Analysis of acoustic-to-articulatory speech inversion across different accents and languages

Ganesh Sivaraman[1], *Carol Espy-Wilson*[1], *Martijn Wieling*[2]

[1]University of Maryland College Park, USA
[2]University of Groningen, Netherlands
{ganesa90, espy}@umd.edu, m.b.wieling@rug.nl

Abstract

The focus of this paper is estimating articulatory movements of the tongue and lips from acoustic speech data. While there are several potential applications of such a method in speech therapy and pronunciation training, performance of such acoustic-to-articulatory inversion systems is not very high due to limited availability of simultaneous acoustic and articulatory data, substantial speaker variability, and variable methods of data collection. This paper therefore evaluates the impact of speaker, language and accent variability on the performance of an acoustic-to-articulatory speech inversion system. The articulatory dataset used in this study consists of 21 Dutch English speakers reading Dutch and English words and sentences, and 22 UK English speakers reading English words and sentences. We trained several acoustic-to-articulatory speech inversion systems both based on deep and shallow neural network architectures in order to estimate electromagnetic articulography (EMA) sensor positions, as well as vocal tract variables (TVs). Our results show that with appropriate feature and target normalization, a speaker-independent speech inversion system trained on data from one language is able to estimate sensor positions (or TVs) for the same language correlating at about $r = 0.53$ with the actual sensor positions (or TVs). Cross-language results show a reduced performance of $r = 0.47$.

Index Terms: Acoustic-to-articulatory speech inversion, Electromagnetic articulography, Tract variables, Cross-accent speech inversion, Cross domain speech inversion

1. Introduction

Speech inversion or acoustic-to-articulatory inversion of speech is the process of mapping the acoustic signal into articulatory parameters. Articulatory information can be used in speech accent conversion [1], speech therapy, language learning, and Automatic Speech Recognition (ASR) [2, 3, 4]. Actual articulatory data is obtained from subjects using techniques such as Electromagnetic Articulometry (EMA), X-ray microbeam, or real-time Magnetic Resonance Imaging (rt-MRI). However these techniques require sophisticated, expensive devices, and obtaining articulatory data is time consuming. Consequently, obtaining this type of data is frequently not practically feasible. Given that acoustic data can be easily obtained, there is a clear use for an accurate speech inversion system which is speaker independent and can accurately estimate articulatory features for any unseen speaker.

The mapping from acoustics to articulations is known to be highly non-linear and non-unique [5]. Speaker variability makes this already challenging problem even more difficult. Most research in speech inversion has been focused on developing accurate speaker-dependent systems. Approaches such as codebook search [6], feedforward neural networks, and Mixture Density Networks [7] have been found to work well for speaker-dependent speech inversion. There have been several attempts to perform speaker independent speech inversion systems [8, 9, 10, 11], but these have been frequently limited to two speakers from the MOCHA-TIMIT dataset [12] (but see [10, 11]).

The goal of this study is to assess how appropriate normalization and deep and shallow neural network techniques may help in creating an adequate speaker-independent acoustic-to-articulatory speech inversion system. To reliably assess the performance of our system, we use articulatory data of more than 40 speakers collected in a research project investigating native and non-native pronunciation of English [13]. Specifically, we focus on two subsets of data collected in this project. The first subset consists of English and Dutch utterances from 21 L1 Dutch speakers (NL data), whereas the second subset consists of English utterances from 22 British English speakers (UK data). Both sets of data contain simultaneously recorded acoustic and electromagnetic articulograpy (EMA) data. Besides using the actual EMA sensor trajectories, we converted the sensor trajectories to Tract Variables (TVs) [14] using geometric transformations (explained in Section 2.2).

We trained separate speech inversion systems on both the NL data as well as the UK data to estimate the EMA sensor positions as well as the TVs. In order to compute the accuracy of the speaker-independent speech inversion systems, we trained and tested them using leave-one-speaker-out cross validation. For the NL data, we trained separate speech inversion systems on exclusively Dutch utterances, English utterances, and both Dutch and English utterances. In the following, we compare the performance of these speaker-independent speech inversion systems across the two datasets. In section 2 we describe the dataset used in our experiments. Section 3 focuses on the speech inversion system, while Section 4 describes the speaker independent, and cross-domain experiments and their results. Section 5 discusses the results and observations followed by conclusions from the results in Section 6.

2. Dataset description

2.1. EMA data

The data used in this study was collected to compare the pronunciation and articulation of English by Dutch speakers to the English pronunciation of native Southern Standard British English speakers (see also [13]). The articulatory data was collected on site (in Groningen, the Netherlands for the Dutch

speakers, and in London, UK for the native English speakers) using an NDI Wave 100 Hz 16-channel articulography device. For the articulatory data collection, three sensors were attached to the midline tongue: one at about half a cm. behind the tongue tip (TT), one about three cm. behind the TT sensor (TB), and the other midway between TT and TB (TM). We further attached three sensors to the lips and two to the teeth: one at the center of the upper lip (at the vermilion border; UL), one at the center of the lower lip (at the vermilion border; LL), and the third in the right corner of the lips (SL). The teeth sensors were attached to the lower incisor (LI) and to the upper incisor (UI). To correct for head movement, we attached four sensors to the head (left and right mastoid process and two at the front of the head), and we used a biteplate with three sensors to rotate all other sensors to a common coordinate system relative to the occlusal plane. The articulatory data was synchronized with the acoustic data, which was collected using a sampling rate of 22.05 kHz (using an Audio Technica AT875R microphone).

In London, we collected data for 22 speakers, whereas we collected data for 21 speakers in Groningen, the Netherlands. For the Dutch speakers, the experiment consisted of two parts. In the first (native Dutch) phase of the experiment, we collected articulatory and acoustic data when the speakers pronounced one paragraph of text (the Dutch version of the North Wind and the Sun), which was followed by the collection of pronunciation data for about 125 words and non-words (in random order, all repeated twice). Each word was preceded and succeeded by a schwa to ensure a neutral articulatory context at the beginning and end of the word pronunciation. In the second (English) phase of the experiment, the participants first pronounced two paragraphs of text (i.e. the North Wind and the Sun, and a paragraph of text used in the Speech Accent Archive [15]), which was followed by about 175 English words and non-words (in random order, each repeated twice, and preceded and followed by the schwa). Finally, if there was still time left, participants were asked to pronounce sentences from the Mocha-TIMIT corpus [12]. For the native English speakers, there was no Dutch phase of the experiment, but the individual words were pronounced both without the schwa context and with the schwa context. In total, this resulted in about 185 minutes of speech for the 21 Dutch speakers (NL data) and 235 minutes of speech for the 22 native English speakers (UK data).

The raw EMA data was corrected for head movement and aligned to the occlusal plane. Missing sensor data (due to sensors which malfunctioned, or came off during the experiment) was estimated using the algorithm outlined in [16]. In short, a probability density of the sensor positions was estimated, and the missing sensor coordinates were approximated using conditional distributions derived from the modeled density [16].

2.2. Conversion of EMA sensors to Tract Variables

The specific EMA data greatly depends on the anatomy of the speaker and the points where the sensors are placed. Vocal tract constriction variables, or tract variables (TVs), are measures of constriction position and location along the vocal tract. Instead of actual coordinates (x: anterior-posterior axis, z: inferior-superior axis) of the sensors, the TVs represent relative positions of the articulators. We converted the EMA sensor trajectories to ten TVs using geometric transformations as shown in Figure 1. The ten TVs were: Lip Aperture (LA), Lip Protrusion (LP), Lip Width (LW), Jaw Aperture (JA), Tongue Tip Constriction Location (TTCL), Tongue Tip

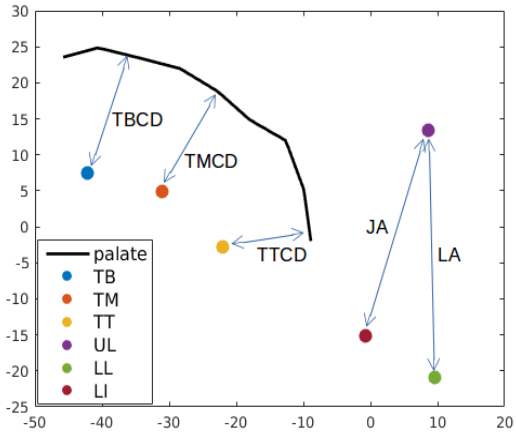

Figure 1: *Transformation of EMA sensor positions to TVs*

Constriction Degree (TTCD), Tongue Middle Constriction Location (TMCL) and Tongue Middle Constriction Degree (TMCD), Tongue Root Constriction Location (TMCL) and Tongue Root Constriction Degree (TMCD). LA was defined as the Euclidean distance between the UL and the LL sensors. LP was defined as the displacement along the x-axis of the LL sensor from its median position. Lip Width (LW) was defined as the Euclidean distance between the SL sensor and the centroid of the UL and LL sensors. JA was defined as the Euclidean distance between the UL sensor and the LI sensor. Two TVs were computed for each tongue sensor - constriction degree and location. Constriction degree for a tongue sensor was defined as the minimum distance between the sensor and the (automatically determined, data-driven) palate trace. This way the TTCD, TMCD, and TBCD TVs were computed from the TT, TM and TB sensor positions and the palate trace. The constriction location for a tongue sensor was defined as the displacement of the sensor along the x-direction from its median position. Thus, TTCL, TMCL, and TBCL were computed from the TT, TM, and TB sensor positions.

3. Speech inversion system

Previous studies (e.g., [17]) have demonstrated that Artificial Neural Networks (ANNs) can be used to reliably estimate sensor and TV trajectories from the speech signal. Once trained, ANNs require low computational resources compared to other methods in terms of both memory requirements and execution speed. In this paper, we trained speech inversion systems using neural networks having one to three hidden layers with the number of hidden layers heuristically determined on the basis of the amount of data. The inputs to the neural network were 13-dimensional MFCCs, which were contextualized with MFCC features from 8 frames on either side. Thus, the input dimension was 13x17 = 221. The outputs of the network were either EMA sensor positions (14 dimensions) or TVs (10 dimensions). The MFCCs were mean and variance normalized separately for every speaker. Similarly, we normalized the means of the EMA sensors to 0 for every EMA recording (to control for minor displacements in case of sensor reattachment) and then normalized the variance speaker wise. We trained neural networks with 300 nodes in each hidden layer. The transfer function for the hidden layer nodes was set to the hyperbolic tangent (tanh), whereas the output nodes used a linear activation function. The networks were trained using

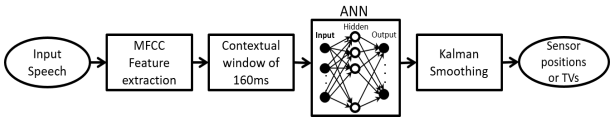

Figure 2: *Block diagram of the speech inversion system*

an Nvidia Titan X GPU using the Keras toolkit [18]. The output of the trained neural network was found to contain high-frequency noise, and was therefore Kalman-filtered to obtain smooth TV / sensor position estimates. Figure 2 shows a schematic representation of our speech inversion system. The performance of our system was evaluated by computing the Pearson product-moment correlation r between the actual and estimated articulatory positions (or tract variables).

$$r = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \quad (1)$$

4. Results

4.1. Leave one speaker out tests

Given the large number of speakers in the UK and NL data, we used leave-one-speaker-out cross-validation (LOCV) to evaluate the speaker-independent speech inversion performance within each dataset. These experiments were performed for both subsets of data separately. The NL data, which consisted of both English and Dutch utterances, was divided into three sets: Dutch utterances (NL_dutch), English utterances (NL_english), and all utterances (NL_all). The UK data only consisted of English utterances (UK_english). The LOCV tests were performed for each of these four sets. Table 1 provides an overview of these systems and the corresponding subsets of data. For the UK dataset, 18 speakers were randomly selected for neural network training, 3 speakers were randomly selected for the validation step (to determine the stopping criterion for the neural network training), and finally the system was tested on the single remaining speaker (i.e. in the LOCV approach, each speaker was included in the test set exactly once). For the Dutch data, a similar approach was used, but with 17 speakers in the training set (as opposed to 18).

The neural networks for the UK_english system had three hidden layers with 300 nodes in each layer. Due to the limited amount of Dutch utterances available in the NL data (see Table 1), the NL_dutch systems were trained with a single hidden layer (with 300 nodes). Similarly, we restricted the number of hidden layers to two (with 300 nodes in each layer) for the NL_english systems. The NL_all systems, which were trained with both the English and Dutch utterances, were given three hidden layers with 300 nodes in each layer. The LOCV experiments were performed separately for estimating EMA sensor positions as well as TVs. For the EMA sensor positions, we estimated the x and z coordinates for all the sensors except for the SL sensor, for which we estimated the x and y (i.e. left-right) positions. The average correlations (on the basis of the LOCV test set results) for the EMA sensor positions are shown in Figure 3, whereas Figure 4 shows the same for the TVs.

4.2. Cross domain experiments

The performance of the speech inversion systems illustrated above shows how well the system has learned to estimate the articulatory patterns of that language. In this section, we report how well our speech inversion system is able to perform in a cross-language setting. For this purpose, we evaluated how well a system trained on the data from the UK English speakers was able to predict the articulatory trajectories of the Dutch speakers (in accented English, Dutch, or both) and vice versa. Instead of training on all data to obtain a new model, we used the best-performing model from the LOCV approach.[1] We evaluated the performance of each of the four systems on all speakers on the basis of the other three subsets of data. For example, the UK_english system was tested on every speaker from the NL dataset (separately for the three subsets of data: Dutch only, non-native English only, or both). The results of these experiments are presented in Table 2 (for the EMA sensor positions) and in Table 3 (for the TVs). Note that the values on the diagonals reflect the average LOCV performance for each

Table 1: *Speech inversion systems and their training data*

System name	Data	Amount of data
UK_english	English utterances from 22 UK English speakers	235 min.
NL_dutch	Dutch utterances from 21 L1 Dutch subjects	60 min.
NL_english	English utterances from 21 L1 Dutch subjects	126 min.
NL_all	English and Dutch utterances from 21 L1 Dutch subjects	186 min.

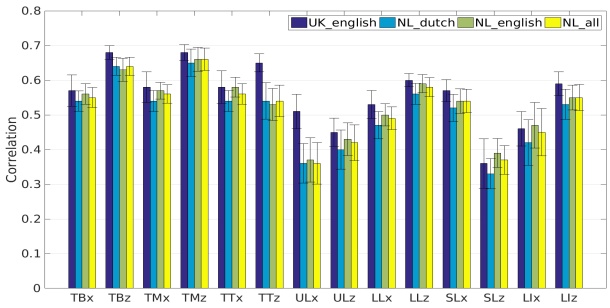

Figure 3: *Average (across all speakers) correlations between actual and estimated EMA sensor positions. Error bars denote two standard errors.*

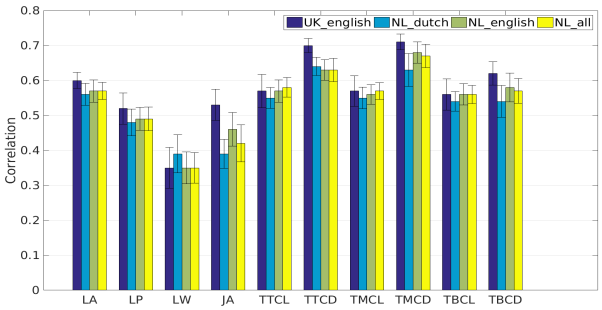

Figure 4: *Average (across all speakers) correlations between actual and estimated tract variables. Error bars denote two standard errors.*

[1] While it is likely that a model on the basis of all data would have been slightly better performing, it is unlikely that this would have impacted the results substantially.

subset of the data shown in Figures 3 and 4.

Table 2: *Average correlations (including standard error) for actual and estimated EMA sensor positions based on different datasets. The left-most column indicates the speech inversion system. The top row indicates the test set. The numbers in the brackets indicate standard errors*

	UK english	**NL dutch**	**NL english**	**NL all**
UK_english	0.56 (0.012)	0.42 (0.015)	0.48 (0.014)	0.45 (0.015)
NL_dutch	0.42 (0.010)	0.51 (0.012)	0.46 (0.012)	0.47 (0.011)
NL_english	0.48 (0.011)	0.48 (0.013)	0.53 (0.011)	0.51 (0.011)
NL_all	0.49 (0.011)	0.51 (0.013)	0.53 (0.011)	0.52 (0.012)

Table 3: *Average correlations (including standard error) for actual and estimated tract variables based on different datasets. The left-most column indicates the speech inversion system. The top row indicates the test set. The numbers in the brackets indicate standard errors*

	UK english	**NL dutch**	**NL english**	**NL all**
UK_english	0.57 (0.012)	0.44 (0.013)	0.51 (0.012)	0.48 (0.014)
NL_dutch	0.43 (0.010)	0.52 (0.012)	0.48 (0.011)	0.49 (0.011)
NL_english	0.50 (0.011)	0.49 (0.013)	0.54 (0.010)	0.52 (0.011)
NL_all	0.51 (0.011)	0.54 (0.011)	0.56 (0.010)	0.54 (0.010)

5. Discussion

In this study we have shown that our system is able to model speaker-independent articulatory positions, with a correlation of about $r = 0.53$. This is substantially lower than the correlation of about $r = 0.62$ reported in [11], but our result do not depend on a specific reference group of speakers. Furthermore, if we exclude the performance with respect to UL and SL for the EMA sensor positions, and the LW tract variable (not included by [11]), these correlations increase to $r = 0.58$. These sensors/tract variables are most difficult to predict given their relatively limited influence on the speech signal.

We only reported correlations rather than mean squared errors, as all experiments are speaker independent and both sensor positions and tract variables were mean and variance normalized. The objective of speaker-independent speech inversion is to accurately capture the trend of the articulatory movements, even though there might be offsets in actual sensor positions due to the anatomical mismatch between training speakers and the speaker used to evaluate the model performance. The performance on the basis of tract variables was only marginally better than the performance based on the EMA sensor positions. As the EMA sensor positions were normalized with respect to their mean and variance, this also (just as tract variables) abstracts away from most anatomical variation.

While cross-language modeling of the trajectories resulted in a lower correlation than the within-language results, the drop in performance was only limited, especially when more data was available (i.e. comparing NL_all to UK_english). Tables 2 and 3 show several evaluations of the speaker-independent speech inversion systems across different test sets. The results in the table highlight the performance of the systems in different mismatch conditions. The native language mismatch condition is highlighted comparing UK_english to NL_all. The NL_all system performs better on the UK_english set than vice versa. This might be due to the fact that the UK data is cleaner (due to being recorded in a soundproof booth) than the NL data. Consequently, the system trained on the clean UK data performs poorly on the NL data. The accent mismatch is highlighted by comparing UK_english to NL_english. We observe that the performance of the UK_english system on the NL_english set is close to the within dataset (NL_english-NL_english) performance. By contrast, the NL_english system performs much worse than the UK_english system on the UK_english dataset. On the one hand, this can be attributed to the higher amount of training data in the UK dataset. On the other hand, the amount of variability in the acoustics and articulatory movements is likely higher for the L2 English speakers (leading to poorer NL_english speech inversion models). Finally, the performance when the language is completely mismatched is shown by the UK_english vs. NL_dutch comparison. Unsurprisingly, we see lower correlations in these comparisons, which can be attributed to both language mismatch as well as a data mismatch. By contrast, the NL_english vs. NL_dutch comparison avoids the problem of mismatched data (i.e. collected at different sites), and their comparison highlights the effect of language mismatch in speech inversion performance (i.e. about 0.05 reduction in the correlation coefficient).

6. Conclusion

The experiments performed in this study shed light on the effects of the amount of training data, the different types of data (i.e. collected in different environments), and different accents and languages on the performance of speech inversion systems. Our results highlight that with appropriate normalizations of the acoustic features and articulatory trajectories, speaker independent systems can estimate the sensor positions and TVs reasonably well with a correlation of about 0.53 with matched training and testing conditions. For mismatched data, the performance drops to about 0.43. Speaker normalization techniques [19, 20] may further improve the performance of these systems. This paper also highlights that data collected using the same protocol may be combined in order to generate improved speech inversion systems, even if the languages are different. In future work, we plan to develop methods for combining data collected with different protocols and potentially even different modalities for the creation of speech inversion systems.

7. Acknowledgements

We would like to thank the University of Maryland Graduate School and the University of Groningen for awarding the International Graduate Research Fellowship to fund this research. This work was made possible by a hardware grant from NVIDIA and a Veni grant for the project “Improving speech learning models and English pronunciation with articulography” awarded to Martijn Wieling by the Netherlands Organisation for Scientific Research (NWO).

8. References

[1] S. Aryal and R. Gutierrez-Osuna, "Accent conversion through cross-speaker articulatory synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2014, pp. 7694–7698.

[2] 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, p. 319, 2002.

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

[4] 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 *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2014, pp. 3017–3021.

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

[6] 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, oct 2010.

[7] K. Richmond, "Trajectory Mixture Density Networks with Multiple Mixtures for Acoustic-Articulatory Inversion," in *Advances in Nonlinear Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 263–272.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] M. Wieling, P. Veenstra, P. Adank, A. Weber, and M. Tiede, "Comparing L1 and L2 speakers using articulography," in *Proceedings of ICPhS 2015*, 2015.

[14] E. L. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, dec 1989.

[15] S. H. Weinberger, "speech accent archive." [Online]. Available: http://accent.gmu.edu/about.php

[16] C. Qin and M. Carreira-Perpiñán, "Estimating missing data sequences in x-ray microbeam recordings," in *INTERSPEECH*, 2010.

[17] 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 of selected topics in signal processing*, vol. 4, no. 6, pp. 1027–1045, sep 2010.

[18] F. Chollet, "Keras," \url{https://github.com/fchollet/keras}, 2015.

[19] 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 *INTERSPEECH*, sep 2016, pp. 455–459.

[20] 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, mar 2017.