Multitask learning based multi-corpus acoustic-to-articulatory speech inversion

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

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

Articulatory kinematics can be measured using various methods like X-ray microbeam, electropalatography Electromagnetic articulometry (EMA), real-time MRI, and ultrasound. All these measurement techniques capture different views of the vocal tract at different degrees of spatial and temporal resolution. Most articulatory studies are limited to one domain of articulatory measurement due to the high variability across measurement domains.

Acoustic-to-articulatory speech inversion is the process estimating the articulatory movements from the acoustic speech signal. Speech inversion systems are data-driven machine learning models usually trained on single articulatory datasets obtained using one measurement technique. The speech inversion systems, like all machine learning systems can learn more generalized mapping (not specific to a speaker or a dataset) if they are trained on larger datasets with more variability. However, most articulatory datasets are small (around 2-5 hours of speech) and contain a small number of speakers (2-10 speakers except the XRMB dataset). Speech inversion systems trained on such small articulatory datasets do not generalize well on unseen datasets [1]. All the articulatory measurement techniques measure similar vocal tract apparatuses performing similar tasks of speech production. Hence, the correlation among these different measurement domains can be leveraged to train a single speech inversion system that generalizes better than those trained on single datasets. Previous studies have shown that discrete articulatory features (AF) derived from phonetic feature classifications are complementary to acoustic features for improving speech recognition accuracy [2, 3, 4]. Since disrete AFs are derived from phone-alignments, large speech datasets with transcriptions can be used as additional data for training the speech inversion system.

In this work we propose a multi-task learning based acoustic-to-articulatory speech inversion system. We combine data from 3 different articulatory datasets and one dataset not containing articulatory recordings for training the speech inversion system in a multi-target learning procedure. The articulatory datasets used in this study are - X-ray microbeam database (XRMB) [5], Haskins Production Rate Comparison (HPRC) database [6], and the MOCHA TIMIT database [7]. The non-articulatory dataset used is Librispeech 100 hour clean speech subset. The articulatory data were converted from raw sensor trajectories (flesh point trajectories for XRMB) to Tract Variables using the procedure described in [8]. The discrete AFs for all the datasets were derived using phone alignment and a lookup table. Each phoneme is uniquely described in terms binary features corresponding to voicing, place of articulation, manner of articulation, front-back, and rounding [2]. MFCC features were chosen as the acoustic features for the speech inversion system.

We experimented with feed-forward and recurrent neural network architectures for the speech inversion system. The model consists of a single neural network architecture with different output layers for different output types (Figure 1). Since the TVs extracted from the articulatory datasets were different (different number of TVs and types), each set of TVs was given a separate output layer. The fourth output layer was used for the discrete AFs. Each epoch of the training consisted of 4 sub-epochs, one for each dataset. When training the network on the articulatory datasets,

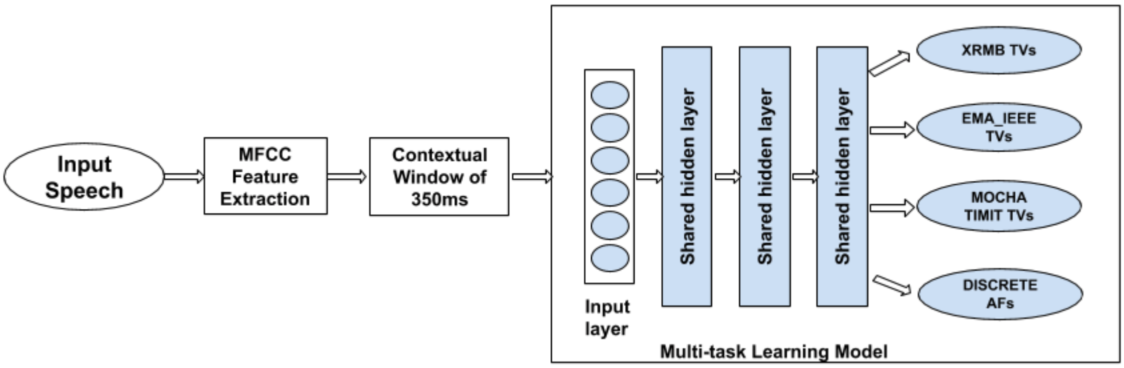

Figure 1: Block diagram of the proposed multi-corpus speech inversion system

errors were back-propagated from the discrete AF output layer as well as the TV output layer corresponding to the dataset. For the Librispeech corpus (non-articualtory data) only the errors from the discrete AF output were used for training. Results using the multi-task speech inversion architecture using only 3 articulatory datasets and no phonetic features showed improved accuracy of the estimated articulatory trajectories and improved cross-corpus performance [8]. This work extends the previous work of the authors by including discrete AFs and non-articulatory datasets for training the system. The proposed speech inversion architecture and training procedure has potential for learning a generalized speaker independent acoustic-to-articulatory mapping.

References:

[1] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, “Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion,” *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 316–329, jul 2019.

[2] K. Kirchhoff, “Robust speech recognition using articulatory information,” 1999.

[3] A. Juneja, “Speech recognition based on phonetic features and acoustic landmarks,” dec 2004.

[4] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, “Speech production knowledge in automatic speech recognition.” *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–42, feb 2007.

[5] J. R. Westbury, “Speech Production Database User ’ S Handbook,” *IEEE Personal Communications - IEEE Pers. Commun.*, vol. 0, no. June, 1994.

[6] 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.

[7] 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.

[8] N. Seneviratne, G. Sivaraman, and C. Espy-Wilson, “Multi-Corpus Acoustic-to-Articulatory Speech Inversion,” in *Interspeech 2019.* ISCA: ISCA, sep 2019, pp. 859–863.