) participated in the data collection. The mean age of the participants was 25.4 $\pm$ 3.6 years. No history of speech, language, or cognitive problems from any participants was reported. Each subject participated in one session in which he or she repeated a list of 132 phrases twice at their habitual speaking rate. The phrases (e.g., *how are you doing?*) that are frequently used in daily life were selected from [14, 20].

2.2. Tongue and lip motion tracking

The Wave system (Northern Digital Inc., Waterloo, Canada), a commercially available electromagnetic tongue and lip motion tracking device, was used to collect the movement data of articulators for all participants. Four small sensors were attached to the surface of each articulator using dental glue (PeriAcryl 90, GluStitch) or tape, including Tongue Tip (TT), Tongue Body Back (TB), Upper Lip (UL), and Lower Lip (LL) as in Figure 1. In addition, another sensor was attached to the middle of forehead (Head Center, HC) for head correction. Our prior work indicated that the four-sensor set consisting of TT, TB, UL, and LL is an optimal set for this application [29]. With this approach, three-dimensional movement data of articulators were tracked and recorded. The sampling rate in Wave recording in this work was 100Hz. The spatial precision of movement tracking is about 0.5mm in the center location of the magnetic field [30].

Before data analysis, the translation and rotation of the HC sensor were subtracted from the motion data of the tongue and lip sensors to obtain head-independent articulatory data. The head translation and rotation removal was automatically done by the Wave system. Figure 1 illustrates the derived 3D Cartesian coordinates system, in which x is left-right direction; y is vertical; and z is front-back direction. It is not expected that articulators have significant lateral movement (x in Figure 1) [31], thus only y and z coordinates of the tongue and lip movement data were used for analysis.

Acoustic data were collected synchronously with the articulatory movement data by built-in microphone in the Wave system. In total, 1,847 utterances (total 31,534 phonemes/39 unique phonemes) for 132 unique phrases were collected from the seven participants.

3. Multiview Representation Learning

3.1. Canonical correlation analysis (CCA)

In multiview representation learning, we have access to different types of observations of the same underlying data, such as articulatory and acoustic features. In this work, the training data consist of pairs of observations $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^{D_x}$ and $\mathbf{y}_i \in \mathbb{R}^{D_y}$ denote the ith input features of simultaneously recorded articulatory and acoustic features, respectively.

CCA is one of the widely used methods for learning a compact representation from multiview data [28, 32]. The objective of CCA is to find linear projection matrices $\mathbf{U} \in \mathbb{R}^{D_x \times L}$ and $\mathbf{V} \in \mathbb{R}^{D_y \times L}$, where $L \leq min(D_x, D_y)$, such that the projections are maximally correlated with each other while the dimensions in the representation are uncorrelated with each other. CCA can be represented by

$$\begin{aligned} &\max_{\mathbf{U},\mathbf{V}} \frac{1}{N} tr\Big(\mathbf{U}^\intercal \mathbf{X}\mathbf{Y}^\intercal \mathbf{V}\Big) \\ &s.t. \mathbf{U}^\intercal \Big(\frac{\mathbf{X}\mathbf{X}^\intercal}{N} + r_x I\Big)\mathbf{U} = \mathbf{V}^\intercal \Big(\frac{\mathbf{Y}\mathbf{Y}^\intercal}{N} + r_y I\Big)\mathbf{V} = I \end{aligned} \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{D_x \times N}, \mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{D_y \times N}$, and $(r_x, r_y) \geq 0$ are regularization parameters. Let $\Sigma_{12} = \frac{1}{N}\mathbf{X}\mathbf{Y}^\intercal$, $\Sigma_{11} = \frac{1}{N}\mathbf{X}\mathbf{X}^\intercal + r_x I$, and $\Sigma_{22} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^\intercal + r_y I$ be the cross- and regularized auto-covariance matrices of the data in the two views. The solution can be

Table 1: *Phoneme error rates (%) with varying the number of hidden layers of DCCA on monophone and triphone systems.*

# of hidden layers		Monophone	Triphone
Art. view	Acs. view		
0	0	63.4	55.1
1	0	63.9	56.0
0	1	62.2	53.7
1	1	64.3	57.0
0	2	62.0	**52.4**
0	3	61.9	52.8

obtained by singular value decomposition of the matrix $\mathbf{T} = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$. The final CCA features (projections) are $\tilde{\mathbf{x}} = \mathbf{U}^\intercal\mathbf{x}$ for the primary (articulatory) view and $\tilde{\mathbf{y}} = \mathbf{V}^\intercal\mathbf{y}$ for the second (acoustic) view. Consequently, the final CCA features for the articulatory view were used for SSR.

3.2. Deep CCA

Suppose we have neural network-based feature mappings $\mathbf{f} : \mathbb{R}^{D_x} \mapsto \mathbb{R}^{d_x}$ for the primary view and $\mathbf{g} : \mathbb{R}^{D_y} \mapsto \mathbb{R}^{d_y}$ for the second view. A K-layer neural network can be represented by $\mathbf{f}(\mathbf{x}) = \mathbf{f}_K((\cdots \mathbf{f}_1(\mathbf{x}; \mathbf{W}_1)\cdots); \mathbf{W}_K)$, where $\mathbf{W}_j$ are the weight parameters of layer j, $j = 1, ..., K$, and $\mathbf{f}_j$ is the non-linear mapping of layer j. In deep CCA (DCCA), we learn weights $\mathbf{W} = \{\mathbf{W}_1, ..., \mathbf{W}_K\}$ as well as linear projection matrices $\hat{\mathbf{U}} \in \mathbb{R}^{d_x \times L}$ and $\hat{\mathbf{V}} \in \mathbb{R}^{d_y \times L}$, where $L \leq min(d_x, d_y)$, that optimize the canonical correlations at the output layers.

$$\begin{aligned} &\max_{\mathbf{W_f}, \mathbf{W_g}, \hat{\mathbf{U}}, \hat{\mathbf{V}}} \frac{1}{N} tr\left(\hat{\mathbf{U}}^\intercal \mathbf{F}(\mathbf{X}; \mathbf{W_f}) \mathbf{G}(\mathbf{Y}; \mathbf{W_g})^\intercal \hat{\mathbf{V}}\right) \\ &s.t.\ \hat{\mathbf{U}}^\intercal\left(\frac{\mathbf{F}\mathbf{F}^\intercal}{N} + r_x I\right)\hat{\mathbf{U}} = \hat{\mathbf{V}}^\intercal\left(\frac{\mathbf{G}\mathbf{G}^\intercal}{N} + r_y I\right)\hat{\mathbf{V}} = I \end{aligned} \tag{2}$$

where $\mathbf{F} = \mathbf{f}(\mathbf{X}) = [\mathbf{f}(\mathbf{x}_1), ..., \mathbf{f}(\mathbf{x}_N)] \in \mathbb{R}^{d_x \times N}$ and $\mathbf{G} = \mathbf{g}(\mathbf{Y}) = [\mathbf{g}(\mathbf{y}_1), ..., \mathbf{g}(\mathbf{y}_N)] \in \mathbb{R}^{d_y \times N}$. Also, $\mathbf{W_f}$ and $\mathbf{W_g}$ are weight parameters for $\mathbf{F}$ and $\mathbf{G}$, respectively.

The projections $(\hat{\mathbf{U}}, \hat{\mathbf{V}})$ can be regarded as adding an extra linear layer on top of non-linear mappings $(\mathbf{f}, \mathbf{g})$, respectively, as in Figure 2. The final DCCA features are $\tilde{\mathbf{x}} = \hat{\mathbf{U}}^\intercal\mathbf{f}(\mathbf{x})$ for the articulatory view. For the optimization of (2), we used stochastic gradient descent (SGD) suggested by [28].

3.3. Application to speaker-independent SSR

To obtain better articulatory representations, we combine our representation learning with speaker normalization techniques such as Procrustes matching and fMLLR, which were successfully applied to speaker-independent SSR [20]. In this work, Procrustes matching was first applied to reduce the inter-speaker physiological differences (tongue and lip orientation). In this method, two dimensional (i.e., y and z coordinates) movement data of TT, TB, UL, and LL were transformed into a "normalized shape", which had a centroid at the origin (0,0) and the centroids of the UL and LL formed a vertical line. Then, L-dimensional DCCA features obtained from the Procrustes matched data were appended to the Procrustes matched data as in a tandem approach [33]. The resulting features are further transformed using fMLLR, which is one of the representative data-driven approaches for feature space normalization. A more detailed explanation of the Procrustes matching and fMLLR can be found in [20].

3.4. Experimental setup

We used 24 dimensional EMA data consisting of 8 static data (2 dimension $\times$ 4 sensors) and their first and second order derivatives with shift size of 10 milliseconds as input features to silent speech recognition systems as baseline. Acoustic features are 39 dimensional MFCCs consisting of 13 static and their first and second derivatives with frame size of 25 milliseconds and shift size of 10 milliseconds. We used mean normalization along each dimension as a default setting. In multiview feature learning, the inputs are articulatory and acoustic features concatenated with a context window of 9 frames around each frame, resulting in 216 dimensional articulatory ($D_x = 216$) and 351 dimensional acoustic inputs ($D_y = 351$) for the (D)CCA models. For DCCA, different neural network architectures (i.e., different numbers of hidden layers) for each view were investigated in a preliminary evaluation. The best configuration in the neural network was 256 hidden units ($d_x = d_y = 256$) at each hidden layer. We used rectified linear units (ReLU) as an activation function. After training a feature transformation from the (D)CCA models, the feature transformation was applied to the articulatory data, giving L-dimensional features. Finally, the L-dimensional features were appended to 24 dimensional articulatory data, producing a final feature vector for each frame.

We evaluated the (D)CCA approaches on both GMM-HMM and DNN-HMM systems. It consisted of 773 tied-state (senone) left-to-right triphone HMMs, where each HMM has 3 states. GMM was trained using maximum likelihood estimation. DNN was trained using EMA data with a context window of 9 frames. The DNN had 3 hidden layers with 512 hidden units at each layer and the 773 dimensional softmax output layer, corresponding to the number of senones of the GMM-HMM system. The parameters were initialized using layer-by-layer generative pre-training based on restricted Boltzmann machines (RBM) and the network was discriminatively trained using backpropagation [34]. The bigram and trigram phoneme language models (LMs) were trained using the IRSTLM toolkit [35] and they were used in decoding. The test perplexities of the bigram and trigram are 10.0 and 3.2, respectively. The training and decoding were performed using the Kaldi speech recognition toolkit [36].

Phoneme error rate (PER) was used as the measure for the SSR performance. Leave-one-subject-out (7 fold) cross validation was used in the experiment. The average performance of cross validations was reported as the overall performance.

4. Results and Discussion

4.1. Evaluation with the number of hidden layers in each view of DCCA

We first examined the effectiveness of DCCA with varying the number of hidden layers in each view. Table 1 presents the performances of DCCA according to the number of hidden layers on monophone and triphone-based GMM-HMM systems with bigram LMs. The number of monophone units is 122 (39 phonemes $\times$ 3 states + 5 states for silence) and the number of trihphone units (senones) is 773. 10-dimensional DCCA features were appended to the Procrustes matched data (Proc.+DCCA) in this experiment. Note that "0" hidden layers for both the articulatory and acoustic views indicate the CCA approach. Interestingly, asymmetric architectures, which are a linear mapping for the articulatory view and a highly nonlinear mapping (deep network) for the acoustic view, produce better performance. This result is consistent with the recent study in

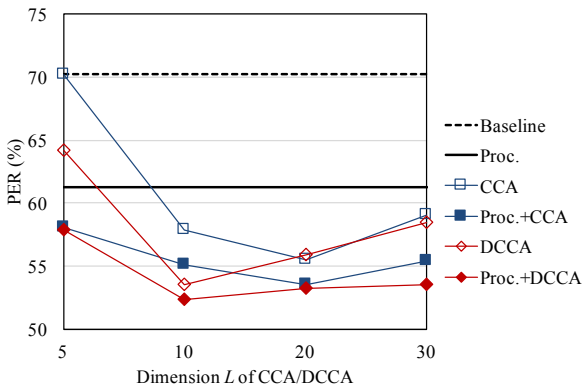

Figure 3: *PERs with the number of (D)CCA dimensions.*

acoustic speech recognition [28]. We obtained the best result (52.4%) on the "0" layer for the articulatory view and the "2" layer for the acoustic view on the triphone system. Therefore, we used this setting as a default of DCCA architecture in the following experiments.

4.2. Evaluation with the number of DCCA dimensions

Figure 3 shows the performances of (D)CCA by varying the number of their dimensions (L) in GMM-HMM based SSR systems with bigram LMs. As can be seen, Procrustes matching outperformed the baseline articulatory movement data. When we used (D)CCA alone, the performances were better than with baseline and Procrustes matching in almost all dimensional settings. Based on the Procrustes matched data, when we appended L-dimensional (D)CCA transformed data, the PERs were further reduced. The best performance was obtained on 10 dimension of DCCA. In the following experiments, 10 dimensions for the CCA and DCCA projections were used as the default setting for the remainder of this paper.

4.3. Effectiveness of DCCA

Table 2 shows the performances of CCA/DCCA on GMM-HMM and DNN-HMM systems with bigram and trigram LMs in terms of the PER. Here, "Proc." indicates Procrustes matching. To evaluate the systems, speaker normalization methods such as Procrustes matching and fMLLR were incorporated with CCA or DCCA features. When appending CCA features to Procrustes matched data (Proc.+CCA), the performance was better than with Procrustes matching and fMLLR (Proc.-fMLLR). When we further applied fMLLR (Proc.+CCA-fMLLR), PER was much reduced on both GMM-HMM and DNN-HMM systems. When we replaced CCA by DCCA (Proc.+DCCA-fMLLR), we were able to obtain the best performance, giving PERs of 40.3% on GMM-HMM and 33.9% on DNN-HMM with trigram LMs, which is close to the performance of acoustic features (MFCC).

4.4. Discussion

These results indicated that incorporating acoustic information in a feature transformation by DCCA is effective for recognizing phonemes, complementing articulatory movements. Also, DCCA can be successfully combined with well-established conventional feature transformation methods and could further improve silent speech recognition performance. Moreover, DNN-HMM outperformed GMM-HMM in all experimental configurations (e.g., CCA and DCCA). This finding is consistent with the literature on acoustic [34] and silent speech recognition [19, 20].

Table 2: *Phoneme error rates (%) of (D)CCA on GMM-HMM and DNN-HMM based SSR systems.*

Method	Bigram	Trigram
GMM-HMM		
Baseline	70.3	65.2
Proc.	61.3	53.6
Proc.-fMLLR	55.9	45.8
Proc.+CCA	55.1	48.8
Proc.+CCA-fMLLR	49.5	42.1
Proc.+DCCA	52.4	45.8
Proc.+DCCA-fMLLR	**47.3**	**40.3**
MFCC	43.6	27.4
DNN-HMM		
Baseline	57.3	47.2
Proc.	53.2	42.9
Proc.-fMLLR	50.8	41.3
Proc.+CCA	46.5	36.6
Proc.+CCA-fMLLR	43.7	34.6
Proc.+DCCA	45.9	35.6
Proc.+DCCA-fMLLR	**42.5**	**33.9**
MFCC	38.1	22.8

Although the evaluation was on normal participants, our approach (DCCA) can be applied to laryngectomees, where both acoustic and articulatory training data will be from healthy speakers. Articulatory movements generally differ between silently articulated (e.g., laryngectomees) and normally spoken speech, and therefore, mapping strategies between them may be considered [37]. Our approach may also have the potential for dysarthric speech recognition [38] as well as whispered speech recognition with articulatory information [39].

5. Conclusions and Future Work

In this paper, we investigated the effectiveness of CCA and DCCA to improve the speaker-independent SSR performance. We used multiview representation learning via DCCA that finds a deep feature mapping that is maximally correlated in articulatory and acoustic views. Experimental results showed that the DCCA-based multiview approach provides significant improvement over the CCA approach and the baseline (DNN-HMM without using CCA). Further, our approach was successfully combined with other transformation approaches such as Procrustes matching and fMLLR for speaker normalization. Future directions include 1) a test of the DCCA approach using a larger dataset collected from more subjects and laryngectomees and 2) applying other deep structures such as convolutional neural network for multiview representation learning-based SSR.

6. Acknowledgements

This work was supported by the National Institutes of Health under an award R03DC013990 and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We thank Joanna Brown, Betsy Ruiz, Janis Deane, Laura Toles, Amy Hamilton, Se-in Kim, Kristin Teplansky, Katie Purdum, and the volunteering participants.

7. References

[1] B. Bailey, J. Johnson, and S. Newlands, *Head & Neck Surgery – Otolaryngology*. Lippincott Williams & Wilkins, 2006.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] J. Wang, A. Samal, J. Green, and F. Rudzicz, "Whole-word recognition from articulatory movements for silent speech interfaces," in *Proc. INTERSPEECH*, 2012.

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

[9] 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.

[10] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech and Language*, vol. 36, pp. 274–293, 2016.

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

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

[13] 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.

[14] 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.

[15] R. Hofe, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore, and S. I. Rybchenko, "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing," *Speech Communication*, vol. 55, no. 1, pp. 22–32, 2013.

[16] 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. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Munich, Germany, 1997, pp. 959–962.

[17] B. Cao, M. Kim, J. van Santen, T. Mau, and J. Wang, "Integrating articulatory information into deep learning-based text-to-speech synthesis," in *Proc. of INTERSPEECH 2017*, Accepted.

[18] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing dtw to the multi-dimensional case requires an adaptive approach," *Data Mining and Knowledge Discovery*, vol. 31, no. 1, pp. 1–31, 2017.

[19] 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.

[20] J. Wang and S. Hahm, "Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training," in *Proc. INTERSPEECH*, 2015, pp. 2415–2419.

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

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

[23] J. R. Westbury, M. Hashi, and M. J. Lindstrom, "Differences Among Speakers in Lingual Articulation for American English /R/," *Speech Commun.*, vol. 26, no. 3, pp. 203–226, 1998.

[24] I. L. Dryden and K. V. Mardia, "Statistical shape analysis," *Journal of Speech, Language, and Hearing Research*, vol. 54, 1998.

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

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

[27] J. C. Vsquez-Correa, J. R. Orozco-Arroyave, R. Arora, E. Nth, N. Dehak, H. Christensen, F. Rudzicz, T. Bocklet, M. Cernak, H. Chinaei, J. Hannink, P. S. Nidadavolu, M. Yancheva, A. Vann, and N. Vogler, "Multi-view representation learning via GCCA for multimodal analysis of Parkinson's disease," in *Proc. ICASSP*, 2017, pp. 2966–2970.

[28] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *Proc. ICASSP*, 2015, pp. 4590–4594.

[29] 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.

[30] 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.

[31] 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.

[32] M. Borga, "Canonical correlation: a tutorial," 2001.

[33] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.

[34] 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.

[35] M. Federico, N. Bertoldi, and M. Cettolo, "Irstlm: an open source toolkit for handling large scale language models," in *INTERSPEECH*, 2008.

[36] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Waikoloa, USA, 2011, pp. 1–4.

[37] M. Wand, M. Janke, and T. Schultz, "Tackling speaking mode varieties in EMG-based speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2515–2526, 2014.

[38] M. Kim, J. Wang, and H. Kim, "Dysarthric speech recognition using kullback-leibler divergence-based hidden markov model," in *Proc. of INTERSPEECH*, 2016, pp. 2671–2675.

[39] B. Cao, M. Kim, T. Mau, and J. Wang, "Recognizing whispered speech produced by an individual with surgically reconstructed larynx using articulatory movement data," in *Proc. ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2016, pp. 80–86.