INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multi-corpus Acoustic-to-articulatory Speech Inversion

Nadee Seneviratne [1], *Ganesh Sivaraman*[2], *Carol Espy-Wilson*[1]

[1]University of Maryland College Park
[2]Pindrop

nadee@umd.edu, ganesa90@gmail.com, espy@umd.edu

Abstract

There are several technologies like Electromagnetic articulometry (EMA), ultrasound, real-time Magnetic Resonance Imaging (MRI), and X-ray microbeam that are used to measure speech articulatory movements. Each of these techniques provides a different view of the vocal tract. The measurements performed using the similar techniques also differ greatly due to differences in the placement of sensors, and the anatomy of speakers. This limits most articulatory studies to single datasets. However to yield better results in its applications, the speech inversion systems should be more generalized, which requires the combination of data from multiple sources. This paper proposes a multi-task learning based deep neural network architecture for acoustic-to-articulatory speech inversion trained using three different articulatory datasets - two of them were measured using EMA, and one using X-ray microbeam. Experiments show improved accuracy of the proposed acoustic-to-articulatory mapping compared to the systems trained using single datasets.
Index Terms:Acoustic-to-articulatory speech inversion, multi-task learning, articulatory phonology, tract variables

1. Introduction

Speech inversion (SI), the process of retrieving the articulatory dynamics composing a speech signal, continues to be a problem of interest in the speech community due to it's contribution in a wide range of applications including robust automatic speech recognition, speech synthesis, pronunciation training, and speech therapy. The non-linearity and non-uniqueness of the acoustic to articulatory mapping [1], makes the problem of speech inversion quite challenging. A lot of solutions have been developed to improve the performance of speech inversion systems.

Initial approaches for speech inversion mostly focused on developing speaker-dependent systems. Such approaches include codebook search [2, 3], feed-forward neural networks [4, 5], and mixture density networks [6]. In order to use articulatory representations in practical applications, it is necessary to train speaker-invariant systems. In recent years there have been several attempts at training speaker independent speech inversion systems [7, 8, 9, 10, 11]. However most of these studies were limited to articulatory data collected from a single corpus. Most works in speech inversion are based on either the U.Wisconsin X-ray microbeam (XRMB) data [12] or the MOCHA TIMIT dataset [13]. The XRMB dataset which is the largest multi-speaker articulatory dataset, contains 56 speakers. The MOCHA TIMIT dataset contains just 2 speakers. Most articulatory datasets contain a limited number of speakers due to the expensive and time consuming process of data collection. There are several such small articulatory datasets containing less than 10 speakers.

Technologies such as Electromagnetic articulometry (EMA), ultrasound, real-time MRI, and X-ray Micro-beam etc. have facilitated the process of acquiring measurements related to movements of speech articulators such as tongue body, tongue tip, upper and lower lips, and jaw. These technologies provide different views of the vocal tract. One limitation is that, depending on the anatomy of the speakers and placement of the sensors or pellets, even similar techniques can provide varying measurements of the same articulatory trajectories. Due to this limitation, most of the studies of speech inversion have been limited to single datasets.

However, machine learning systems, especially deep neural networks (DNNs) can generalize only when they are trained with a large amount of data. Thus, for training robust speaker independent speech inversion systems, it is essential to combine data from different sources collected with different instruments and methods. In this paper, we propose to generalize the speech-inversion process by using a multi-task learning model to develop a multi-corpus SI system.

This paper is organized as follows. Section 2 provides an overview of the three speech corpora used to train the SI system. Section 3 explains the architecture and training procedure of the multi-corpus SI model using a DNN. The different experiments used to evaluate the performance of the multi-corpus SI system against single-corpus SI systems and the corresponding results are discussed in section 4. Section 5 summarizes the quantitative results and improvement of the performance of the proposed solution.

2. Datasets Description

For this work we used three speech databases to develop the multi-corpus speech inversion system.

2.1. X-Ray Microbeam (XRMB) dataset

The original University of Wisconsin XRMB database [12] comprises of naturally spoken isolated sentences and short read paragraphs collected from 32 male and 25 female subjects. These speech utterances were recorded along with trajectory data captured by X-ray microbeam cinematography of the mig-sagittal plane of the vocal tract using pellets placed on several articulators: upper (UL) and lower (LL) lip, tongue tip (T1), tongue blade (T2), tongue dorsum (T3), tongue root (T4), mandible incisor (MANi), and (parasagittally placed) mandible molar (MANm). However some of the articulatory recordings were marked as mistracked samples in the database due to pellets either falling off or being mistracked. We eliminated these samples from the database prior to further processing.

The absolute position of the articulators depend on the anatomy of the speaker's vocal tract. Given that the X-Y positions of the pellets strongly depend on the anatomy of the speakers and variability of pellet placements, the measurements can vary significantly across speakers. Hence, to better repre-

http://dx.doi.org/10.21437/Interspeech.2019-3168

sent vocal tract shape, relative measures were used to calculate the Tract Variables (TVs) from the X-Y positions of the pellets. TVs lead to a relatively speaker independent representation of speech articulation and characterize salient features of the vocal tract area function [14]. The TVs also provides a theoretical framework for speech production analysis and articulatory phonology [15]. Thus using geometric transformations defined in the Task Dynamic model of speech production, the XRMB trajectories were converted to TV trajectories as outlined in [16]. The transformed XRMB database is comprised of 4 hours of speech data from 21 males and 25 females with corresponding six TV trajectories: 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).

2.2. EMA-IEEE dataset

The EMA-IEEE dataset is composed of recordings from 4 female and 4 male subjects reciting 720 phonetically balanced IEEE sentences [17] at normal and fast production rates [18]. The recordings were done using a 5-D electromagnetic articulometry (EMA) system (WAVE; Northern Digital). Each sentence was first produced at speaker's preferred 'normal' speaking rate and then producing a 'fast' repetition of the same, without making errors. Sensors were placed on the tongue (tip (TT), body (TB), root (TR)), lips (upper (UL) and lower (LL)) and mandible, together with reference sensors on the left and right mastoids, and upper and lower incisors (UI, LI). These EMA trajectories were obtained at 100 Hz and then were low-pass filtered at 5 Hz for references and 20 Hz for articulator sensors. Synchronized audio was recorded at 22050 Hz.

The following geometric transformations were used to obtain 9 TVs (namely LA, LP, Jaw Angle (JA), TTCL, TTCD, Tongue Middle Constriction Location (TMCL), Tongue Middle Constriction Degree (TMCD), TBCL and TBCD).

$$LA[n] = \sqrt{(LL_x[n] - UL_x[n])^2 + (LL_z[n] - UL_z[n])^2} \quad (1)$$

$$LP[n] = LL_x[n] - \underset{m \in allutterances}{median} LL_x[m] \quad (2)$$

$$JA[n] = \sqrt{(LI_x[n] - UL_x[n])^2 + (LI_z[n] - UL_z[n])^2} \quad (3)$$

$$TTCD[n] = \underset{m \in (-50,0)}{Min} Dist(TT, pal(x)) \quad (4)$$

$$TTCL[n] = \underset{m \in allutterances}{median} TT_x[m] - TT_x[n] \quad (5)$$

LA (eq. 1) is defined as the Euclidean distance between the UL and the LL sensors. LP (eq. 2) is the displacement along the x-axis of the LL sensor from its median position. JA (eq. 3) was computed as the Euclidean distance between the UL sensor and the LI sensor. For each tongue sensor, two TVs were computed to obtain the degree and location of constriction. Constriction degree is defined as the minimum distance between the sensor and the palate trace as shown in equation 4 which was used to compute TTCD, TMCD, and TBCD using TT, TM, and TB sensor positions and the pellet trace. The location of constriction

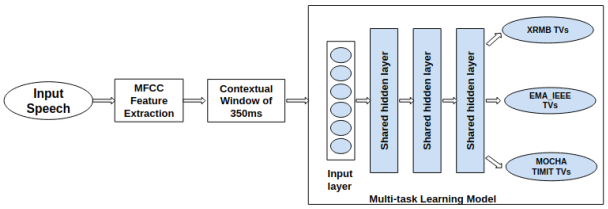

Figure 1: *Block diagram of the multi-corpus speech inversion system*

(eq. 5) for a tongue sensor was defined as the displacement of the sensor along the x-direction from its median position. This way, TTCL, TMCL, and TBCL were computed from the TT, TM, and TB sensor positions.

2.3. MOCHA-TIMIT dataset

The Multichannel Articulatory (MOCHA) database [19] consists of speech data and EMA data recorded simultaneously for one male and one female subjects speaking British English. The EMA data was downsampled to 100 Hz from 500 Hz as described in [20]. Using the same approach as in section 2.2, the EMA sensor position data was converted to 9 TVs.

3. Speech Inversion (SI) System

3.1. Feature extraction

Mel-Frequency Cepstral Coefficients (MFCCs) were chosen as the acoustic input features for the speech inversion systems. Using a 20ms Hamming analysis window with a 10ms frame shift, 13 cepstral coefficients were extracted for each frame. For training, the MFCCs and TVs were normalized to have zero mean and unit variance per speaker. This normalization was done per utterance for testing. To construct the feature vector, MFCCs were contextualized by concatenating every other feature frame in a 340ms window leading to having 8 frames in either side of the current frame (17 frames in total). To keep the current analysis frame centered when concatenated, two frames were skipped when splicing the frames. 17 was found to be the optimal splice width by previous studies which explored different lengths of feature contextualization [21].

3.2. Model architecture and training

A feed forward deep neural network (FF-DNN) architecture was used to build the multi-corpus SI system. Using a multi-task learning approach, the model was trained to learn three different sets of TVs corresponding to speech samples in the three databases XRMB, EMA-IEEE, and MOCHA-TIMIT (three tasks). The hidden layers of the model are shared by these three output tasks. A schematic of the model architecture is given in figure 1.

MFCCs described in section 3.1 were used as the acoustic feature input to the model. To train the FF-DNN, contextualized MFCC features were used. The input layer dimensionality was 221 nodes (13 coefficients x 17 frames). The three tasks of estimating TVs for XRMB, EMA-IEEE, and MOCHA-TIMIT speech utterances had 6, 9, and 9 output nodes respectively. Each dataset was divided into training, development, and testing sets prior to experiments without any overlap of speakers. For the XRMB dataset, utterances from 36 speakers were randomly allocated for training (80% of the data) and the development

and testing sets have 5 speakers each (3 males, 2 females). For the EMA-IEEE dataset, training subset contained 576 sentences while the test and cross-validation sets contained 72 sentences each. MOCHA TIMIT dataset was distributed with 735, 91, and 92 utterances for training, validation and testing subsets respectively.

The objective of the training was to minimize the mean squared error between the ground-truth TVs and the estimated TVs. Network parameters were optimized using Adam optimizer (Keras default) along with batch normalization and dropouts (0.1 for the input layer and 0.2 for subsequent layers). The generated TVs were speaker-specific normalized.

3.3. Kalman smoothing

Kinematic constraints of human speech production system make the articulatory trajectories low-pass in nature. However the TVs directly estimated from the neural networks are often noisy. Thus, to remove the high frequency noise components in the TV estimations, a Kalman filter was used. In this work, we fixed the parameters of the Kalman smoother to operate as a low-pass filter (Kalman smoothing parameters were not estimated). The output of the lowpass filter is taken as the estimated TV output.

3.4. Performance measurement

The performance evaluation metric for the model was chosen to be the Pearson Product Moment Correlation (PPMC) between the estimated TV and the corresponding ground-truth TV.

4. Experiments and Results

4.1. Performance of single-corpus SI systems

The purpose of this experiment was to understand the performance of the SI system when being trained with a single corpus at a time, so that the set of results could be used as a baseline. To choose the best performing model, for each dataset, separate feed-forward neural networks were trained with 2, 3, 4, and 5 hidden layers with 512 and 1024 nodes in each layer. Beyond 5 hidden layers, the performance saturated and hence the DNN was limited to 5 hidden layers. Table 1 provides the average correlation values of the cross-validation set used during the training process.

Table 1: *Average correlations of TVs estimated by single-corpus SI systems for the best performing models*

Dataset	Model architecture	Validation set average corr.
XRMB	5 hidden layers, 512 nodes each	0.789
EMA-IEEE	5 hidden layers, 1024 nodes each	0.826
MOCHA-TIMIT	5 hidden layers, 1024 nodes each	0.730

4.2. Performance of multi-corpus SI system

In this experiment, the multi-task learning model was trained using all three databases. The joint model was trained for 100 epochs in total. During one global epoch, the three output tasks for XRMB, EMA-IEEE, and MOCHA-TIMIT datasets were learned one after the other for one epoch at a time. This way, the weights of shared hidden layers were tuned using speech samples from multiple datasets. The number of hidden layers was changed from 3 to 5 and the number of nodes in each layer was changed from 512 to 1024 to choose the optimal modal. The best performing model architecture was found to have 5 hidden layers with 1024 nodes in each hidden layer. The average correlations for the validation sets of the three databases were **0.768**, **0.819**, and **0.780** respectively.

4.3. Cross correlation evaluation for single-corpus SI systems

Once the best models for separate single-corpus SI systems were obtained, we performed cross correlation evaluation for test sets of three databases. For this, the models were evaluated on test data from its respective database which was used to train the model as well as on the test data from remaining databases. The test utterances used in this case were not present in either training or validation sets. As the TVs defined for the XRMB database were different to those of EMA-IEEE and MOCHA-TIMIT databases, some TVs were ignored when they are not common in all three datasets. Therefore when evaluating XRMB test samples on other two models, JA, TMCL, and TMCD TVs were ignored while keeping the correspondence between other TVs with similar names.

Table 2: *Cross correlations of TVs of test samples evaluated on best performing single-corpus models*

	Best Model - XRMB	Best Model - EMA-IEEE	Best Model - MOCHA-TIMIT
XRMB test set	0.779	0.543	0.460
EMA-IEEE test set	0.453	0.821	0.540
MOCHA-TIMIT test set	0.475	0.608	0.735

From Table 2, it can be seen that when evaluated on test data from an alternative database, the average correlation between the estimated TVs and actual TVs are lower compared to the evaluation on test data from the corresponding database.

4.4. Cross correlation evaluation for multi-corpus SI system

We performed a similar experiment as described in section 4.3 to evaluate cross correlation using the multi-task learning model optimized using all three corpora. According to Table 3, the cross correlation results improved significantly compared to the results in section 4.3 (Figure 2). The relative improvements (as a percentage) compared to the results in section 4.3 are included along with the absolute average correlation values. It can be seen that the correlations scores are reduced in the scenarios where a particular test set was evaluated on its corresponding output (eg: XRMB test set on XRMB output of the joint model), compared to the model which was solely trained only on the data from that dataset. This is expected as the multi-corpus SI system will be tuned to model articulatory information represented in multiple datasets, making it more generalized. It is possible that the model could lose information of some articulatory features which are too specific to a certain database. However this reduction is quantitatively negligible compared to the percentage improvement of cross correlation scores.

Table 3: *Cross correlations of TVs of test samples evaluated on multi-corpus model*

	XRMB output	EMA-IEEE output	MOCHA-TIMIT output
XRMB test set	0.761 (-2.3%)	0.581 (6.9%)	0.596 (29.5%)
EMA-IEEE test set	0.576 (27.1%)	0.812 (-1.1%)	0.724 (34.2%)
MOCHA-TIMIT test set	0.576 (21.3%)	0.692 (13.9%)	0.781 (6.3%)

Tables 4 and 5 include the correlations between individual estimated and ground-truth TVs which are common in all three

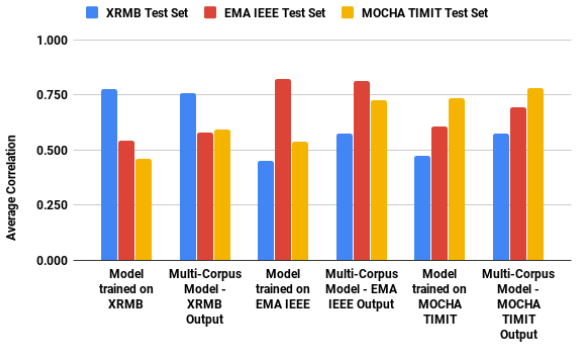

Figure 2: *Cross correlation results for different models*

datasets. By using the multi-corpus SI system, the correlations of TBCL and TBCD TVs could be improved significantly.

Table 4: *Cross correlations of individual TVs estimated for test samples evaluated on single-corpus models*

Model	Test Data Set	LA	LP	TTCL	TTCD	TBCL	TBCD
Single Corpus XRMB	XRMB	0.859	0.601	0.705	0.907	0.863	0.736
	EMA-IEEE	0.579	0.253	0.451	0.704	0.286	0.445
	MOCHA-TIMIT	0.632	0.174	0.499	0.702	0.324	0.519
Single Corpus EMA-IEEE	XRMB	0.771	0.347	0.413	0.809	0.445	0.474
	EMA-IEEE	0.828	0.763	0.841	0.864	0.758	0.806
	MOCHA-TIMIT	0.688	0.412	0.554	0.731	0.542	0.608
Single Corpus MOCHA TIMIT	XRMB	0.633	0.293	0.343	0.734	0.454	0.304
	EMA-IEEE	0.535	0.409	0.522	0.640	0.519	0.431
	MOCHA-TIMIT	0.790	0.619	0.715	0.803	0.691	0.718

Table 5: *Cross correlations of individual TVs estimated for test samples evaluated on multi-corpus model*

Model	Test Data Set	LA	LP	TTCL	TTCD	TBCL	TBCD
Multi Corpus XRMB	XRMB	0.846	0.570	0.688	0.891	0.851	0.718
	EMA-IEEE	0.781	0.375	0.537	0.816	0.335	0.610
	MOCHA-TIMIT	0.779	0.313	0.504	0.809	0.374	0.676
Multi Corpus EMA-IEEE	XRMB	0.808	0.378	0.456	0.848	0.483	0.512
	EMA-IEEE	0.816	0.755	0.832	0.856	0.752	0.793
	MOCHA-TIMIT	0.782	0.482	0.672	0.807	0.629	0.710
Multi Corpus MOCHA TIMIT	XRMB	0.796	0.381	0.526	0.858	0.466	0.549
	EMA-IEEE	0.751	0.595	0.749	0.809	0.654	0.690
	MOCHA-TIMIT	0.836	0.667	0.769	0.839	0.739	0.785

Figures 3 and 4 show the estimated Tract Variables, LA, TBCD, and TTCD for an example utterance estimated by the single-corpus SI systems and multi-corpus SI system respectively. In both figures, 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. With the multi-corpus model, the TBCD TVs have become more aligned. Also, the estimated TTCD TVs from the multi-corpus model seems to correlate well with the ground truth TV as well as with each other compared to those in Figure 3.

5. Conclusion

This paper presented the development of a multi-corpus SI system using different sets of articulatory trajectories provided by the three databases, XRMB, EMA-IEEE, and MOCHA-TIMIT. A multi-task learning based DNN was trained using the contextualized MFCC features as the input acoustic features and corresponding TV information as the model outputs. As given in Table 3, for XRMB test set, even though the joint model slightly reduced the correlation for the XRMB output (2.3%), when tested on EMA-IEEE and MOCHA-TIMIT outputs, the

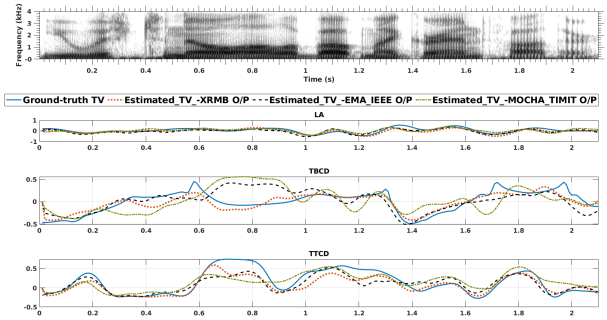

Figure 3: *TV plots for the utterance "You wished to know all about my grandfather" estimated using single-corpus models. Solid blue Line - actual TV (from XRMB database), red dotted line - estimated TV from model trained on XRMB data, black dashed Line - estimated TV from model trained on EMA-IEEE data, green dash-dot line - estimated TV from model trained on MOCHA-TIMIT data*

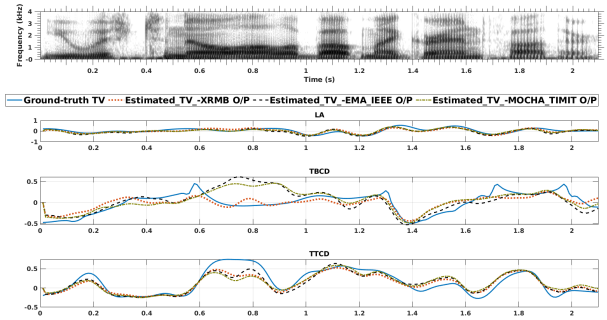

Figure 4: *TV plots for the utterance "You wished to know all about my grandfather" estimated using multi-corpus model. Solid blue Line - actual TV (from XRMB database), red dotted line - estimated TV from XRMB output, black dashed Line - estimated TV from EMA-IEEE output, green dash-dot line - estimated TV from MOCHA-TIMIT output*

correlations increased by 6.9% and 29.5% respectively compared to single-corpus SI systems. When EMA-IEEE data was evaluated on its corresponding output, the correlation was decreased by 1.1%, but the correlations increased by 27.1% and 34.2% when evaluated on XRMB and MOCHA-TIMIT outputs respectively. Test correlations for the MOCHA-TIMIT data improved for all three outputs in the joint model: 21.3%, 13.9%, and 6.3% for XRMB, EMA-IEEE, and MOCHA-TIMIT respectively. Thus, it can be believed that the proposed multi-corpus SI system perform better in generalizing articulatory dynamics of speech samples in multiple databases. In future, we plan to investigate different neural network architectures with different acoustic feature representations which could improve the performance of the system.

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] 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.

[3] S. Ouni and Y. Laprie, "Design Of Hypercube Codebooks For The Acoustic-To-Articulatory Inversion Respecting The Non-Linearities Of The Articulatory-To-Acoustic Mapping," in *Eurospeech*, dec 2002.

[4] M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi, "On the use of neural networks in articulatory speech synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, p. 1109, feb 1993. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/93/2/10.1121/1.405559

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

[6] 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. [Online]. Available: http://link.springer.com/10.1007/978-3-540-77347-4_23

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

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

[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] 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. [Online]. Available: http://ieeexplore.ieee.org/document/7814297/

[11] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, sep 2016, pp. 455–459. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/1399.html

[12] J. R. Westbury, "Speech Production Database User ' S Handbook," *IEEE Personal Communications* -, vol. 0, no. June, 1994.

[13] 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.

[14] 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.

[15] C. P. Browman and L. Goldstein, "Articulatory Phonology : An Overview *," *Phonetica*, vol. 49, pp. 155–180, 1992.

[16] 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. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.3682038

[17] E. Rothauser, W. Chapman, and N. Guttman, "IEEE Recommnded Pratice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, sep 1969.

[18] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017. [Online]. Available: https://doi.org/10.1121/1.4987629

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

[20] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 153–172, 2003.

[21] V. Mitra, "Articulatory Information For Robust Speech Recognition," Ph.D. dissertation, University of Maryland, College Park, 2010.