Interspeech 2018
2-6 September 2018, Hyderabad

Noise Robust Acoustic to Articulatory Speech Inversion

Nadee Seneviratne[1], *Ganesh Sivaraman*[1], *Vikramjit Mitra*[1], *Carol Espy-Wilson*[1]

[1]University of Maryland College Park

nadee@terpmail.umd.edu, ganesa90@gmail.com, vikramjitmitra@gmail.com, espy@isr.umd.edu

Abstract

In previous work, we have shown that using articulatory features derived from a speech inversion system trained using synthetic data can significantly improve the robustness of an automatic speech recognition (ASR) system. This paper presents results from the first of two steps needed for exploring if the same will hold true for a speech inversion system trained with natural speech. Specifically, we developed a noise robust multi-speaker acoustic to articulatory speech inversion system. A feed forward neural network was trained using contextualized mel-frequency cepstral coefficients (MFCC) as the input acoustic features and six tract-variable (TV) trajectories as the output articulatory features. Experiments were performed on the U. Wisc. X-ray Microbeam (XRMB) database with 8 noise types artificially added at 5 different SNRs. Performance of the system was measured by computing the correlation between estimated and actual TVs. The performance of the multi-condition trained system was compared to the clean-speech trained system. The effect of speech enhancement on TV estimation was also evaluated. Experiments showed a 10% relative improvement in correlation over the baseline clean-speech trained system.

Index Terms: noise robust speech inversion, vocal tract variables, deep neural networks, articulatory features

1. Introduction

Speech inversion refers to the inverse problem of retrieving the articulatory dynamics responsible for the produced speech signal. It is a topic of interest that has applications in speech therapy, pronunciation training, robust Automatic Speech Recognition (ASR), and speech synthesis. The non-linearity and the non-uniqueness of the inverse mapping make speech-inversion a challenging task [1]. In building a good speech inversion system, it is imperative that we use the most suitable representations for acoustic and articulatory features as well as the algorithm appropriate for the complex non-linear mapping.

Previous studies have explored which acoustic features best represents the signal for the speech-inversion task. It was shown that Mel-Frequency Cepstral Coefficients (MFCCs) perform better than Perceptual Linear Prediction (PLP) and mel-spectrum (MELSPECT) features for the TV estimation process [2, 3].

Most research in speech inversion has been focused on developing accurate speaker-dependent systems. Approaches such as codebook search [4], feedforward neural networks, and Mixture Density Networks [5] have been found to work well for speaker-dependent speech inversion. There have been several attempts to perform speaker independent speech inversion systems [6, 7, 8, 9], but most of such studies have been limited to two speakers from the MOCHA-TIMIT dataset [10].

Articulatory features have been shown to provide noise robustness to ASR systems [11, 12, 13, 14, 15, 16]. Mitra et.al. (2017) [16] trained Deep Neural Network (DNN) and Convolutional Neural Network (CNN) based speech inversion systems on a clean and noise added multi-speaker simulated synthetic speech dataset generated using the Task Dynamics and Applications (TADA) system [17]. They showed that the articulatory features estimated by the multi-condition-trained speech inversion system combined with Gammatone filterbank features through a hybrid convolutional neural network (HCNN) architecture reduced the word error rates for each of the clean, noisy, and channel-mismatched conditions, providing state-of-the-art results on the Aurora 4 database. The same architecture provided significant improvement in performance for the SWB-1 speech recognition task when articulatory features were used with the Damped Oscillator Cepstral Coefficients (DOCC) features, compared to using the DOCC features alone [18].

Motivated by previous work on synthetic speech trained inversion system, in this paper, we focus on training a speaker independent noise robust speech inversion system on natural speech articulatory data. The speech data collected during real-time MRI [19] contains a significant amount of noise due to the MRI machine and speech enhancement methods have been proposed to reduce the noise in the recorded speech of rt-MRI articulatory datasets [20]. The noise robust speech inversion proposed in this paper can be applied to train speech inversion systems on the re-MRI datasets without requiring a speech enhancement preprocessor. To the best of our knowledge, the work presented in this paper is the first attempt at training a noise robust speaker independent speech inversion system using real speech data. We used the multi-speaker XRMB dataset for our experiments. We generated a noisy version of the XRMB dataset by electronically adding 8 different noise types at 5 SNRs. Further details about the dataset generation is explained in section 2. We used a feedforward neural network architecture to learn the mapping from acoustics to articulatory trajectories. Section 3 describes the architecture of the speech inversion system. Experiments were performed to evaluate the performance of the multi-condition trained speech inversion system under different noise types and SNRs. A speech inversion system which was built previously [2] by training only on clean speech was also used in these experiments for comparison purposes. We also evaluated the efficacy of a log-Minimum Mean Squared Error (log-MMSE) based speech enhancement system as a preprocessor for estimating the TVs from noisy speech. The experimental details and the results are discussed in section 4. The conclusions are given in section 5.

2. Dataset Description

In order to make the trained neural network model robust to noise, it is imperative that the model is trained on an input dataset which represents noise in different situations as well as in different intensity levels. As such a noisy speech database was not readily available, we used the Wisconsin X-Ray Microbeam database with noise being added artificially.

10.21437/Interspeech.2018-1509

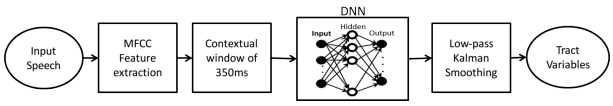

Figure 1: *Block diagram of the speech inversion system*

The XRMB recordings originally comprise of naturally spoken utterances from 32 male and 25 female subjects along with X-ray microbeam cinematography of the mid-sagittal plane of the vocal tract with pellets placed at points along the vocal tract. The trajectory data are recorded for the individual articulators: Upper Lip, Lower Lip, Tongue Tip, Tongue Blade, Tongue Dorsum, Tongue Root, Lower Front Tooth (Mandible Incisor), Lower Back Tooth (Mandible Molar). We call these trajectories as pellet trajectories. A common problem with articulatory recordings of this type is the mis-tracking of pellets or the pellets falling off while recording. Such problems were encountered in the XRMB recordings and were marked as miss-tracked segments. These segments were removed from the database before using it for our analysis.

The X-Y positions of the pellets are closely tied to the anatomy of the speakers. The quantification of the vocal tract shape is better performed by the location and degree of these constrictions than the X-Y positions of the pellets. Moreover, the absolute positions of the articulators are dependent on the anatomy of the speaker's vocal tract. The TVs specify the salient features of the vocal tract area function more directly than the pellet trajectories [21] and are relatively speaker independent. They also provide us a theoretical framework to analyze speech production with the theoretical framework of articulatory phonology. Hence, the pellet trajectories were converted to TV trajectories using geometric transformations as outlined in [22]. The transformed XRMB database consists of 21 males and 25 females, with a total of 4 hours of speech data with corresponding 6 TV trajectories. The TVs obtained from the seven pellet trajectories were Lip Aperture (LA), Lip Protrusion (LP), Tongue Body Constriction Location (TBCL), Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Location (TTCL) and, Tongue Tip Constriction Degree (TTCD).

The noise signals that were artificially added to the clean speech data of the XRMB database, were obtained from the AURORA database recordings representing different places: babble, car, exhibition hall, restaurant, street, airport, train station, and train. These noise signals were added to the clean speech signals at Signal-to-Noise Ratio (SNR) levels of 0dB, 5dB, 10dB, 15dB, and 20dB. We call this noise added version of the XRMB dataset as XRMB_noisy dataset.

The resultant noise embedded speech database was then used to train the noise-robust speech inversion system which is described in the next section.

3. Speech Inversion System

The development of the Deep Neural Network to be trained on the acoustic features to estimate the articulatory features is described here. The block diagram of the speech inversion system is shown in Figure 1. The different steps involved in the system are elaborated in the following subsections.

3.1. Feature extraction

Mel Frequency Cepstral Coefficients (MFCCs) are used as acoustic features for the input to the speech inversion system. When computing MFCCs, 13 cepstral coefficients were extracted from a 20ms Hamming analysis windows with a 10ms frame shift. The MFCCs and TVs were mean and variance normalized to have zero mean and unit variance per utterance. The mean and variance normalization were performed per utterance instead of using global statistics because different noise types might affect the cepstral coefficients in different ways resulting in average statistics that don't normalize the differences across noise types. The MFCCs were contextualized by concatenating every other feature frame in a 340ms window. This resulted in 8 frames of MFCCs on either side of each frame being concatenated to form the contextualized MFCC features (total of 17 frames including the current frame). By skipping two frames when splicing the frames, the current analysis frame is centered when concatenating every other frame. Previous studies [23] have explored different lengths of feature contextualization and found a splice width of 17 frames to be optimal.

3.2. DNN training

The input layer (which accepts contextualized MFCCs) of the feed-forward DNN has 221 nodes (13 MFCCs x 17 frames) and the output layer (which generates estimated TVs) has a dimensionality of 6 nodes. The input dataset was divided into training, development, and testing sets so that the training set has utterances from 36 speakers and the development and testing sets have 5 speakers each (3 males,2 females). Around 80% of the total number of utterances were present in the training set. The development and testing sets have a nearly equal number of utterances. This allocation was done in a completely random manner.

Similar to prior experiments [2], the trained DNNs had 1024 neurons in all the hidden layers. DNNs with 3, 4 and 5 hidden layers were trained and the model performing better on the cross-validation set was chosen as the optimal model.

The DNN was trained on mini batches as the total amount of input MFCC features were quite high to be trained as a whole batch. The training was performed to minimize the mean squared error between the actual TVs and the estimated TVs. The Adam optimizer was used for optimizing the network parameters. Speaker specific normalization was performed on the TVs when they are being generated.

3.3. Kalman smoothing

The estimated TVs from the DNN are often noisy. Articulatory trajectories (such as TVs) are low-pass in nature due to the kinematic constraints of human speech production system. The noise in the TV estimates were reduced by low-pass filtering the estimated TVs using a Kalman filter. In our work, we did not estimate the Kalman smoothing parameters, instead we fixed the parameters of the Kalman smoother to operate as a low-pass filter. The output of this is taken as the estimated TV output.

3.4. Performance measurement

Once the best performing DNN is obtained, the test data set was used to measure the performance of the model. The Pearson Product Moment Correlation (PPMC) between the estimated TV and the corresponding ground-truth TV was computed as

the performance evaluation metric.

$$PPMC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4. Experiments and Results

First, to identify the best performing DNN architecture, we trained the model having 3, 4, and 5 hidden layers keeping 1024 neurons in each hidden layer. The average correlations between actual and estimated TVs across all TVs on the cross validation set were 0.702, 0.707, and 0.710 respectively for 3, 4, and 5 layers. Since the 5 layer model generated the highest correlation, the rest of the experiments were conducted using the 5-layer DNN model. Adding further layers to the network led to non-convergence of the training. Hence we stopped at 5 hidden layers.

Next, several experiments were conducted to evaluate the performance of the multi-condition trained speech inversion system and also to compare the improvement of the results, keeping the speech inversion system trained only on the clean speech data as the baseline.

4.1. Performance of multi-condition trained speech inversion system across noise conditions

The multi-condition trained speech inversion system was evaluated on the test set of the XRMB_noisy dataset. The test was performed on all noise types and SNRs on the test set. Correlations between actual and estimated TVs were computed. Table 1 shows the correlations for each TV averaged over all SNR levels and noise types. The average correlation on the XRMB_noisy test set was 0.70949. Note that none of the speakers from the test set were included in the training set.

Figure 2 depicts the increase of average correlation as the SNR level is increases. Average Correlation for the TVs estimated on clean version of the test set using the noise robust speech inversion system is also included in the graph which is 0.741. These correlation values are averaged over all 6 TVs.

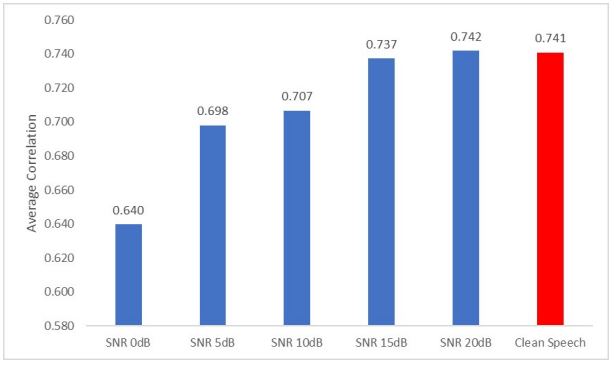

Figure 2: *Average correlations on the XRMB_noisy test set for different SNR levels.*

Figure 3 indicates the variation of the average correlation depending on the noise type. It can be observed that when the noise is more non-stationary, (Eg: Subway, Exhibition, and, Restaurant), the correlations are lower compared to the relatively more stationary noise types like car, train and airport. As expected, the correlation on the clean speech subset (red bar in Figure 3) outperforms the noisy test subsets.

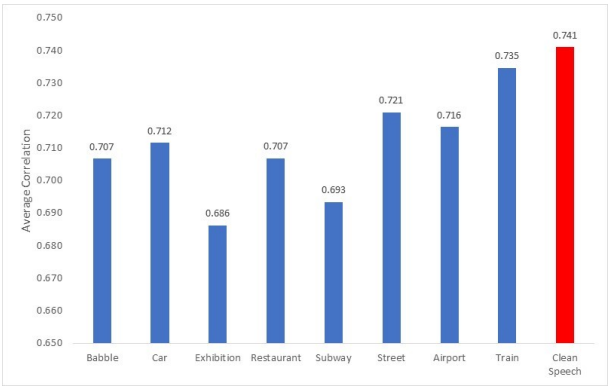

Figure 3: *Average correlations on the XRMB_noisy test set for different noise types*

4.2. Performance of clean trained inversion system on noisy test set

The clean speech trained inversion system, was evaluated on the XRMB_noisy test set. The results are shown in Table 1. The average correlation is 0.640. Thus, the multi-condition trained system in section 4.1 shows a relative improvement of 10.83% in correlation for the TV estimation of noisy speech using this noise robust inversion system.

4.3. Effect of speech enhancement

We used an implementation of the minimum mean-square error log-spectral amplitude estimator in [24] to enhance the noisy speech signals before extracting the MFCCs (by lowering the residual noise level). The effect of this pre-processing technique is investigated on the clean speech trained inversion system, and the results are compared with those of sections 4.1 and 4.2. The computed correlations are given in Table 1 and the average correlation in this case was only 0.597. This could be due to the distortions in the speech introduced by the speech enhancement algorithm.

Table 1: *Average correlations of TVs estimated by multi-condition trained speech inversion (SI) system and clean trained SI system on different test conditions*

Model	Test Set	LA	LP	TBCL	TBCD	TTCL	TTCD	Average
Multi-condition SI	Noisy	0.793	0.530	0.818	0.626	0.635	0.855	0.710
Clean Speech SI	Noisy	0.710	0.476	0.757	0.546	0.571	0.782	0.640
Clean Speech SI	Enhanced	0.668	0.430	0.725	0.497	0.525	0.740	0.597
Multi-condition SI	Clean	0.828	0.554	0.845	0.680	0.654	0.885	0.741
Clean Speech SI	Clean	0.856	0.613	0.866	0.745	0.707	0.907	0.782

4.4. Performance on clean speech data

Finally, we compared the correlations of the estimated TVs of clean speech data obtained from the multi-condition trained inversion system along with those generated by the speech inversion system trained only on clean speech. The last two rows of Table 1 summarizes the results in this section. The TVs estimated from the clean trained system had an average correlation of 0.782 while those estimated by the multi-condition trained system had an average correlation of 0.741. Thus the multi-condition training results in a relative degradation in the performance on clean speech by 5.2%

Figures 4 and 5 show the estimated Tract Variables, LA, TBCD, and TTCD for an example utterance estimated by the

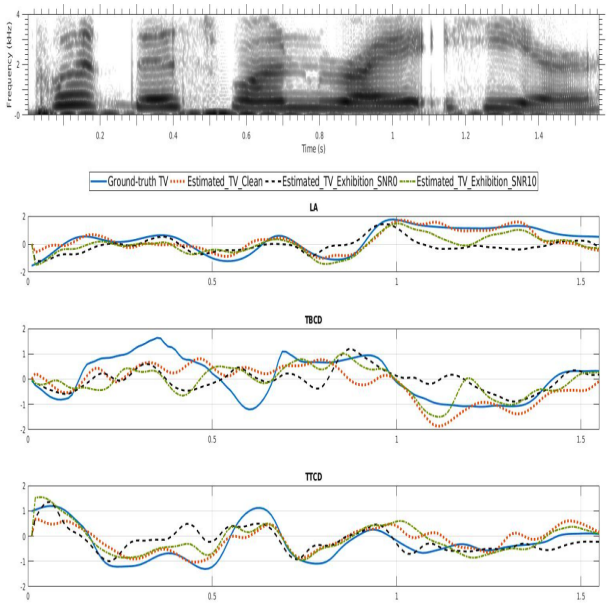

Figure 4: *TV plots for the utterance "Put this one right here" for exhibition noise - estimated using multi-condition trained speech inversion system. Solid blue Line - actual TV, red dotted line - estimated TV from clean signal, black dashed Line - estimated TV from noisy signal (0dB), green dash-dot line - estimated TV from noisy signal (10dB)*

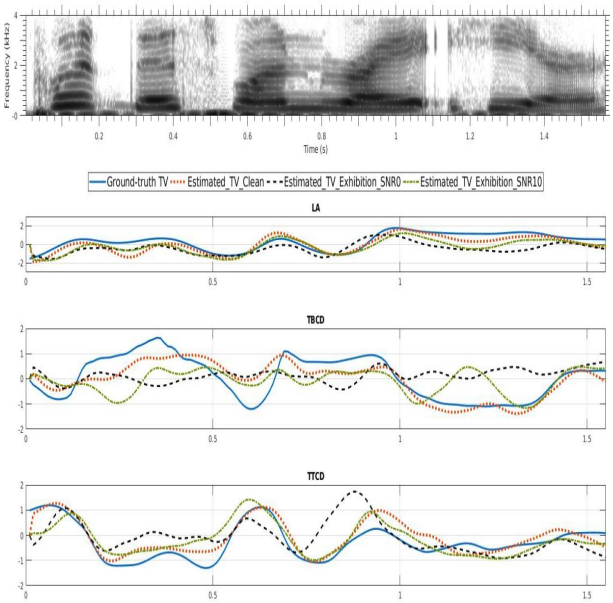

Figure 5: *TV plots for the utterance "Put this one right here" for exhibition noise - estimated using clean speech trained speech inversion system. Solid blue line - actual TV, red dotted line - estimated TV from clean signal, black dashed line - estimated TV from noisy signal (0dB), green dash-dot line - estimated TV from noisy signal (10dB)*

multi-condition trained speech inversion system and the clean speech trained system respectively. We chose to display these TVs for the clean speech signal and for signals with noise added at 0dB SNR and 10 dB SNR, along with the actual TV for the noise type Exhibition. It can be seen that the estimated TVs for LA and TTCD align with the corresponding ground-truth TVs well, compared to TVs for TBCD. As the chosen utterance doesnt contain velar consonants where TBCD would be a critical constriction, there is more variability in TBCD (i.e., different people will choose to different things with the tongue body given it is not needed to produce any of the consonants). When the TTCD TV estimations for the two cases are considered, it can be seen that the multi-condition trained systems TV at 0 dB aligns more with the actual TV compared to the same TV of the clean speech trained system. This is clearly seen in the time region from 0.7 to 0.9 where there is a tongue tip constriction for the /n/ and the following /r/ in "one right".

5. Conclusion

This paper presented the development of a noise robust acoustic to articulatory speech inversion system using a multi-condition-trained DNN. Experiments were performed on the noise added XRMB dataset. A DNN based speech inversion system was trained with contextualized MFCCs as the input and 6 TVs as the output. Results show that the correlation of TVs estimated by the multi-condition-trained system improves by 10.83% compared to the TVs estimated by the clean speech trained system. We performed speech enhancement preprocessing on the noisy speech data to determine whether the speech enhancement nullifies the gains obtained by multi-condition training. Testing the clean trained system on speech enhanced XRMB noisy test set revealed that the speech enhancement did not provide any improvement in the correlation compared to not performing any speech enhancement. This was a perplexing result as it showed that the clean speech trained system works better on noisy speech compared to enhanced speech. This result is probably due to speech distortions introduced by the speech enhancement. In the future we plan to see if less aggressive speech enhancement yields better results, perform a cross corpus test of the noise robust speech inversion system and apply the noise robust speech inversion system for robust ASR.

6. Acknowledgements

This work was made possible by a hardware grant from NVIDIA.

7. References

[1] C. Qin and M. Á. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping." *INTERSPEECH*, 2007.

[2] G. Sivaraman, "Articulatory representations to address acoustic variability in speech," Ph.D. dissertation, University of Maryland College Park, 2017.

[3] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Retrieving tract variables from acoustics: A comparison of different machine learning strategies," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1027–1045, 2010.

[4] P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion." *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–72, 2010.

[5] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proceedings of the*

Annual Conference of the International Speech Communication Association, INTERSPEECH. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 2, pp. 577–580.

[6] A. Afshan and P. K. Ghosh, "Improved subject-independent acoustic-to-articulatory inversion," *Speech Communication*, vol. 66, pp. 1–16, 2015.

[7] L. Girin, T. Hueber, and X. Alameda-Pineda, "Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 3, pp. 662–673, 2017.

[8] A. Ji, "Speaker Independent Acoustic-To-Articulatory," Ph.D. dissertation, Marquette University, 2014.

[9] A. Ji, M. T. Johnson, and J. J. Berry, "Parallel Reference Speaker Weighting for Kinematic-Independent Acoustic-to-Articulatory Inversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1865–1875, 2016.

[10] A. A. Wrench, "A Multichannel Articulatory Database and its Application for Automatic Speech Recognition," in *In Proceedings 5 th Seminar of Speech Production*, 2000, pp. 305–308.

[11] K. Kirchhoff, "Robust speech recognition using articulatory information," 1999.

[12] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 303–319, 2002.

[13] G. Sivaraman, V. Mitra, and C. Espy-Wilson, "Fusion of acoustic, perceptual and production features for robust speech recognition in highly non-stationary noise," in *CHiME-2013*, Vancouver, 2013, pp. 65–70.

[14] V. Mitra, G. Sivaraman, H. Nam, C. Y. Espy-Wilson, and E. Saltzman, "Articulatory features from deep neural networks and their role in speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*. IEEE, may 2014, pp. 3017–3021.

[15] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "Integrating articulatory data in deep neural network-based acoustic modeling," *Computer Speech and Language*, vol. 36, pp. 173–195, 2016.

[16] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, and M. Tiede, "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Communication*, vol. 89, pp. 103–112, 2017.

[17] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, p. 2430, 2004.

[18] V. Mitra, W. Wang, C. Bartels, H. Franco, and D. Vergyri, "Articulatory Information And Multiview Features For Large Vocabulary Continuous Speech Recognition," in *ICASSP*, 2018.

[19] S. Narayanan, E. Bresch, P. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time MRI articulatory corpus for speech research," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. International Speech Communication Association, 2011, pp. 837–840.

[20] C. Vaz, V. Ramanarayanan, and S. Narayanan, "A two-step technique for MRI audio enhancement using dictionary learning and wavelet packet analysis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.

[21] R. S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," *Speech Communication*, vol. 14, no. 1, pp. 19–48, 1994.

[22] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "Recognizing articulatory gestures from speech for robust speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2270–2287, 2012.

[23] V. Mitra, "Articulatory Information For Robust Speech Recognition," 2010.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.