A STUDY ON ROBUSTNESS OF ARTICULATORY FEATURES FOR AUTOMATIC SPEECH RECOGNITION OF NEUTRAL AND WHISPERED SPEECH

Gokul Srinivasan

Birla Institute of Technology and Science
Electrical and Electronics Engineering
Pilani, India

Aravind Illa, Prasanta Kumar Ghosh

Indian Institute of Science
Electrical Engineering
Bangalore, India

ABSTRACT

Traditionally, automatic speech recognition (ASR) systems are trained on acoustic representations of neutral speech. As a result, their performance degrades when tested with whispered speech. In this work, we explore the robustness of articulatory features in ASR of neutral and whispered speech. We use acoustic, articulatory, and integrated acoustic and articulatory feature vectors in matched and mismatched train-test cases. The results suggest that the articulatory data is useful in ASR of both neutral and whispered speech, especially in the mismatched train-test cases. When we concatenate acoustic and articulatory feature vectors and deploy it to the mismatched train-test case where the model is trained with neutral speech and tested with whispered speech, a relative improvement in phone error rate of 27.2% is observed compared to when only acoustic features are used. This suggests that articulatory data contains information complementary to acoustic representations. A phone specific recognition error is also presented which illustrates phones where adding articulatory information gives maximum benefit.

***Index Terms*—** Automatic speech recognition, neutral speech, whispered speech, articulatory data

1. INTRODUCTION

With advancements in deep neural networks (DNN), modern automatic speech recognition (ASR) systems have shown state-of-the-art performances in acoustically matched train-test cases. However, ASR performance is often affected by mismatch in train-test cases due to various factors including surrounding environment (noise, reverberation, loudness, etc.) [1], speech rate, accent, and dialect variations [2]. Further, speech can also be produced with different modality, such as whispered speech, in contrast to normally phonated (neutral) speech. Examples of such scenarios are private conversations, and when the user has a pathological condition [3, 4]. Without large amounts of whispered training data in addition to neutral speech, it is challenging to train an ASR system that is robust enough to recognize both neutral and whispered speech [5].

The challenges are due to the differences in acoustic characteristics and articulation between neutral and whispered speech. Whispered speech lacks vocal-fold vibrations, and hence, pitch [6, 7]. It has noise-like characteristics and lower signal-to-noise ratio (SNR) than neutral speech [8]. There is also a shift of formants in low frequencies, compared to those in neutral speech [9, 10]. This results in a mismatch in acoustic characteristics between neutral and whispered speech. Therefore, ASR systems trained with acoustic representations of neutral speech do not perform well when tested with whispered speech.

Previous works have shown that articulatory data is useful for ASR [11, 12, 13, 14], particularly when there is background noise [15], in conversational speech [16], and when the speaker has a pathological condition [17]. Results reported by Szu-Chen Jou et al. [18] suggest that having articulatory feature representation helps in the recognition of whispered speech. However, the articulatory features were derived using IPA phonological features and then adapted, rather than measured directly, which could have limited the evaluation procedure [18]. In this work, we conduct a study on the robustness of articulatory data for the ASR of neutral and whispered speech in matched and mismatched train-test cases, with direct measurement of articulatory data using Electromagnetic Articulography (EMA) [19]. Table 1 shows the definition of the different matched and mismatched train-test cases considered in this study. While Beiming et al. [20] used articulatory movement data from a single patient with a surgically reconstructed larynx for automatic whispered speech recognition, there has been no systematic investigation with directly measured articulatory data from multiple subjects in ASR of whispered speech.

It has been found that there are some differences in the articulation of neutral and whispered speech [21, 22, 23, 24]. However, we believe that the relative invariance of articulatory data between neutral and whispered speech, compared to acoustic representations [15], could be exploited to improve the ASR performance in the mismatched train-test cases by using articulatory data in addition to acoustic features. We present the results when three different types of feature vectors are used for ASR, for all four train-test cases. The three types of feature vectors are - (1) acoustic features in the form of mel frequency cepstrum coefficients (MFCCs), (2) articulatory features (AFs), and (3) integrated acoustic and articulatory features (AF+MFCCs). Experiments are performed with $\sim$8 hours of parallel acoustic and articulatory data from neutral and whispered speech in matched and mismatched train-test cases as listed in Table 1.

The rest of the paper is organized as follows: Section 2 contains details regarding data collection. Section 3 is divided into two subsections – section 3.1 contains specifics of the experimental setup

		Test	
		Neutral	Whispered
Train	Neutral	Matched (Neu-Neu)	Mismatched (Neu-Whi)
	Whispered	Mismatched (Whi-Neu)	Matched (Whi-Whi)

Table 1: Matched and mismatched train-test cases for ASR in this study.

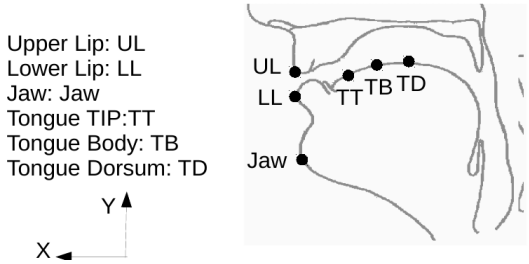

Fig. 1: A diagram showing placement of sensors which record articulatory movements in the mid-sagittal plane [24]

and section 3.2 contains discussions on the results obtained and an example to illustrate the benefits of articulatory data at the phone level. Section 4 contains the conclusion of the paper and scope for further study.

2. DATASET

Parallel acoustic and articulatory data were recorded for neutral and whispered speech. 10 subjects comprising 5 male and 5 female subjects participated in the data collection, in an age group of 20–28 years. None of the subjects were reported to have speech disorders. An informed consent was obtained from all subjects before collecting the data. The data collection was approved by the ethics committee of the Indian Institute of Science (IISc), Bangalore.

Articulatory movements were recorded with AG501, an Electromagnetic articulograph [25], which is a state-of-the-art instrument for recording articulatory movements. In this study, eight sensors were placed at different articulators and the sampling rate was set to 250Hz. Two sensors were attached behind the ears and were used for head motion correction. The rest of the six sensors were used to record articulatory movements in the midsagittal plane. Three of these sensors were placed outside the oral cavity, on the upper lip (UL), lower lip (LL), and jaw (Jaw). The remaining three were placed inside the oral cavity on the tongue tip (TT), tongue body (TB), and tongue dorsum (TD), following the guidelines provided in [26]. A diagram indicating the placement of sensors is shown in Fig. 1. Movement of each sensor in the midsaggital plane was captured using X and Y coordinates of the positional data provided by the AG501. From these X and Y coordinates of six sensors, we obtained a 12 dimensional fecture vector with elements denoted by UL_x, UL_y, LL_x, LL_y, Jaw_x, Jaw_y, TT_x, TT_y, TB_x, TB_y, TD_x, and TD_y.

The 460 phonetically balanced English sentences from the MOCHA-TIMIT dataset [27] were chosen as stimuli for all the recordings. The phonetic transcription of the dataset consists of 39 ARPABET symbols used for evaluation of models on the TIMIT dataset [28], along with $[td]$, $[kd]$, $[pd]$, $[dd]$, $[gd]$, $[bd]$ which denote unreleased stops, and $[ts]$, which is a voiceless alveolar affricate. To avoid erroneous pronunciation, and insertion or deletion of words, the subjects were familiarized with the sentences prior to recording. The data was collected in an AG501 recording facility at IISc, Bangalore. Subjects were given ample time to accustom themselves to the presence of sensors while they speak. This was done by conversing with the subjects for some time, and making them comfortable reading sentences. Audio was recorded with a t.bone EM9600 shotgun unidirectional electret condenser microphone [29] at a sampling frequency of 48 kHz. Acoustic and articulatory data were recorded simultaneously for every utterance. There was careful scrutiny during the recording, and the subject was asked to repeat the utterance in case of any error or ambiguity. The recording time for a single subject was $\sim$2 hours. Due to the lengthy recording duration, the recordings of neutral speech and whispered speech were done in two separate sessions for the convenience of the subjects. As whispered speech has low intensity, we used sound pressure level calibration during whispered speech recordings [30]. A single tone of known intensity and a TES-1350A sound level meter were used to measure the sound pressure levels for every 100 sentences. Care was taken to place the sensors in almost identical positions in the two recording sessions. Subjects were allowed to take a break in the middle of the recording as many times as they wanted. In the event that the sensors came off the articulators, proper care was taken to re-glue them in the identical position. In order to avoid error due to mismatch in the location after re-glueing, average sensor location was subtracted separately for each utterance in the data processing step. The total duration of the recorded data after the removal of silences before and after every sentence turned out to be 224 minutes and 238 minutes for neutral and whispered speech, respectively. Mean and standard deviation (SD) of duration of recorded data per subject was 22.43 ($\pm$2.63) minutes for neutral speech and 23.85 ($\pm$2.42) minutes for whispered speech.

3. EXPERIMENTS AND RESULTS

3.1. Experimental setup

The speech recorded using AG501 was down-sampled to 16 kHz. Then, acoustic features comprising 13 MFCCs were computed using a window size of 20ms with a frame shift of 10ms. This was followed by cepstral mean subtraction and variance normalization [31]. To avoid high-frequency noise due to EMA measurement error, the articulatory data was low-pass filtered with a cut-off frequency of 25Hz, as most of the energy of the articulatory movements of all chosen articulators lies below 25Hz [32]. The articulatory data was downsampled to 100Hz so that the feature vector is frame synchronized with the MFCC feature vector. Motivated by articulatory phonology [33], from the articulatory trajectories of lips, we derived lip protrusion ($LPro$) and lip aperture (LA) as two articulatory features [34]. $LPro$ captures the horizontal distance of lips from front teeth, obtained from horizontal trajectories of UL_x and LL_x as $LPro = (UL_x + LL_x)/2$. LA indicates the vertical distance between upper lip and lower lip, $LA = |UL_y - LL_y|/2$. $LPro$ and LA along with Jaw_x, Jaw_y, TT_x, TT_y, TB_x, TB_y, TD_x, and TD_y result in a 10-dimensional AF vector. As the average position of articulators could change across utterances, we subtracted the mean and divided by SD for each dimension of the AF vector within each utterance. On concatenating AF and MFCC feature vectors, we obtained a 23-dimensional AF+MFCC feature vector.

The acoustic model used was based on the model proposed by Veselý et al. for the TIMIT dataset [35]. Our model was built using the Kaldi Speech Recognition toolkit. The DNN consisted of an input layer, six hidden layers, and an output layer. The input layer, hidden layers, and output layer had dimensions of 440, 2048, and 3370, respectively. Neurons in all layers except the output layer used the sigmoid activation function. The neurons in the output layer used the softmax activation function. The DNN was trained with LDA-MLLT-fMLLR features [36].

All the experiments were performed using a 10-fold cross-validation setup. The 460 sentences spoken by each speaker were divided into 10 sets containing the same number of sentences, out of which eight sets were used for training, one for validation, and one

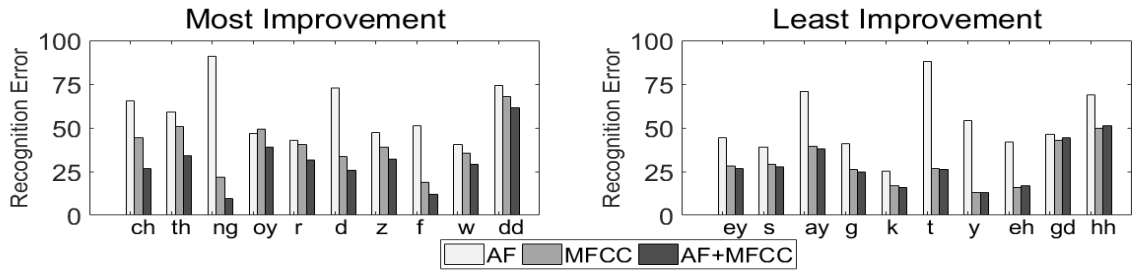

Fig. 2: Neu-Neu train-test case: phones arranged in decreasing order of improvement after the addition of AFs to MFCCs.

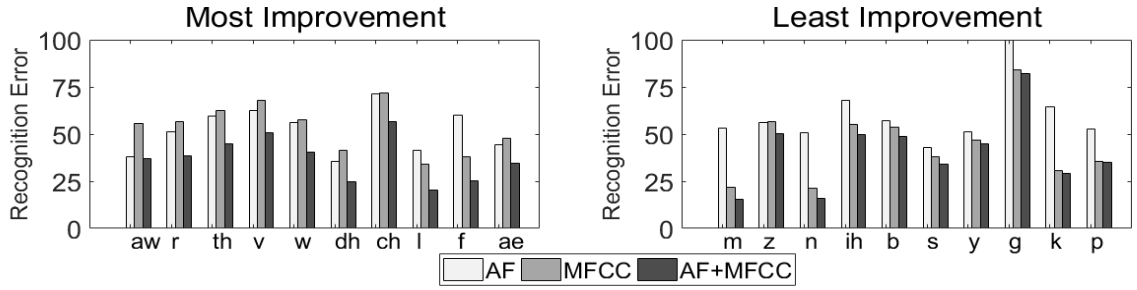

Fig. 3: Whi-Whi train-test case: phones arranged in decreasing order of improvement after the addition of AFs to MFCCs.

Train-test case	AF	MFCC	AF+MFCC
Neu-Neu	47.0 ($\pm$1.22)	29.3 ($\pm$0.81)	27.7 ($\pm$1.21)
Whi-Whi	50.6 ($\pm$0.81)	42.9 ($\pm$0.82)	37.7 ($\pm$1.07)
Neu-Whi	55.0 ($\pm$0.73)	66.3 ($\pm$1.10)	48.3 ($\pm$1.11)
Whi-Neu	51.2 ($\pm$0.83)	46.1 ($\pm$0.80)	37.8 ($\pm$0.85)

Table 2: PER (in %) for all train-test cases. Values in brackets indicate SD across 10 folds (in %)

for testing in a round-robin fashion. The mean durations of training, testing, and validation sets were 190.8, 23.85, and 23.85 minutes, respectively, for neutral speech, and 179.44, 22.43, and 22.43 minutes, respectively, for whispered speech. The performance of the ASR system for all matched and mismatched cases have been reported as the phone error rate (PER) averaged across the ten folds, along with its SD. As the recorded dataset was not large enough (2015 unique words) to build a word-level language model (LM), we built a phone-level bigram LM. For measuring the performance of the models on individual phones, we computed the phone recognition error. It is computed as: Recognition Error $\% =$

$$\left(\frac{\text{Substitution Errors} + \text{Insertion Errors} + \text{Deletion Errors}}{\text{Total number of occurrences of the phone across all folds}}\right) \times 100,$$

where substitution, insertion, and deletion errors are summed across all folds for the phone in consideration. PER and recognition errors are reported with respect to the 46 unique phones in the dataset.

3.2. Results and Discussion

Matched train-test case: The first two rows of Table 2 report the PER of matched train-test cases. When MFCCs were used as features, there was a relative increase in PER of 46.4% in Whi-Whi compared to Neu-Neu. This was predominantly due to lack of voicing cues in whispered speech [5]. When only AFs were used as a feature vector, there was an increase in PER compared to the MFCCs case in both matched train-test cases. We also observed that there was a relative increase of 7.6% in PER from the Neu-Neu to Whi-Whi train-test case when only the AF vector was used. We are unclear about the aspects that could have led to the drop in AF performance in Whi-Whi compared to Neu-Neu, which requires further investigation. In the case where AF+MFCCs were used, we observed a relative reduction in PER for both Neu-Neu (5.4%) and Whi-Whi (12.1%) cases, compared to the MFCC feature vector. Also, the relative increase in PER from the Neu-Neu to Whi-Whi case was 36.1%, which is 10.3% less than the corresponding increase in PER in the MFCC case.

We sorted phones in descending order with respect to the reduction in recognition error from the AF+MFCCs to the MFCCs case. The recognition errors of ten phones at the top (most reduction in PER) and bottom (least reduction in PER) of this sorted list are shown in Fig. 2 (Neu-Neu) and Fig. 3 (Whi-Whi) for the different feature vectors. If the recognition error reduced significantly after the addition of AFs, it means that substantial information about that phone was encoded in AFs that was not present in MFCCs.

Mismatched train-test case: The last two rows of Table 2 report the PER of mismatched train-test cases. While comparing the performance degradation of the mismatched cases to their matched counter parts, we observed that there was a relative increase in PER of 57.3% (Neu-Neu vs Whi-Neu) and 54.5% (Whi-Whi vs Neu-Whi) when the MFCC feature vector was used. When AFs were used, we observed that the PER increased relatively by 8.0% in both Neu-Neu vs Whi-Neu and Whi-Whi vs Neu-Whi cases. This drop could be due to differences in articulation between neutral and whispered speech [21, 22, 23, 24]. Note that there was no severe degradation in PER due to mismatch of train-test cases when AFs were used (8.0%), compared to MFCCs ($\sim$55.9%). This could be because variance in features across the modes of speech is lesser in articulatory space compared to acoustic space [15]. After concatenating AFs with MFCCs, we observed that PER is reduced compared to the MFCC

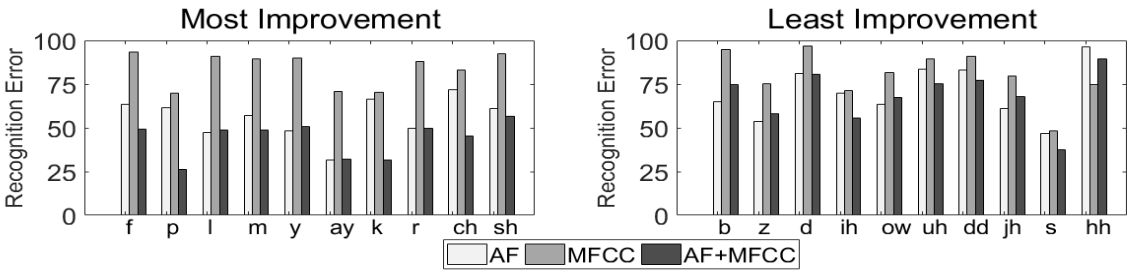

Fig. 4: Neu-Whi train-test case: phones arranged in decreasing order of improvement after the addition of AFs to MFCCs.

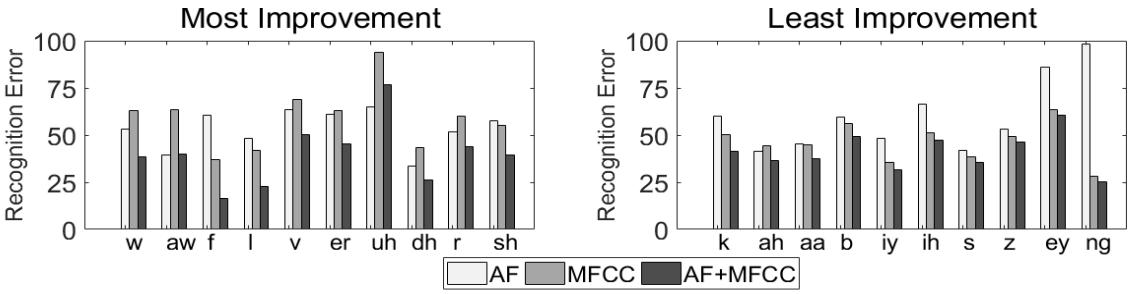

Fig. 5: Whi-Neu train-test case: phones arranged in decreasing order of improvement after the addition of AFs to MFCCs.

Substitution	AF	MFCC	AF+MFCC
$[m]$ ->$[n]$	41	148	32
$[ey]$ ->$[ih]$	192	192	151
$[w]$ ->$[f]$	212	207	132

Table 3: Count of few phone substitutions for Whi-Whi case using models trained on MFCC, AF, and AF+MFCC.

case. In particular, we observed a relative improvement of 27.2% in the Neu-Whi and 18.0% in the Whi-Neu case. We attribute this improvement to AFs capturing a significant amount of information that was independent of the mode of speech. This result could help ASR models, which are often trained only on neutral speech, to recognize whispered speech more accurately.

Fig. 4 and Fig. 5 show plots similar to Fig. 2 and Fig. 3 for Neu-Whi and Whi-Neu cases respectively. From Fig. 4, we see that there was complementary information being captured by AFs that reduced the recognition error for all phones except $[hh]$ in the Neu-Whi case. The recognition errors dropped by more than 30% for all top ten phones (most improvement) in Fig. 4 after AFs were added to MFCCs, compared to using just MFCCs.

To illustrate the benefit of AFs at the phone level, we present few examples of phone substitutions in Table 3 for the Whi-Whi case. From Table 3, it is clear that the number of substitutions of $[m]$ with $[n]$ dropped from 148 (MFCC) to 32 (AF+MFCC). This could be because articulatory differences between whispered $[m]$ and $[n]$, such as place of articulation [37], were captured by the AFs. Similarly, the number of substitutions of $[ey]$ with $[ih]$ dropped from 192 (MFCC) to 151 (MFCC+AF). This drop possibly occurred due to AFs capturing the change of positions of speech articulators during the glide between the two vowel sounds $[eh]$ and $[ih]$ in the diphthong $[ey]$ [38], which could have complemented the information present in the MFCCs. The number of substitutions of $[w]$ with $[f]$ dropped by 75 from the MFCC to AF+MFCC case. Due to the absence of voicing in whispered speech, the difference in acoustics between $[w]$ and $[f]$ may have become smaller. But, their distinct articulatory representations [38] could have helped the model to differentiate between them better.

4. CONCLUSIONS

A study was conducted to ascertain the usefulness of articulatory information in ASR of neutral and whispered speech. PER was found to vary less with mismatch in train-test cases when AFs alone were used, compared to MFCCs. This suggests that AFs are useful for ASR irrespective of the mode of speech. When AFs were concatenated with MFCCs to obtain an augmented feature vector, we found marginal improvements in PER in the matched train-test cases when compared to using only MFCCs. From recognition errors of individual phones, we observed that AFs carry information that is complementary to MFCCs. This result was very prominent in the Neu-Whi case where MFCC failed to capture significant information that was independent of the mode of speech. There was a substantial reduction in relative PER (27.2%) after the addition of AFs to MFCCs in the Neu-Whi case. This result could potentially help traditional ASR systems recognize whispered speech more accurately. Future work would involve conducting experiments similar to those done in this paper using articulatory data predicted from acoustic-to-articulatory inversion (AAI) [24], as opposed to directly measured articulatory data which is not always practical. We reported the empirical evaluation results for the individual phone recognition errors, however, further investigation is required to ascertain the nature of the complementary information carried by AFs, and to provide physiological explanations.

Acknowledgement: We thank Nisha Meenakshi G for her help in the recording process, Anurag Das and Avni Rajpal for their guidance related to Kaldi, and the Pratiksha Trust for their support.

5. REFERENCES

[1] Takahiro Fukumori, Masato Nakayama, Takanobu Nishiura, and Yoichi Yamashita, "Estimation of speech recognition performance in noisy and reverberant environments using PESQ score and acoustic parameters," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–4, 2013.

[2] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al., "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[3] Hajime Hirose, "Pathophysiology of motor speech disorders (dysarthria)," *Folia Phoniatrica et Logopaedica*, vol. 38, no. 2-4, pp. 61–88, 1986.

[4] H. R. Sharifzadeh, I. V. Mcloughlin, and F. Ahmadi, "Regeneration of speech in voice-loss patients," in *IFMBE 13th International Conference on Biomedical Engineering*, 2009, pp. 1065–1068.

[5] Dorde T. Grozdić and Slobodan T. Jovičić, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 12, pp. 2313–2322, 2017.

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

[7] Robert W. Morris and Mark A. Clements, "Reconstruction of speech from whispers," *Medical Engineering and Physics*, vol. 24, no. 7-8, pp. 515–520, 2002.

[8] Robert W. Morris, *Enhancement and Recognition of Whispered Speech*, Ph.D. thesis, Georgia Institute of Technology, 2003.

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

[10] Masahiro Matsuda and Hideki Kasuya, "Acoustic nature of the whisper," in *Sixth European Conference on Speech Communication and Technology, EUROSPEECH*, 1999, vol. 1, pp. 137–140.

[11] 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.

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

[13] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1913–1924, Sep. 2011.

[14] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[15] Katrin Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, University of Bielefield, 1999.

[16] Florian Metze, *Articulatory Features for Conversational Speech Recognition*, Ph.D. thesis, Karlsruhe Institute of Technology, 2005.

[17] Emre Yılmaz, Vikramjit Mitra, Chris Bartels, and Horacio Franco, "Articulatory features for ASR of pathological speech," in *INTERSPEECH*, 2018, pp. 2958–2962.

[18] Szu-Chen Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, vol. 1, pp. 1009–1012.

[19] Paul W. Schönle, Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader, and Bastian Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," 1987.

[20] Beiming Cao, Myung Jong Kim, Ted Mau, and Jun Yi Wang, "Recognizing whispered speech produced by an individual with surgically reconstructed larynx using articulatory movement data.," *Workshop on Speech and Language Processing for Assistive Technologies*, vol. 2016, pp. 80–86, 2016.

[21] G.N.Meenakshi and P.K.Ghosh, "Reconstruction of articulatory movements during neutral speech from those during whispered speech.," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3352–3352, 2018.

[22] 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.

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

[24] Aravind Illa, Prasanta Kumar Ghosh, et al., "A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5075–5079.

[25] "3D Electromagnetic Articulograph," `http://www.articulograph.de/`, last accessed: 20/10/2018.

[26] Ashok Kumar Pattem, Aravind Illa, Amber Afshan, and Prasanta Kumar Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," *Computer Speech & Language*, vol. 47, pp. 157–174, 2018.

[27] Alan A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research.," *Phonus.*, vol. 5, pp. 1–13, 2000.

[28] J. S. Garofolo, Lori Lamel, W. M. Fisher, Jonathan Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1992.

[29] "EM 9600 shotgun microphone," `http://www.tbone-mics.com/en/product/information/details/the-tbone-em-9600-richtrohr-mikrofon/`,last accessed: 20/10/2018.

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

[31] Fu-Hua Liu, Richard M Stern, Alejandro Acero, and Pedro J Moreno, "Environment normalization for robust speech recognition using direct cepstral comparison," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, 1994, vol. 2, pp. 61–64.

[32] 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.

[33] Catherine P. Browman and Louis Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[34] Prasanta Kumar Ghosh, Shrikanth S. Narayanan, Pierre L. Divenyi, Louis Goldstein, and Elliot Saltzman, "Estimation of articulatory gesture patterns from speech acoustics.," in *INTERSPEECH*, 2009, pp. 2803–2806.

[35] Karel Veselý, Arnab Ghoshal, Lukáš Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," in *INTERSPEECH*, 2013, pp. 2345–2349.

[36] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý, "The Kaldi Speech Recognition Toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[37] George A. Miller and Patricia E. Nicely, "An analysis of perceptual confusions among some English consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[38] Alfred Charles Gimson and Susan Ramsaran, *An Introduction to the Pronunciation of English*, vol. 4, Edward Arnold London, 1970.