A COMPARATIVE STUDY OF ACOUSTIC-TO-ARTICULATORY INVERSION FOR NEUTRAL AND WHISPERED SPEECH

Aravind Illa, Nisha Meenakshi G, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

ABSTRACT

Whispered speech is known to have different characteristics in acoustics and articulation compared to neutral speech. In this study, we compare the accuracy with which the articulation can be recovered from the acoustics of both types of speech, individually. Acoustic-to-articulatory inversion (AAI) is performed with twelve articulatory features using the deep neural network (DNN) with data obtained from four subjects. We consider AAI in matched and mis-matched train-test conditions, where the speech types in training and test are identical and different respectively. Experiments in matched condition reveal that the AAI performance for whispered speech drops significantly compared to that for neutral speech, only for jaw, tongue tip and tongue body, consistently, for all four subjects. This indicates that the whispered speech encodes information about the rest of the articulators to a degree similar to that of the neutral speech. Experiments in the mis-matched condition show a consistent drop in the AAI performance compared to the matched condition. This drop in performance from matched to mis-matched condition is found be the highest for upper lip which indicates that the upper lip movement could be encoded differently in whispered speech compared to that in neutral speech.

Index Terms— acoustic-to-articulatory inversion, neutral speech, whispered speech, electromagnetic articulography,

1. INTRODUCTION

Acoustic-to-articulatory inversion (AAI) is the task of recovering articulatory movement from acoustic representations. AAI has been shown to be useful for speech synthesis [1], and human computer interaction applications [2, 3, 4], and automatic speech recognition [5], especially in cases of noisy, spontaneous, or pathological speech [6, 7, 8, 9]. A number of techniques have been proposed in the literature for AAI including codebook based procedures [10, 11], statistical modeling of the acoustic-articulatory map such as Gaussian mixture model (GMM) [12], mixture density network (MDN) [13], a trajectory hidden-Markov model (HMM) [14], generalized smoothness criterion (GSC) [15] and neural network-based modeling of the acoustic-to-articulatory mapping [9, 16] including deep neural network (DNN) [17, 18].

All the works on AAI in the literature have been primarily on neutral speech. To the best of our knowledge, there is no reported result on AAI for whispered speech. Whispered speech often appears in private communication [19] and pathological situations [20, 21, 22]. There are several differences in the acoustics and the articulation between the whispered and the neutral speech. For example, there is no pitch in whispered speech [23, 24] and, hence, it sounds like unvoiced speech [25]. It also differs from neutral speech by the shift of formants in low frequencies [26, 27]. Similarly, there are differences in articulation during whispering compared to neutral speech [28, 29]. For example, whispered speech induces a constriction in the false vocal folds region [27] unlike neutral speech. Several studies on the articulatory space of whispered consonants reveal that hyper-articulation occurs while whispering to ensure intelligibility. Specifically, exaggerated movements of the tongue have been reported using Electro-palatography [30, 31]. Differences in the lip kinematics in whispered bilabial consonants compared to their neutral counterparts have also been reported [32]. Due to such differences in acoustics and articulation, the acoustic-to-articulatory map in whispered speech could be different from the that in the neutral speech. It remains unclear how such differences could impact the AAI performance for whispered speech.

In this work, we perform a comparative study between the AAI performance for neutral and whispered speech. The goal of the study is to quantify the accuracy with which different articulators are recovered from acoustics in these two types of speech. We also examine how an AAI model trained using a neutral acoustic-articulatory map performs for inversion of whispered acoustics (mis-matched train-test condition) and vice-versa. The study is conducted using acoustic and articulatory data of four subjects for both neutral and whispered speech. The articulatory data is obtained using Electromagnetic articulography (EMA). Experiments of AAI, separately for each subject reveal that there is a significant drop in the AAI performance for three out of twelve articulatory features in the case of whispered speech compared to neutral speech. Similarly, experiments with mis-matched train-test condition show that the AAI performance drops by $\sim$20% (relative) compared to matched train-test condition.

2. DATASET

To perform this study, we collected acoustic and articulatory movement data for both neutral and whispered speech. A total four subjects comprising three males (M1, M2, M3) and one female (F1) participated in the data collection for this study. The age of M1, M2, M3 and F1 was 19, 22, 24 and 28 years respectively. The native language of M1, M2, M3 and F1 was American English, Kannada, Bengali and Tamil respectively. None of the subjects were reported to have any speech disorders. Prior to data collection, an informed consent was obtained from each subject. The data collection was approved by the ethics committee of the Indian Institute of Science (IISc), Bangalore.

Movements of the articulators were recorded with an Electromagnetic articulograph, namely, AG501 [33], which is a widely used machine for articulatory movements recording. AG501 has 24 channels to measure both the displacement and angular orientation of a maximum of 24 sensors in horizontal, vertical and lateral directions at a sampling rate varying from 250Hz to 1250Hz. In this study, we use eight sensors placed at different articulators to get the data at a rate of 250Hz. Out of eight sensors, two are connected at the

back of two ears which are utilized for head correction. The remaining six sensors are used to record the articulatory movements in the midsagittal plane. Three sensors are attached on the articulators outside the oral cavity (upper lip (UL), Lower Lip (LL) and Jaw) and the rest are placed inside the oral cavity (Tongue Tip (TT), Tongue Body (TB) and Tongue Dorsum (TD)). A schematic diagram of the placement of different sensors are shown in Fig. 1. It should be noted that, the sensor movements along the midsaggittal plane are captured by the X and Y co-ordinates of the positional data provided by AG501, which we use in the present study. Thus we obtain a twelve dimensional articulatory feature vector representing sensor positions namely, $UL_x, UL_y, LL_x, LL_y, Jaw_x, Jaw_y, TT_x, TT_y, TB_x, TB_y, TD_x, TD_y$.

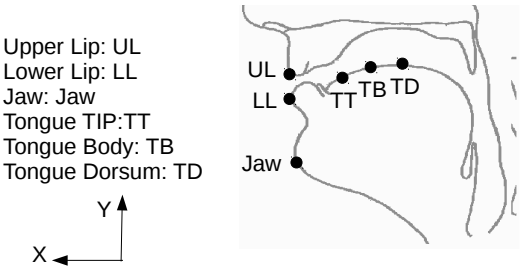

Fig. 1. A schematic diagram demonstrating the placement of six EMA sensors

In order to get a phonetically balanced dataset, the 460 English sentences from MOCHA-TIMIT [34] are chosen as the stimuli for the data collection. Since the native language of each subject is different, they are familiarized with the 460 sentences prior to recording. This is done to avoid any pronunciation error, word insertion and deletion during recording.

The data collection is carried out in a sound proof AG501 recording facility at the institute. After the attachment of sensors onto the subjects' articulators, subjects are provided with sufficient time to get adjusted to speaking comfortably in the presence of sensors. This is done through natural conversation and by reading a few exemplar sentences. The sentences are projected onto a screen in front of the subject at a distance of three meters during recording. The subjects could navigate across different sentences by themselves. A t.bone EM9600 shotgun, unidirectional electret condenser microphone [35] is placed near the subject to simultaneously record the audio data at a rate of 48kHz, synchronous with the articulatory movement data. Simultaneous audio and articulatory movement recording is done sentence by sentence. Each utterance is carefully scrutinized during the recording and the subjects are asked to repeat in case of any errors. Since, for a set of 460 stimuli, the average time taken for the recording turned out to be $\sim$ 2 hours, the entire stimuli of 460 sentences is recorded in neutral and whispered speech in two different sessions for the convenience of the subjects. Since, whispered speech has lower intensity, there is a need for sound pressure level calibration [36]. For this, we use a pure tone of known intensity and a TES-1350A sound level meter to obtain the sound pressure levels for every 100 sentences. As the recordings of neutral and whispered speech are done in two different sessions, proper care is taken to place the sensors in almost the same positions for the both recordings. The subjects are given breaks whenever they reported tiredness. For M1, M2, M3, and F1 the duration of the collected data after removing silences before and after the sentences turned out to be 18.90, 24.38, 21.83, 20.23 minutes for neutral speech and 21.87, 25.85, 24.52, 21.57 minutes for whispered speech, respectively.

3. ACOUSTIC-TO-ARTICULATORY INVERSION

Acoustic-to-articulatory inversion (AAI) is a regression problem, where the relationship between input (acoustic features) and output (articulatory features) is known to be nonlinear [18]. Since a DNN can efficiently learn such a non-linear mapping, we follow a strategy similar to that proposed by Wu et al [18]. Consider a DNN with L layers, such that, the first and the last layers correspond to the input layer and the output linear regression layer, respectively. Therefore, given an input vector $\mathbf{x}$ at the first layer, we obtain the predicted output vector $\mathbf{y}_L$ at the output layer, L. The output of the l^{th} hidden layer $\mathbf{y}_l$, given the weight matrix $\mathbf{W}_l$ and hidden bias $\mathbf{b}_l$ is given by,

$$\mathbf{y}_l(\mathbf{x}) = \phi(\mathbf{n}_l(\mathbf{x})), l = 2, \ldots, L-1, \quad (1)$$

such that,

$$\mathbf{n}_l(\mathbf{x}) = \mathbf{W}_l \mathbf{y}_{l-1}(\mathbf{x}) + \mathbf{b}_l, \quad (2)$$

where, ϕ is the activation function. We define $\mathbf{d}$ to be the desired output vector for training the DNN. Following an approach similar to Wu et al. [18], we define the objective function to be minimized as the mean squared error between the desired $\mathbf{d}$ and the predicted $\mathbf{y}_L$. The weights of the DNN are learnt by the back-propagation algorithm. The weights are updated using ADAM [37], a first-order gradient-based optimization method, based on adaptive estimates of lower-order moments, suitable for stochastic objective functions. ADAM is a computationally efficient algorithm with less memory requirements. DNN is implemented by using keras [38] and theano [39] libraries.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

The AAI is performed separately for neutral and whispered speech in a subject dependent manner in which non-overlapping training and test data are taken from the same subject. For every subject AAI is performed in a 10-fold cross-validation setup where the entire set of 460 sentences are divided into 10 groups among which eight groups are used for training, one group for validation and the remaining one for testing. The recorded speech is downsampled to 16kHz and silence before and after the sentence in every recorded utterance is removed since, during silence, articulators can take any position causing more variability in the inverse mapping. As acoustic features, we compute a 39-dim Mel frequency cepstral coefficients (MFCC) vector for a window size of 20ms and a frame shift of 10ms followed by cepstral mean subtraction and variance normalization [40]. In order to utilize the context information in the acoustic features for the DNN training, MFCCs from five frames before and after every frame are concatenated resulting in a 429-dim feature vector. The articulatory position data have high frequency noise resulting from EMA measurement error, but the articulatory movements are predominantly low-pass in nature [15]. Hence, the 12-dim articulatory movement data is low-pass filtered with a cut-off frequency of 25Hz as most of the energy of the articulatory movements is below 25Hz for all articulators. The articulatory data is further downsampled to 100Hz to obtain frame synchronized MFCC and articulatory feature vectors. Since the average position for each sensor could change from utterance to utterance [41], we subtract the mean and divide by the standard deviation (SD) within every utterance for each dimension of the articulatory feature vector.

For DNN, we have chosen a configuration of 3-hidden layers with 300 units in each layer. The sigmoid function is used as the activation function ϕ. 429-dim MFCC vector is given as the input to

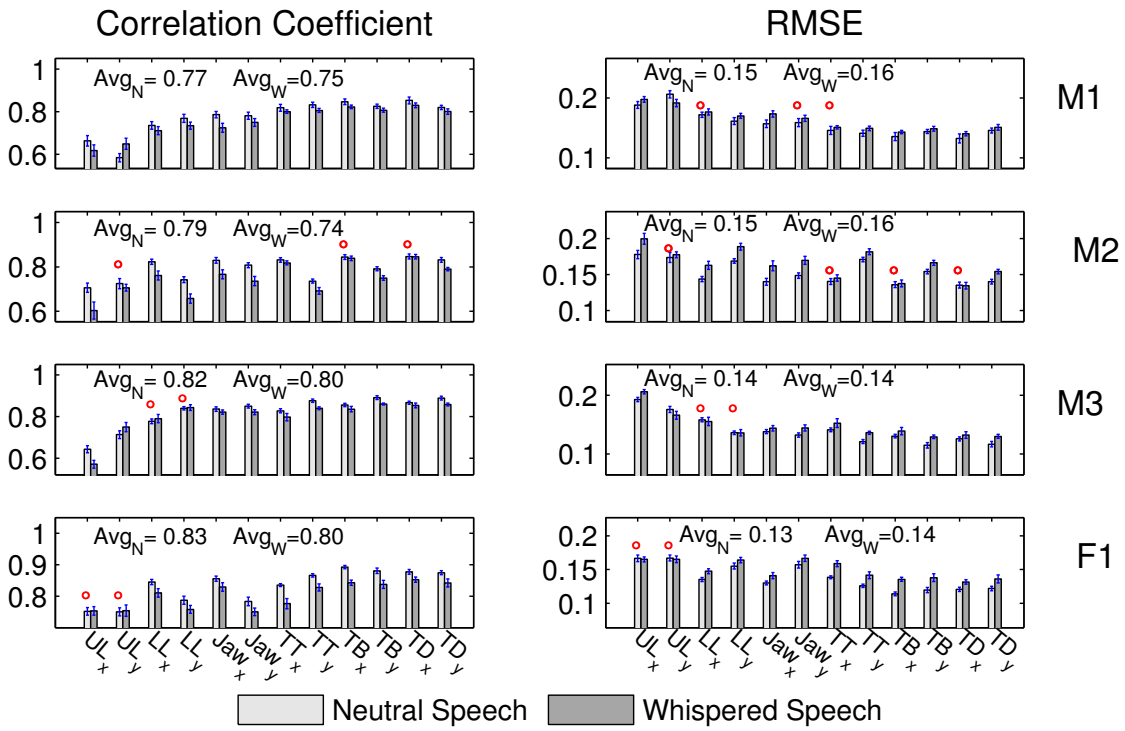

Fig. 2. Correlation coefficient (ρ) and RMSE for AAI for neutral and whispered speech for four subjects in matched condition. Avg_N and Avg_W denote ρ (column 1) and RMSE (column 2) averaged over all articulators. ○ denotes the articulatory features for which there is no significant (p <0.01) difference in AAI performance between neutral and whispered speech types.

first layer. The output layer is a linear regression layer with 36 units corresponding to 12 articulatory features along with their velocity and acceleration coefficients. The parameters for ADAM are chosen as follows: learning rate=0.001, β_1=0.9, β_2=0.999 (β_1, $\beta_2 \in [0, 1)$: exponential decay rates for the moment estimates), $\epsilon = 1e\text{-}08$ and batch size of 128. For each fold, DNN weights are learnt using the MFCCs and articulatory data from the training set.

For comparative study of neutral and whispered speech, we choose two evaluation metrics, namely, the Root Mean Square Error (RMSE) and the correlation coefficient (ρ) [15] for each articulator separately. The predicted articulatory trajectory from DNN is typically jagged in nature since no inter-frame smoothness criterion explicitly employed while training the DNN. However, realistic articulatory trajectories are smooth in nature. Thus, we low-pass filter each articulatory trajectory predicted by the trained DNN. The cut-off frequency of the low-pass filter is learnt using the validation set. For this purpose, a range of cut-off frequencies from 5 to 25Hz with a step of 1Hz is chosen. For each articulatory feature, the best cut-off frequency is chosen by finding that frequency for which the RMSE of the corresponding feature is minimum in the validation set.

Apart from examining AAI performance in a matched condition (i.e., both train and test are either whisper or neutral data), we examine the accuracy with which an AAI model, trained with one type of speech, performs when the other type of speech is presented as the test case (mis-matched condition). Suppose, for the i-th articulator, ρ_o^i and ρ_c^i be the correlation coefficients in the matched and mis-matched conditions respectively. We compute the percentage drop in correlation coefficient (PDCC) as follows: $PDCC_i = \frac{(\rho_o^i - \rho_c^i)}{\rho_o^i} \times 100$. PDCC is not reported for RMSE since the sensor positions may not be identical in neutral and whispered speech recordings and, hence, the RMSE between original and predicted articulatory positions may not be directly comparable.

4.2. Results and discussion

Matched train-test condition: The barplot in Fig. 2 shows the average and SD of ρ and RMSE of AAI for twelve articulators separately for four subjects considered in this study. The bar height indicates the average value across ten folds while the errorbar shows SD. For comparing the AAI performance between the neutral and the whispered speech, two bars are plotted adjacent to each other for each articulatory feature – the light and dark color bars denote the neutral and whispered speech cases respectively. It is clear that there is a significant drop in the ρ and significant increase in the RMSE values when AAI is performed for the whispered speech compared to the neutral speech for most of the articulatory features for all subjects. For a few articulators, there is a significant drop in performance across all subjects. For example, there is a significant ($p < 0.01$) drop in ρ for Jaw_x, Jaw_y, TT_x, TT_y, TB_y and TD_y consistently for all subjects with an a relative drop of 5.1%, 5.1%, 3.6%, 4.4%, 3.9%, and 3.6% respectively. Similarly, there is a significant ($p < 0.01$) increase in RMSE for Jaw_x, TT_y, and TB_y consistently for all subjects with a relative increase of 9.8%, 9.4%, and 10% respectively. This indicates that, irrespective of the subject, both ρ and RMSE for Jaw_x, TT_y and TB_y deteriorate when in the whispered speech is used for AAI compared to when the neutral speech is used. Interestingly, we see a subject specific deterioration in the two metrics across different articulators. For example, the largest

drop in ρ occurs for Jaw_x (7.9%), UL_x (14.5%), UL_x (11.2%), and TT_x (7.0%) for M1, M2, M3, and F1 respectively. Similarly, the largest increase in RMSE occurs for UL_y (7.0%), TD_x (0.5%), UL_y (5.6%) and UL_y (1.18%) for these four subjects respectively. Although all subjects are fluent in English, variability in the results, across subjects, could be due to the effect of different native languages (L1), which requires further investigation. The ρ and RMSE averaged over all articulators are also shown in the respective subplots in Fig. 2; these are denoted by Avg_N and Avg_W, for neutral and whisper respectively. It is clear that, on average, the ρ drops by 0.02, 0.05, 0.02, and 0.03 (absolute) and RMSE increases by 0.01, 0.01, 0.00, and 0.01 (absolute) for M1, M2, M3, and F1 respectively. This suggests that although whispered speech lacks voicing and is less intelligible compared to neutral speech, the information about the articulatory movements could be encoded in the spectral characteristics of whispered speech.

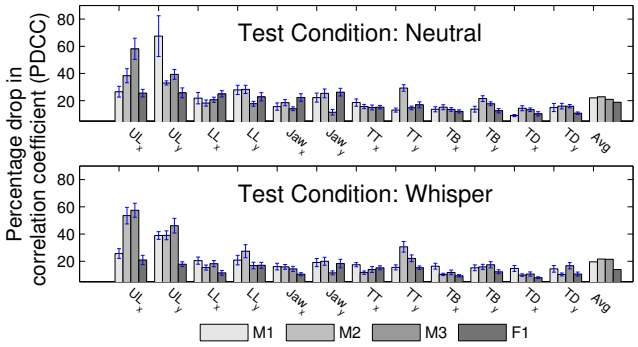

Fig. 3. PDCC for different articulatory features in mis-matched condition compared to matched condition for all four subjects. The top and bottom row correspond to neutral and whisper test conditions. The last four bars in each row show the average of PDCC across twelve articulatory features.

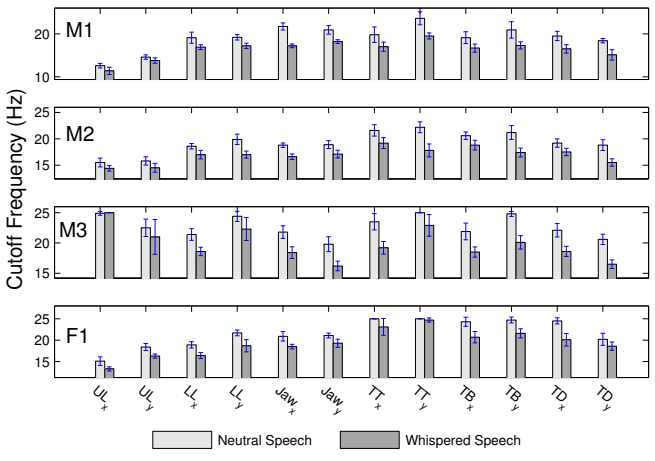

Fig. 4. Cut-off frequency optimized for each articulator on the validation set for both neutral and whispered speech in matched condition for four subjects. Bar height shows the cut-off frequency averaged over 10 folds while errorbars indicate SD.

Mis-matched train-test condition: Fig. 3 shows the PDCC values for each articulator for both neutral and whisper test conditions in case of four subjects. The bar height indicates the PDCC value averaged over 10 folds, while the errorbar indicates the SD. It is interesting to note that there is a consistent drop in the correlation coefficients from the matched condition to the mismatched condition for both neutral and whisper test cases. This indicates that the acoustic-to-articulatory map for neutral and whispered speech are different and one may not be used to predict the articulatory motion in case of the other. The PDCC averaged across twelve articulators (last group of four columns in Fig. 3) indicates that the average PDCCs are approximately 20% and they are similar for neutral and whisper test conditions. Among twelve articulatory features, the PDCC is more for UL_x and UL_y compared to others. This suggests that the movement of UL could be significantly different in neutral and whispered speech. This could also be due to the differences in the manner in which the UL movement is encoded in neutral and whispered speech.

Smoothness of articulator movement: The optimal cut-off frequencies obtained from the validation set for different articulators are shown in Fig. 4 for both neutral and whispered speech for four subjects. It is clear that the optimal cut-off frequency varies across different articulators. This is because different articulators are smooth to different degrees. Even for the same articulator the optimal cut-off frequency varies across subjects. However, it is interesting to observe that consistently for all articulators and subjects, the optimal cut-off frequency for neutral speech is higher than that for the whispered speech. This indicates that the articulatory dynamics for whispered speech is more smooth compared to that for the neutral speech. This is also reflected in the reduced speaking rate of the subjects for the whispered speech compared to that for the neutral speech. For example, the average neutral phoneme rates (in phonemes per second) are 12.53, 9.83, 10.86, and 11.79 for M1, M2, M3, F1 respectively, while these are 10.81, 9.19, 9.66, and 11.05 for whispered speech. This could be because of exaggerated articulatory movements [30] and increase in the utterance duration [36] that characterize whispered speech.

5. CONCLUSIONS

We perform a comparative study of AAI for neutral and whispered speech using three male and one female subjects. We observe that the articulatory movement is smoother for whispered speech compared to that for neutral speech. It is also found that the AAI performance drops significantly for Jaw_x, TT_y, TB_y in the case of whispered speech compared to the neutral speech. Drop in the AAI performance is observed for the whispered speech when the acoustics and articulation of neutral speech are used for training and vice-versa. This suggests that the acoustic-to-articulatory mapping of whispered speech is different from that of the neutral speech. Experiments also reveal that although the information of the articulatory movements is retained in whispered speech, it is encoded differently, compared to that in neutral speech. Further investigation is required to examine the manner in which articulation during whispered speech could be different from that for neutral speech and develop an adaptation technique for both acoustics and articulation so that AAI on the whispered (neutral) speech could be improved when the AAI model is trained using the neutral (whispered) speech acoustics and articulation.

Acknowledgement: We thank all subjects who participated in EMA data collection.

6. REFERENCES

[1] Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang, "Integrating articulatory features into hmm-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[2] Chuang Ding, Lei Xie, and Pengcheng Zhu, "Head motion synthesis from speech using deep neural networks," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9871–9888, 2015.

[3] Jia Jia, Zhiyong Wu, Shen Zhang, Helen M Meng, and Lianhong Cai, "Head and facial gestures synthesis using pad model for an expressive talking avatar," *Multimedia Tools and Applications*, vol. 73, no. 1, pp. 439–461, 2014.

[4] Jia Jia, Shen Zhang, Fanbo Meng, Yongxin Wang, and Lianhong Cai, "Emotional audio-visual speech synthesis based on pad," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 3, pp. 570–582, 2011.

[5] Jiping Sun and Li Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *The Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1086–1101, 2002.

[6] Igor Zlokarnik, "Adding articulatory features to acoustic features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3246–3246, 1995.

[7] Alan Wrench and Korin Richmond, "Continuous speech recognition using articulatory data," *Proceedings of the International Conference on Spoken Language Processing*, pp. 145–148, 2000.

[8] Joe Frankel, Korin Richmond, Simon King, and Paul Taylor, "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces," *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China (CD-ROM) 2000.

[9] Katrin Kirchho, *Robust speech recognition using articulatory information*, Ph.D. thesis, University of Bielefeld, 1999.

[10] Slim Ouni and Yves Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[11] Bishnu S Atal, Jih Jie Chang, Max V Mathews, and John W Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, 1978.

[12] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[13] Korin Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping.," in *in Proceedings of the ICSLP, Pittsburgh*, 2006, pp. 577–580.

[14] Le Zhang and Steve Renals, "Acoustic-articulatory modeling with the trajectory hmm," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[15] Prasanta Kumar Ghosh and Shrikanth Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[16] Simon King and Paul Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000.

[17] Benigno Uria, Steve Renals, and Korin Richmond, "A deep neural network for acoustic-articulatory speech inversion," in *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[18] Zhiyong Wu, Kai Zhao, Xixin Wu, Xinyu Lan, and Helen Meng, "Acoustic to articulatory mapping with deep neural network," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9889–9907, 2015.

[19] Szu-Chen Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, March 2005, vol. 1, pp. 1009–1012.

[20] R Netsell and B Daniel, "Dysarthria in adults: physiologic approach to rehabilitation," *Archives of physical medicine and rehabilitation*, vol. 60, no. 11, pp. 502–508, November 1979.

[21] H. Hirose, "Pathophysiology of motor speech disorders (dysarthria)," *Folia Phoniatr Logop*, vol. 38, pp. 61–88, 1986.

[22] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, *Regeneration of Speech in Voice-Loss Patients*, pp. 1065–1068, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[23] Vivien C Tartter, "Whats in a whisper?," *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[24] Robert W Morris and Mark A Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7, pp. 515–520, 2002.

[25] G. N. Meenakshi and P. K. Ghosh, "Robust whisper activity detection using long-term log energy variation of sub-band signal," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1859–1863, Nov 2015.

[26] Siobodan T Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *Acta Acustica united with Acustica*, vol. 84, no. 4, pp. 739–743, 1998.

[27] Masahiro Matsuda and Hideki Kasuya, "Acoustic nature of the whisper.," in *Eurospeech*, 1999, vol. 99, pp. 133–136.

[28] Man Gao, *Tones in whispered Chinese: articulatory features and perceptual cues*, Ph.D. thesis, University of Victoria, 2002.

[29] Gordon E Peterson, "Parameters of vowel quality," *Journal of Speech, Language, and Hearing Research*, vol. 4, no. 1, pp. 10–29, 1961.

[30] Hirohide Yoshioka, "The role of tongue articulation for /s/ and /z/ production in whispered speech," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, 2008.

[31] Megan Jo Osfar, "Articulation of whispered alveolar consonants," M.S. thesis, University of Illinois at Urbana-Champaign, 2011.

[32] Masahiko Higashikawa, Jordan Green, Christopher Moore, and Fred Minifie, "Lip kinematics for /p/ and /b/ production during whispered and voiced speech," *Folia Phoniar Logop*, vol. 55, pp. 1–9, 2003.

[33] "3d electromagnetic articulograph, available online: http://www.articulograph.de/, last accessed:7/9/2016," .

[34] A. Wrench, "MOCHA-TIMIT," speech database, Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, 1999.

[35] "Em 9600 shotgun microphone, available online: http://www.tbone-mics.com/en/product/information/details/the-tbone-em-9600-richtrohr-mikrofon/, last accessed:23/12/2016," .

[36] Chi Zhang and John HL Hansen, "Analysis and classification of speech mode: whispered through shouted.," in *INTERSPEECH*, 2007, pp. 2289–2292.

[37] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[38] Franois Chollet, "keras," https://github.com/fchollet/keras, 2015.

[39] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016.

[40] Fu-Hua Liu, Richard M Stern, Alejandro Acero, and Pedro J Moreno, "Environment normalization for robust speech recognition using direct cepstral comparison," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. IEEE, 1994, vol. 2, pp. II–61.

[41] Korin Richmond, *Estimating articulatory parameters from the acoustic speech signal*, Ph.D. thesis, University of Edinburgh, 2002.