Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion

Ganesh Sivaraman,[1,a)] Vikramjit Mitra,[1] Hosung Nam,[2] Mark Tiede,[3] and Carol Espy-Wilson[1]
[1]*Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20740, USA*
[2]*Korea University, Seoul, South Korea*
[3]*Haskins Laboratories, New Haven, Connecticut 06511, USA*

(Received 7 April 2018; revised 22 May 2019; accepted 20 June 2019; published online 23 July 2019)

Speech inversion is a well-known ill-posed problem and addition of speaker differences typically makes it even harder. Normalizing the speaker differences is essential to effectively using multi-speaker articulatory data for training a speaker independent speech inversion system. This paper explores a vocal tract length normalization (VTLN) technique to transform the acoustic features of different speakers to a target speaker acoustic space such that speaker specific details are minimized. The speaker normalized features are then used to train a deep feed-forward neural network based speech inversion system. The acoustic features are parameterized as time-contextualized mel-frequency cepstral coefficients. The articulatory features are represented by six tract-variable (TV) trajectories, which are relatively speaker invariant compared to flesh point data. Experiments are performed with ten speakers from the University of Wisconsin X-ray microbeam database. Results show that the proposed speaker normalization approach provides an 8.15% relative improvement in correlation between actual and estimated TVs as compared to the system where speaker normalization was not performed. To determine the efficacy of the method across datasets, cross speaker evaluations were performed across speakers from the Multichannel Articulatory-TIMIT and EMA-IEEE datasets. Results prove that the VTLN approach provides improvement in performance even across datasets. © *2019 Acoustical Society of America*.
https://doi.org/10.1121/1.5116130

[JFL] Pages: 316–329

I. INTRODUCTION

Speech inversion or acoustic-to-articulatory inversion of speech has been a widely researched topic in the last 40 years. Speech Inversion is the process of mapping the acoustic signal into articulatory parameters. If estimated accurately, articulatory information can be applied to speech accent conversion (Aryal and Gutierrez-Osuna, 2014), speech therapy (Cavin, 2015; Preston *et al.*, 2014), language learning, Automatic Speech Recognition (ASR) (Kirchhoff *et al.*, 2002; Mitra, 2010; Mitra *et al.*, 2014b), and detection of depression from speech (Helfer *et al.*, 2013; Mitra *et al.*, 2014a). Real articulatory data is obtained from subjects using techniques like Electromagnetic Articulometry (EMA) (Schönle *et al.*, 1987), X-ray microbeam (Westbury, 1994), and real-time Magnetic Resonance Imaging (rt-MRI) (Narayanan *et al.*, 2004). However, these techniques require sophisticated devices and are expensive and time consuming. Obtaining real articulatory data is not practically feasible for real world applications like ASR. Only the acoustic data is available from the speaker. Hence, it is essential to develop speech inversion systems that are speaker independent and can accurately estimate articulatory features for any unseen test speaker. The mapping from acoustics to articulations is known to be highly non-linear and non-unique (Qin and Carreira-Perpiñán, 2007). Adding speaker variability to the already challenging problem makes it even more difficult. Most research in speech inversion has been focused on developing accurate speaker dependent systems. Based on a comprehensive study of the speech inversion techniques, the speaker dependent techniques can be classified into three categories: (1) codebook based approaches (Atal *et al.*, 1978; Ouni and Laprie, 2005) in which a codebook of acoustic and corresponding articulatory patterns is constructed from the training data, (2) analytical approaches involving articulatory models such as Maeda's model (Krstulović, 2001; Laprie and Mathieu, 1998), and (3) statistical modeling (parametric and non-parametric) of acoustic to articulatory mapping like Gaussian Mixture Model (GMM) (Toda *et al.*, 2004), Mixture density networks (MDN) (Richmond, 2006), multilinear regression and principal components analysis (Mokhtari *et al.*, 2007), Hidden Markov models (HMM) (Hiroya and Honda, 2004), generalized smoothness criteria (Ghosh and Narayanan, 2010), and neural networks (King and Taylor, 2000; Kirchhoff, 1999; Mitra *et al.*, 2010, 2014b).

There have been a few attempts to perform speaker independent speech inversion (Afshan and Ghosh, 2015; Ghosh and Narayanan, 2011; Ji, 2014), which have been limited to two speakers from the Multichannel Articulatory-TIMIT (MOCHA-TIMIT) dataset (Wrench, 2000). In Ghosh and Narayanan (2011), a subject independent speech inversion system is developed by representing acoustic features

[a)]Electronic mail: ganesa90@gmail.com

with respect to a generic acoustic space trained on a dataset containing multiple speakers. While testing, the test speaker's acoustic features are matched with the training speaker's acoustic features with respect to the generic acoustic space. The generic acoustic space normalizes the mismatch between the training and test speakers. Afshan and Ghosh (2015) extended the subject independent inversion system by incorporating speaker adaptation techniques. They proposed supervised and unsupervised ways of clustering the generic acoustic space to perform speaker normalization. They performed adaptation of the GMM based generic acoustic space using maximum likelihood linear regression (MLLR) that is commonly used in speaker adaptation of ASR acoustic models. They found that availability of phone transcriptions of the adaptation data improves the performance of the speaker adaptation. In Ji (2014), an unseen speaker's acoustic feature space is approximated with a weighted combination of the acoustic spaces of the training speakers by a maximum likelihood based weighting scheme. This method is unsupervised and does not assume availability of phone transcriptions. Hueber *et al.* (2015) presents a Gaussian mixture regression based speaker adaptation scheme for a GMM based speech inversion system. To the best of our knowledge, there has not to date been any effort in performing speaker adaptation for artificial neural network (ANN) based speech inversion systems. Modeling speaker variability and training speaker independent speech inversion models has been a challenge facing acoustic-to-articulatory speech inversion. Mokhtari *et al.* (2000) modeled inter-speaker variability in acoustic-to-articulatory mapping in terms of three components: structure, setting, and strategy. Their objective was to model the speaker variability in vocal tract area functions estimated using linear prediction based methods. They defined the "structure" component of speaker variability as the mean vocal tract length of each speaker.

This paper aims to minimize the speaker variability in the acoustic space attributed to the vocal tract length differences between speakers for performing acoustic-to-articulatory inversion. This paper presents a Vocal Tract Length Normalization (VTLN) based approach to speaker adaptation for speech inversion. The VTLN is approximated as a non-linear warping of the frequency axis in the filterbank analysis to adapt a test speaker's acoustic space to a target speaker. The non-linear warping function of VTLN is optimized to increase the acoustic similarity between two speakers. The objective of the warping function is to normalize the variations in the formant frequencies arising due to varying lengths of vocal tract. VTLN does not necessarily normalize the differences in vocal tract lengths. The technique also does not explicitly compute the vocal tract lengths of the speakers. It was commonly used as a speaker adaptation technique in HMM based ASR. To our knowledge, this paper is the first ever to perform VTLN for speaker adaptation in speech inversion. This method does not assume availability of any phonetic transcripts or a supervised acoustic model for performing the adaptation.

The objective of this paper is to normalize acoustic data from multiple speakers towards the acoustic space of a target speaker. Diagonal covariance GMMs are trained for each speaker. Given a test speaker's utterance, a piecewise linear frequency warping is applied to the frequency axis of the mel-filterbank to adapt the acoustic space of the test speaker towards that of the target speaker. The parameter of the piece-wise linear warping function is determined such that the warping maximizes the likelihood of the test speaker in the target speaker's acoustic space (GMM). More details about this adaptation procedure are provided in Sec. V.

The key contributions of this paper are as follows:

(1) Estimating articulatory constriction variables instead of flesh point trajectories or EMA sensor trajectories. Most works in the speech inversion literature have focused on estimating actual X-Y positions of EMA sensors or the X-ray Microbeam (XRMB) pellets from acoustics. In this paper, we present methods to convert the raw articulatory measurements to tract variables (TVs) and train systems to estimate the TVs from speech. This is significant because tract variables represent goal-directed synergies among the articulators, which are thus inherently more stable than the articulatory positions themselves, subject as they are to coarticulatory pressures.
(2) Speaker independent speech inversion system trained on large number of speakers (46 speakers) from the Wisconsin XRMB database (Westbury, 1994). To the best of our knowledge, this is the first time a speaker independent speech inversion system has been trained on such a large number of speakers.
(3) A novel unsupervised speaker adaptation technique based on VTLN to adapt a test speaker towards a target speaker for speech inversion.
(4) Cross speaker evaluation of speaker dependent speech inversion systems. The proposed VTLN speaker adaptation improved the performance of the speech inversion systems in mismatched speaker and gender scenario.
(5) Cross-corpus speech inversion experiments. The proposed VTLN based speaker adaptation improved the performance of speech inversion even in the highly challenging cross-corpus experiments. To the best of our knowledge, this is the first time cross-corpus speech inversion experiments have been performed.

We perform experiments on three different datasets: (1) Wisconsin XRMB database (Westbury, 1994), (2) EMA-IEEE dataset (Tiede *et al.*, 2017), and (3) MOCHA-TIMIT dataset (Wrench, 2000).

Our speech inversion system is a neural network based system that maps contextualized acoustic features to vocal tract constriction variables (TVs). The architecture of the speech inversion system is fixed across all our experiments, except for the number of hidden layers and nodes in the neural networks. We describe the architecture of the speech inversion system in Sec. III.

We first train a speaker independent speech inversion system using all the data from the University of Wisconsin XRMB database (Westbury, 1994) dataset (46 speakers). We use a deep neural network with 5-hidden layers to map contextualized Mel Frequency Cepstral Coefficients (MFCCs) to TVs. The details of the speaker independent speech inversion and the results are presented in Sec. IV.

Section VI presents cross-speaker speech inversion experiments on the XRMB dataset. Due to the complexity of the experiment and the large number of mismatched speaker trials, we performed the experiments on a randomly selected subset of the XRMB dataset consisting of ten speakers. We trained speaker dependent speech inversion systems with single hidden layer neural networks for all the ten speakers and then evaluated the mismatched speaker performance of the systems. In the mismatched speaker tests, we also applied the VTLN based speaker adaptation to adapt test speakers to the target speaker. We present the cross-speaker experiments and their results in Sec. VI.

We performed leave-one-speaker-out experiments on the ten-speaker subset of the XRMB dataset. In this experiment too, due to the small amount of data available per speaker, we used shallow single-hidden-layer neural networks for training the speech inversion systems. Separate experiments were performed for each speaker in which the acoustic features from the other nine speakers were transformed using the VTLN approach. The transformed acoustic features were then used to train a speech inversion system. The performance of the system trained on VTLN adapted acoustic features was compared to the performance of speaker dependent systems. More details of the leave-one-speaker-out experiments speech inversion system training and the experiments are provided in Sec. VII.

In most studies in the literature, speech inversion systems are trained and evaluated on the same dataset. However, for all practical purposes speech inversion systems are used on speech utterances previously unseen by the inversion system. This paper recognizes this as a necessary challenge to address and performs cross-corpus speech inversion experiments and evaluates the strength of the VTLN adaptation for cross corpus evaluation. We perform the cross-corpus speech inversion experiments on the MOCHA-TIMIT and EMA-IEEE datasets. It is observed that the VTLN adaptation procedure provides a 8.15% relative improvement in correlation on average compared to the performance without speaker adaptation in a multi-speaker cross-corpus evaluation.

A summary of all the experiments and their results are presented in Sec. IX.

II. ARTICULATORY DATASETS

This section describes the articulatory datasets used in the experiments performed in this paper.

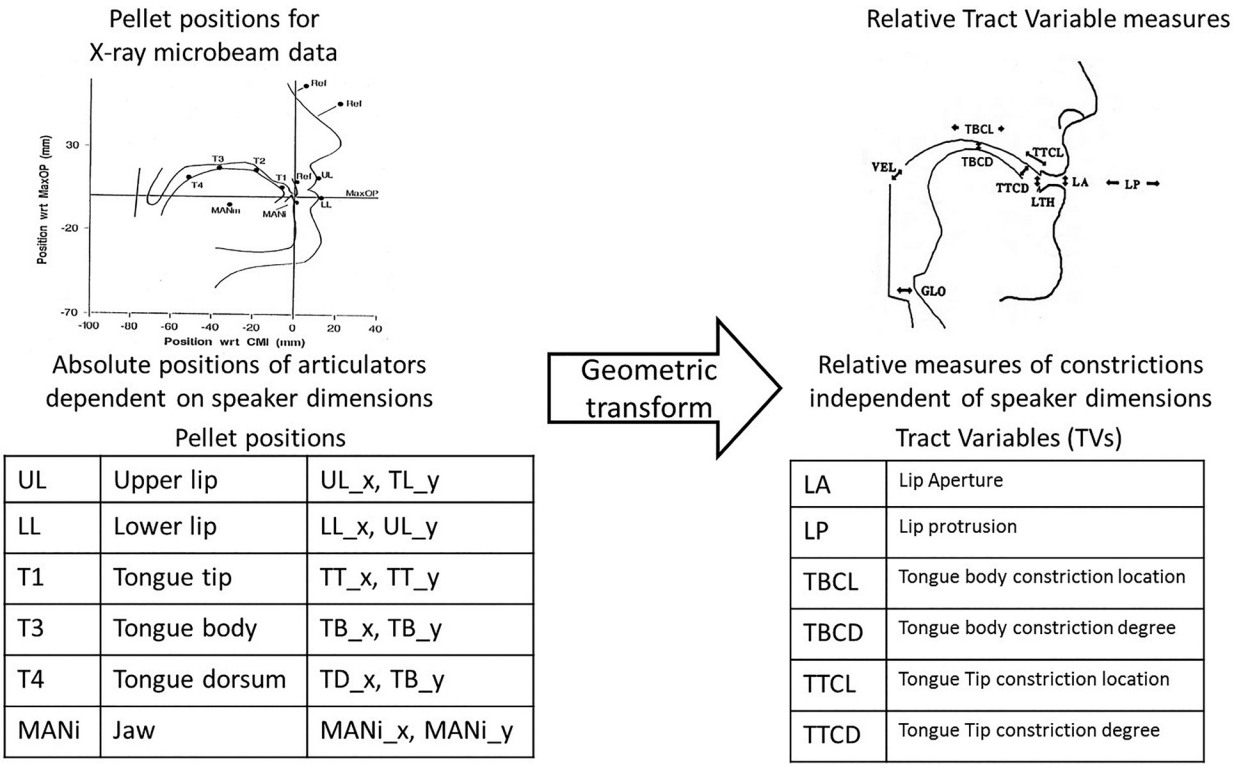
A. XRMB

The Wisconsin XRMB database (Westbury, 1994) consists of naturally spoken utterances: isolated sentences and short read paragraphs. Speech audio was collected from 32 males and 25 females along with X-ray microbeam cinematography of the mid-sagittal plane of the vocal tract with tracked pellets placed at four points on the tongue, upper, and lower lips. Figure 5.2 in Westbury (1994) shows the placement of the pellets on the articulators as a midsiggital view of the vocal tract. Trajectory data were recorded for pellets placed midsagittally on these 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).

A common problem with articulatory recordings of this type is the mistracking of pellets or the pellets falling off while recording, which are marked as mistracked samples in the XRMB database. These samples were removed from the database before using it for our analysis.

1. Converting XRMB pellets to TVs

The X-Y positions of the pellets are closely tied to the anatomy of the speakers and can therefore vary considerably across speakers for the same sound and may also vary considerably due to small differences in the pellet placement. Speech production involves the shaping of the supralaryngeal vocal tract filter by producing constrictions at different places along the vocal tract using the articulators. Hence, the quantification of the vocal tract shape is better performed by the location and degree of these constrictions, which are relative measures compared to the absolute measures of 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 (McGowan, 1994) and provide a relatively speaker independent representation of speech articulation. Developed within the context of Articulatory Phonology (Browman and Goldstein, 1992), they also provide a useful theoretical framework for the analysis of speech production with the theoretical framework of articulatory phonology. TVs characterize the location and degree of vocal tract constrictions without reference to the specific synergies that achieve them; for example, the "lip aperture" (LA) tract variable represents the degree of occlusion at the acoustic terminus, without the need to specify the individual contributions of the jaw, upper and lower lips. Because of these advantages, the XRMB trajectories were converted to TV trajectories using geometric transformations as outlined in Mitra *et al.* (2012). As defined in the Task Dynamic model (TADA) of speech production, the hard palate was approximated with a large circle using curve fitting through the palate trace. The TADA model also approximates the tongue body as a smaller circle within the larger circle approximating the palate. Hence, the tongue body was approximated with a circle passing through the pellets T2, T3, and T4 at each time step. The tongue tip was modeled separately by the segment T2-T1. In this manner, at each time step, the pellet X-Y positions were converted to TVs. Out of the total of 57 speakers in the XRMB dataset, 46 were successfully converted to TVs. The remaining speakers could not be transformed due to a higher proportion of mis-tracked segments in the pellet trajectories. The transformed XRMB database consists of 21 males and 25 females, with a total of 4 h of speech data with corresponding six TV trajectories. The TVs obtained from the seven pellet trajectories were: 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). A rough schematic of the transformation is shown in Fig. 1.



Pellet positions

UL	Upper lip	UL_x, TL_y
LL	Lower lip	LL_x, UL_y
T1	Tongue tip	TT_x, TT_y
T3	Tongue body	TB_x, TB_y
T4	Tongue dorsum	TD_x, TB_y
MANi	Jaw	MANi_x, MANi_y

Tract Variables (TVs)

LA	Lip Aperture
LP	Lip protrusion
TBCL	Tongue body constriction location
TBCD	Tongue body constriction degree
TTCL	Tongue Tip constriction location
TTCD	Tongue Tip constriction degree

FIG. 1. Schematic of transformation of XRMB database from pellets to TV trajectories. Parts of the figure are taken from Saltzman and Munhall (1989) and Westbury (1994).

B. EMA-IEEE dataset

A five-dimensional (5 D) EMA system (WAVE; Northern Digital) was used to record the 720 phonetically balanced IEEE sentences (Rothauser *et al.*, 1969) from eight speakers (four males, and four females) at normal and fast production rates (Tiede *et al.*, 2017). Participants produced each sentence twice, first at their preferred "normal" speaking rate followed by a "fast" production (for a subset of the sentences two normal rate productions were elicited). They were instructed to produce the "fast" repetition as quickly as possible without making errors. EMA trajectories were obtained at 100 Hz from sensors 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). The data were low-pass filtered at 5 Hz for references and 20 Hz for articulator sensors, corrected for head movement and aligned to the occlusal plane. Synchronized audio was recorded at 22050 Hz, using a directional shotgun microphone placed 50 cm from the speaker's mouth. In this paper, we have used only the normal rate utterances from the EMA-IEEE dataset for cross-corpus speech inversion experiments detailed in Sec. VIII.

1. Conversion of EMA sensor positions to TVs

The EMA sensor trajectory data was converted to nine TVs using geometric transformations. The nine TVs were: LA, LP, JA, TTCL, TTCD, Tongue Middle Constriction Location (TMCL), Tongue Middle Constriction Degree (TMCD), Tongue Root Constriction Location (TRCL), and Tongue Root Constriction Degree (TRCD).

LA was defined as the Euclidean distance between the UL and the LL sensors as shown in Eq. (1)

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

LP was defined as the displacement along the x axis of the LL sensor from its median position as shown in Eq. (2)

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

JA was defined as the Euclidean distance between the UL sensor and the LI sensor as shown in Eq. (3)

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

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 palate trace as shown in Eq. (4). The palate trace was one of the measurements taken as part of the data collection for each subject. The palate trace provides only the height (z-coordinate) of the hard palate along the anterior-posterior axis (x axis). The function $pal(x)$ shown in Eq. (4) gives the z-coordinate of the palate for 100 points within the range $x \in (-50, 0)$. Computationally, pal(x) is just an array of z-coordinates of the hard palate for different values of x.

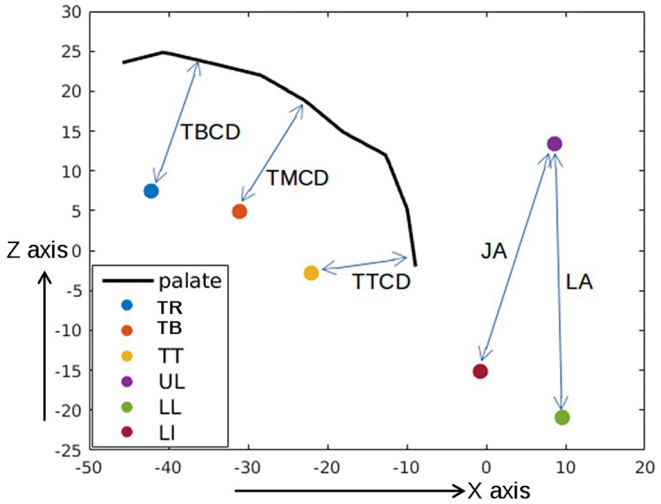


FIG. 2. (Color online) Transformation of EMA sensor positions to TVs.

$$TTCD[n] = \underset{x\in(-50,0)}{\text{Min}} \left\{ \sqrt{(TT_x[n]-x)^2 + (TT_z[n]-pal(x))^2} \right\}. \quad (4)$$

The same way the 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 as shown in Eq. (5).

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

The same way, TTCL, TMCL, and TBCL were computed from the TT, TM, and TB sensor positions.

C. MOCHA-TIMIT dataset

The Multichannel Articulatory (MOCHA) database (Wrench, 2000) contains acoustic and simultaneous EMA data from one male and one female speaker of British English. The EMA data from the MOCHA database was smoothed and downsampled from 500 to 100 Hz as described in Richmond *et al.* (2003). We converted the EMA sensor position data to nine TVs using the same approach described in Sec. II B 1 (Fig. 2).

III. SPEECH INVERSION SYSTEM DESCRIPTION

Mitra *et al.* (2010) explored various machine learning approaches to acoustic-to-articulatory speech inversion. Based on the comparison of the different machine learning algorithms (Mitra *et al.*, 2010), we chose ANNs to be the best suited approach for estimating TVs from speech. This is a function mapping approach to speech inversion where the frame wise input acoustic features are mapped to frame wise measurements of TVs which represent the instantaneous configuration of the vocal tract. With the advent of Deep Neural Networks (DNN), faster learning strategies and higher computational power, it has been shown that deep architectures can represent certain families of functions more efficiently than shallow ones (Bengio and Lecun, 2007). Hence, we explore feedforward DNNs for learning the mapping from acoustics to TVs. This section describes the acoustic-to-articulatory speech inversion system architecture that has been used throughout this work.

A DNN can have M inputs and N outputs; hence, a nonlinear complex mapping of M vectors into N different functions can be achieved. In such an architecture, the same hidden layers are shared by all N outputs, giving the DNN the implicit capability to exploit any correlation that the N outputs may have amongst themselves. The feed-forward DNN used in our study to estimate the TVs from speech were trained with back propagation using a stochastic gradient descent algorithm.

The system shown in Fig. 3 outlines the blocks involved in the speech inversion system design. The details of the speech inversion system are given in Secs. III A and III B.

A. Feature extraction

The utterances were downsampled to 8 kHz. The input features to the neural network were varied and compared. We experimented with different acoustic features: MFCC, Perceptual Linear Prediction (PLP) and mel-spectrum (MELSPECT). Single hidden layer neural networks to estimate TVs were trained for each feature type and the best performing feature was chosen for fine tuning. For MFCCs, 13 cepstral coefficients were extracted using a Hamming analysis window of 20 ms with an inter-frame interval of 10 ms. The TVs and MFCCs were mean and variance normalized to have zero mean and a variance of 0.25. As described in Sec. IV, two different methods of mean and variance normalization were performed and compared. The mean and variance normalization was performed separately for every speaker in the database. This ensured some normalization of inter-speaker variations in measurements of acoustics and articulations. The MFCCs were then contextualized by concatenating every other feature frame within a 350 ms window. Since the articulatory movements are smooth and slow varying compared to acoustics, concatenation of adjacent MFCC frames is essential to learn a mapping from acoustics to TVs. After performing experiments by varying the feature splicing

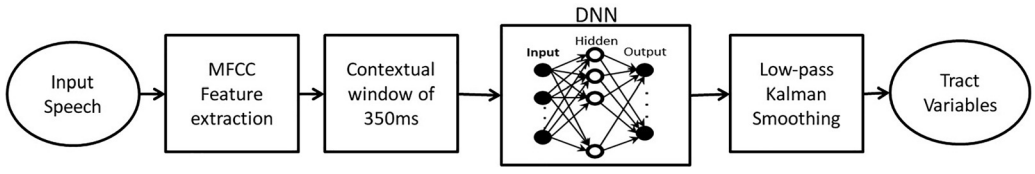


FIG. 3. Block diagram of the speech inversion system.

widths from 60 to 500 ms, the splicing width of 350 ms was found to be the best performing splicing width. This amounted to eight frames of MFCCs on either side of each frame being concatenated to form the contextualized MFCC features. While splicing the frames, we skipped two frames, thus concatenating every other frame within a 35 frame window centered at the current analysis frame. The experiments with other features were performed by adding the same amount of context as for MFCCs.

B. DNN Training

The dimension of the input to the neural network was 221 for MFCC features (=13 MFCCs × 17 frames) and the output dimension was 6 (= 6 TVs). The speakers in the dataset were split into train, validation, and test sets. Thirty-six speakers were assigned for training, and five each for validation and test sets. The splitting of speakers was random such that the training set consisted of no more than 80% of the utterances and the test and validation sets contained nearly an equal number of utterances. Note that the number of utterances from each speaker is not the same due to mistracked segments. A 3 hidden layer neural network was trained. First, a DNN with 1024 neurons in each hidden layer was trained with different acoustic features as inputs. The best performing feature on the XRMB validation set was selected and then the network parameters like number of hidden layers and number of neurons in each layer were tuned. Networks with different numbers of hidden-layer neurons (128–1024) were trained, and among them the best performing network on the validation set was chosen. It was observed that the outputs of the neural network were not as smooth as the original TVs. TVs being vocal tract movements are necessarily smooth signals (Hogden, 1996). Hence, a low-pass Kalman smoothing was performed to remove estimation noise by the neural network. The performance of the TV estimator was measured by computing the Pearson Product Moment Correlations (PPMC) of the estimated TVs with the groundtruth TVs on the test set

$$r = \frac{\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{1}^{n}(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}. \quad (6)$$

The Kalman smoothed TVs showed high correlation with the original TVs and lower mean squared error (MSE).

IV. SPEAKER INDEPENDENT SPEECH INVERSION EXPERIMENT ON MULTI-SPEAKER XRMB DATASET

To begin with, we experimented with different types of acoustic features for the speech inversion system in order to figure out the best feature representation. We performed this experiment to select the best feature by fixing the neural network architecture to a 3-hidden-layer network. Later, once we select the best acoustic feature, we fine tune the number of hidden-layers in the neural network. As described in Sec. III, 3-hidden-layer neural networks with 1024

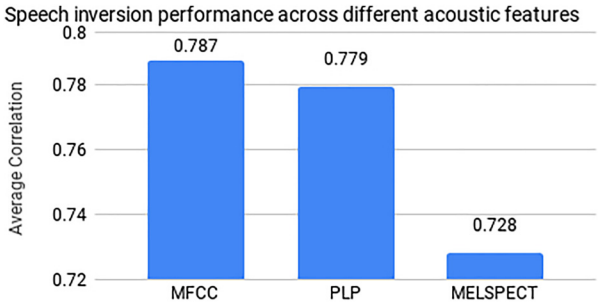


FIG. 4. (Color online) Average correlation results across six TVs for different input features (MFCC, PLP, and MELFB).

neurons each were trained to estimate TVs using three different types of acoustic features. The acoustic features we considered for our experiment were MFCC, PLP, and MELSPECT features. The MFCC and PLP features were 13 dimensional cepstral coefficients per frame. The MELSPECT feature contained 40 mel-filterbank energies for every frame. For each of these features, the analysis frame width was 20 ms and the shift was 10 ms. The input features were contextualized by concatenating eight frames on either side. The results on the XRMB cross validation set from these experiments are presented in Fig. 4. The results are Pearson correlations between actual and estimated TVs.

Based on the results shown in Fig. 4, the TV estimator performed best with MFCCs. As a result, MFCCs were used for all further experimentation. We next focused on tuning the DNN parameters for the MFCC feature based speech inversion system. We trained DNNs with 1, 2, 3, 4, and 5 hidden layers with 128, 256, 512, 1024, and 2048 neurons in each layer. Thus, we trained 25 such DNNs for mapping contextualized MFCCs to TVs. We computed the correlation between actual and estimated TVs for the validation set and selected the best performing configuration. Figure 5 shows the plot of the correlations for different network configurations. Based on the plot, we can see that a 5 hidden-layer DNN with 512 nodes in each layer performed the best. The 5 hidden layer model with 2048 nodes in each layer failed to train due to limited data. The performance of the networks beyond 5 hidden layers saturated, and hence we limited our DNN to 5 hidden layers.

We experimented with two different types of feature and target normalizations: global normalization and speaker

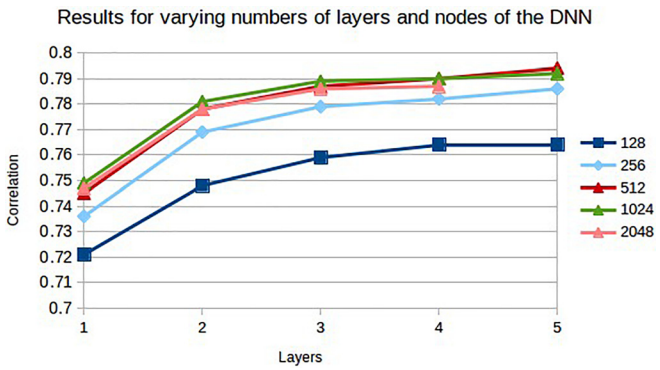


FIG. 5. (Color online) Results of varying DNN parameter (number of hidden layers and number of nodes per hidden layer) on XRMB validation set.

TABLE I. Correlation results for the final XRMB speech inversion system.

	LA	LP	TBCL	TBCD	TTCL	TTCD	Average
Crossval set	0.809	0.678	0.873	0.761	0.769	0.877	0.794
Test set	0.856	0.613	0.866	0.745	0.707	0.907	0.782

specific normalization. In the global mean and variance normalization scheme, all the MFCCs and TVs from the XRMB database were normalized with the global mean and variance estimated from all the utterances. In the speaker-specific normalization approach, the MFCCs and TVs were mean and variance normalized separately for each speaker. We found that the correlations on development set were 12.75% better with speaker specific normalization relative to the global normalization. Hence, for all of our experiments henceforth, we normalize the MFCCs and TVs in a speaker specific manner.

After performing the fine tuning of the speech inversion system, the final best performing neural network architecture was a 5 hidden layer DNN with 512 nodes in each layer. The feature and target normalization chosen was SPKNORM. We will call this speech inversion system XRMB TV estimator (alternatively, as XRMB speech inversion system) and it will be used for various other experiments in the upcoming sections. The Pearson correlation results of the XRMB speech inversion system are shown in Table I. Figure 6 shows example plots of the estimated and actual TVs for three utterances from the XRMB test set.

V. SPEAKER NORMALIZATION TO COMBAT ACOUSTIC VARIABILITY

There is a significant amount of speaker specific component in the acoustic as well as the articulatory domains. The representation of articulatory features as TVs reduces the dependence of articulatory domain on speaker anatomy, but the acoustic features still contain speaker differences arising due to differences in pitch, vocal tract length, speaking rate, and prosody. This section presents a VTLN based approach to speaker adaptation for speech inversion. VTLN is a popular speaker adaptation technique in ASR which has so far not been applied to speech inversion.

VTLN (Eide and Gish, 1996) uses a piecewise linear warping function applied to the frequency axis in the filterbank analysis. The warping of the frequency axis is aimed at reducing the cross-speaker differences in the range of formant frequencies (of phonemes) arising due to differences in vocal tract length. The technique does not explicitly involve the estimation of a speaker's vocal tract length. The objective of VTLN is to maximize the "similarity" (in a probabilistic sense) between the acoustic features of two speakers. This is a commonly adopted approach for speaker adaptation in speech recognition. We applied VTLN in a maximum likelihood framework to adapt the acoustic features of the mismatched speakers to the target speaker. In order to perform VTLN, a speaker dependent acoustic space using GMM was trained on each of the ten speakers.

The experiments that follow are performed on a set of ten speakers from the University of Wisconsin XRMB database (Westbury, 1994). The articulatory features are represented by six TV trajectories as described earlier in Sec. II A. Using a leave-one-out methodology, separate experiments were performed for each speaker in which the acoustic features from the other nine speakers were transformed using the VTLN approach. The transformed acoustic features were then used to train a speech inversion system. The performance of the system trained on VTLN adapted acoustic features was compared to the performance of speaker dependent systems. The performances of the individual systems were compared using the correlation between the estimated and the actual TVs on the target speaker's test set.

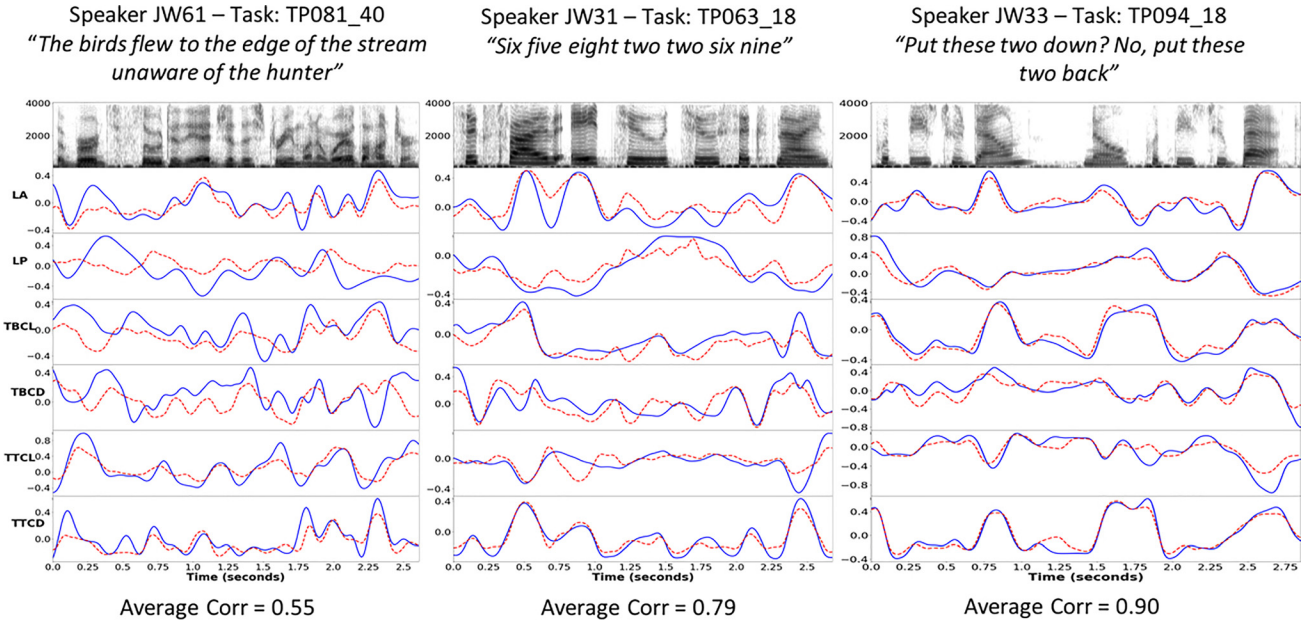


FIG. 6. (Color online) Example plots of estimated (dashed red line) and actual (solid blue line) TVs for three test set utterances. The average correlations between estimated and actual TVs for the example in the left pane, middle pane, and right pane are 0.55, 0.79, and 0.90, respectively. The speakerID, utteranceID and the sentence spoken are given on top of each pane.

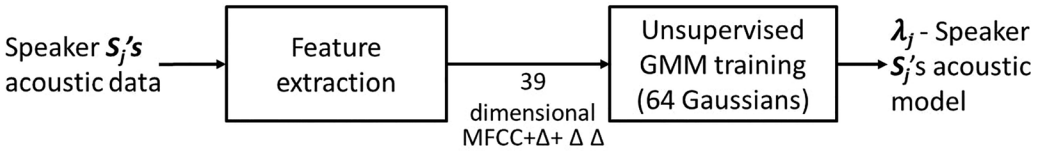


FIG. 7. Training of GMM speaker acoustic spaces.

More details of the speech inversion system training and the experiments are provided in the upcoming sections.

A. Speaker acoustic spaces

In this paper, we define speaker acoustic spaces as probability distributions (Gaussian mixture) that approximately fit the acoustic features of a particular speaker. The 13 dimensional MFCCs with their slope and acceleration were used as acoustic features for modeling the speaker acoustic spaces. GMMs with 64 Gaussian components were trained on the 39 dimensional MFCC + Δ + ΔΔ features. While the speech inversion system accepts MFCC features contextualized with eight frames on either side (221 dimensional feature vector), the speaker acoustic space is modeled on MFCC + Δ + ΔΔ features. This is because the GMMs can be trained faster and efficiently on features with uncorrelated components. With uncorrelated feature components, we can also use the diagonal covariance matrix for the Gaussians in the GMM which can be trained faster. The diagonal covariance GMMs were trained iteratively by increasing the number of Gaussians from 2 to 64 by doubling the number of components in each stage. The GMM training routines were obtained from the MSR Identity Toolbox v1.0 (Sadjadi *et al.*, 2013). Thus, such GMMs were trained for each of the ten speakers chosen for the cross-speaker evaluation. Figure 7 shows the block diagram of the system used to train unsupervised speaker acoustic spaces. The training is unsupervised because we do not use any kind of phone alignments for training phone-wise GMM like in HMM based ASR. Instead, we let the GMMs fit the distribution of the acoustic features for each speaker. Each model λ_i is a 64 component GMM modeling the distribution of MFCCs for speaker S_i.

B. Maximum likelihood based VTLN

VTLN aims to compensate the effects of different vocal tract lengths by warping the frequency spectrum in the filterbank analysis before the computation of the cepstral coefficients. This warping can be implemented by a simple piecewise linear warping function as shown in Fig. 8. The warping factor α determines the nature of the warping function. The warping is implemented between the lower boundary of frequency analysis (f_L) and the upper boundary of frequency analysis (f_U). In all experiments, we fixed f_L at 60 Hz and f_U at 3200 Hz. The parameters were selected based on the recommended default values for HTK's implementation of VTLN. Varying f_L and f_U would provide a wider range of non-linear warping function, however, the large search space would make it intractable for a grid search. Hence, we fixed f_L at 60 Hz and f_U at 3200 Hz. In order to adapt the acoustic features of speaker S_i to speaker S_j, a single warping factor α_{ij} is used for all utterances from speaker S_i. The warping factor α_{ij} is determined by a maximum likelihood approach as outlined below.

Let the GMM acoustic model for speaker S_j be λ_j, and the warped acoustic features for the tth time frame of an utterance of speaker S_i to the target speaker S_j be $x_{ij}(t)$. Then, the most likely warping factor α_{ij} is given by

$$\alpha_{ij} = \arg\max_{\alpha} \sum_{t=1}^{N} \log\left(P(x_{ij}(t)|\lambda_j, \alpha)\right). \tag{7}$$

In Eq. (7), $\sum_{t=1}^{N} \log\left(P(x_{ij}(t)|\lambda_j, \alpha)\right)$ is the log likelihood of the transformed features of speaker S_i with respect to speaker S_j's acoustic model. The conditional probability $P(x_{ij}(t)|\lambda_j)$ for a given α value is computed by converting as follows:

- Apply the frequency warping function corresponding to parameter α on speaker S_i's filterbank spectrum.
- Compute the cepstral coefficients x_{ij} from the warped frequency spectrum of speaker S_i.
- Compute the likelihood $P(x_{ij}(t)|\lambda_j)$ of the transformed cepstral coefficients x_{ij} given speaker S_j's GMM acoustic model.

The optimal α_{ij} is obtained by sweeping the value of α_{ij} from 0.8 to 1.2 in steps of 0.025. As discussed in Zhan *et al.* (1997), most of the VTLN warping factors lie between 0.8 and 1.2 for speaker adaptation in a large vocabulary ASR task. This forms our basis for searching for the optimal warping factor in the range of 0.8 to 1.2. Using the optimal α_{ij}, we compute the speaker adapted acoustic features for speaker S_i adapted to speaker S_j.

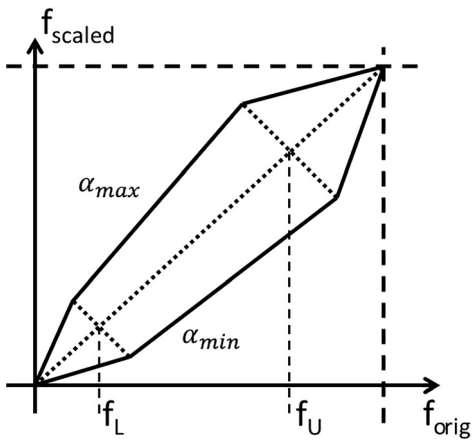


FIG. 8. Frequency warping function implemented in HTK toolkit (Young *et al.*, 2009).

TABLE II. Correlation results for speaker dependent speech inversion systems.

	Spk ID	Gender	LA	LP	TBCL	TBCD	TTCL	TTCD	Average
Spkr 1	JW12	M	0.837	0.821	0.908	0.828	0.792	0.905	0.848
Spkr 2	JW14	F	0.826	0.698	0.927	0.840	0.864	0.902	0.843
Spkr 3	JW24	M	0.824	0.769	0.907	0.773	0.764	0.827	0.811
Spkr 4	JW26	F	0.814	0.825	0.908	0.785	0.804	0.900	0.839
Spkr 5	JW27	F	0.795	0.796	0.878	0.774	0.733	0.893	0.811
Spkr 6	JW31	F	0.851	0.782	0.922	0.850	0.809	0.906	0.853
Spkr 7	JW40	M	0.779	0.551	0.906	0.749	0.833	0.869	0.781
Spkr 8	JW45	M	0.834	0.785	0.896	0.804	0.845	0.866	0.838
Spkr 9	JW54	F	0.758	0.529	0.879	0.760	0.884	0.848	0.776
Spkr 10	JW59	M	0.806	0.769	0.909	0.806	0.815	0.882	0.831

VI. CROSS-SPEAKER EXPERIMENTS ON THE XRMB DATASET

This section examines whether the acoustic and articulatory variability across speakers affects the performance of the speaker independent speech inversion, and if so, what is the impact on the performance. In order to explore the speaker variability, speaker dependent systems were trained on each of the ten speakers (five males and five females). The correlation results for the speaker dependent systems for the ten chosen speakers are shown in Table II. Comparing the numbers from Tables I and II, we observe that a speaker dependent speech inversion system is more accurate compared to a speaker independent system. However, the performance of the speaker dependent systems across speakers is mediocre. We tested each speaker dependent system using the test sets of the remaining nine speakers. Figure 9 shows the average correlation across the six TVs for the cross speaker tests performed on the speaker dependent systems. The mismatched speaker test correlations highlight the inter-speaker variability of the acoustic and articulatory spaces. As shown in Fig. 9, the cross-speaker performance of the speaker-dependent systems also showed a clear trend of gender dependence where matched gender trials had a better correlation by an absolute value of 0.2 than the mismatched gender trials.

We also evaluated the VTLN based speaker adaptation approach by applying the adaptation to cross-speaker trials. We evaluated each speaker dependent (SD) speech inversion system on the test sets of the other nine speakers. For each cross-speaker trial, we adapted the test speaker's evaluation data to match the target speakers acoustic space using the maximum likelihood based VTLN speaker adaptation procedure explained in Sec. V B. The median optimal VTLN warping factor for male target speakers was 0.925 (across all trials for the five males in the ten speaker subset). The median optimal warping factor for female target speakers was 1.125. For the matched gender trials, the median warping factor was 1.0, indicating that the VTLN adaptation is most effective in cross-gender trials.

We evaluated the speaker adapted MFCC features of the test speakers with the target speaker's inversion model. We computed the correlation between the actual and estimated TVs using the Pearson correlation. Figure 9 shows the average correlation across the six TVs for cross speaker trials before and after VTLN adaptation. We have plotted the average correlations for the mismatched gender and matched

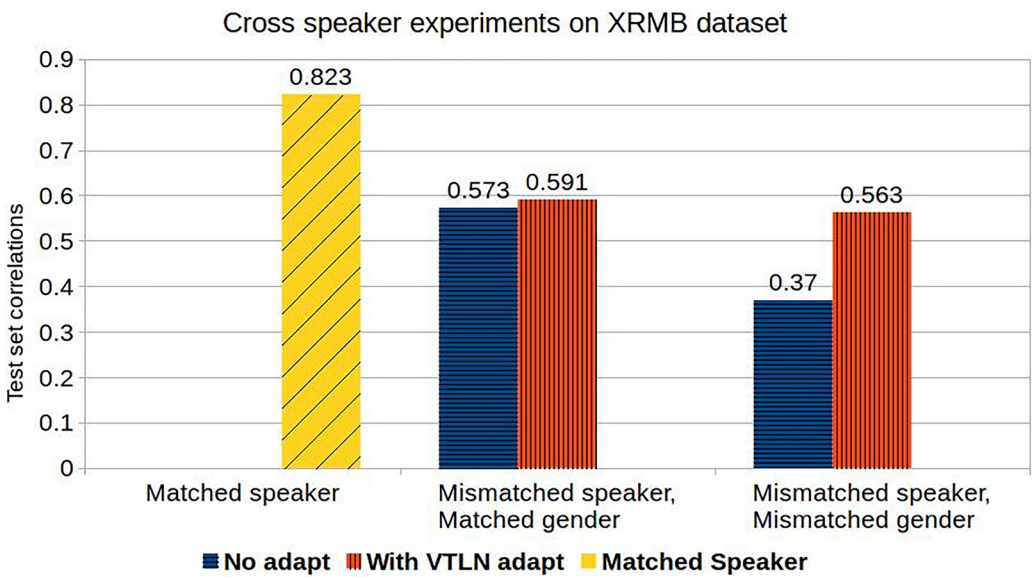


FIG. 9. (Color online) Average correlations of cross-speaker experiments on the XRMB dataset. Bars indicate average correlation across all trials (ten matched speaker trials and 90 mismatched speaker trials).

gender trials. We observe that, in the mismatched gender trials, the VTLN based speaker adaptation improves the average correlation by 52.16% relative to the case when no adaptation was performed, whereas the improvement due to VTLN adaptation in the matched gender trials is 3.14% relative to the case where no adaptation was performed.

Figure 10 shows a detailed visualization of the average correlations before and after VTLN adaptation for each trial in the cross-speaker experiment. The bar plots in Fig. 9 shows the average of the correlations from the matched and mismatched speaker trials shown in Fig. 10.

VII. LEAVE-ONE-SPEAKER-OUT EXPERIMENTS ON THE XRMB DATASET

We trained speech inversion systems using a single hidden layer feed-forward neural network. Since only small amounts of data were available for each speaker, single hidden layer networks were chosen as the architecture. The inputs to the neural network were the 13 dimensional MFCCs contextualized with MFCC features from eight frames on either side. Thus, the input dimension was $13 \times 17 = 221$. The outputs of the network were six dimensional TVs. We trained the neural networks with 300 nodes in the hidden layer. The number of nodes was chosen based on a pilot experiment conducted on a couple of speakers by varying the number of nodes from 100 to 500. The network with 300 nodes in the hidden layer performed best in terms of correlation. We did not extend the parameter sweep (of selecting optimal number of hidden layer nodes) to all ten speakers because that would greatly complicate the experiment. The outputs of the trained neural network were found to be noisy. The outputs were smoothed using a Kalman smoothing technique to obtain smooth TV estimates. Figure 3 shows the block diagram of our speech inversion system. Note that the speaker dependent experiments were performed with a shallow single hidden layer neural network and not a DNN, unlike what is shown in the block diagram in Fig. 3 for the speaker independent speech inversion system.

A. Speaker transformed datasets

Using the VTLN method described in Sec. V B, each speaker's data were transformed to each of the other nine speakers' data. Thus, for each speaker, we have ten sets of data—one set which is the original data for the speaker, and another nine sets obtained by transforming the remaining nine speakers' acoustic features to the target speaker using VTLN. In this way, we created 90 transformed datasets tailored to each of the ten speakers' acoustic spaces.

B. Speech inversion systems trained on speaker transformed datasets

We trained four types of speech inversion systems for each speaker as described in Sec. III. The following are the descriptions of the different inversion systems trained.

- SD: 10 SD speech inversion systems.
- *Sys1*: For each speaker, data from the other nine speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, ten such systems were trained. For example, for speaker "a," data from S_b,…,S_j was randomly sampled to match the amount of data in S_a.
- *Sys2*: For each speaker, VTLN transformed data from the other nine speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, ten such systems were trained. For example, for speaker "a," data from S_{ba},…,S_{ja} was randomly sampled to match the amount of data in S_a.
- *Sys3*: For each speaker, data from the target speaker and the VTLN transformed data from the other nine speakers were randomly chosen to match the amount of data from the target speaker and an inversion system was trained. In total, ten such systems were trained. For example, for speaker "a," data from S_a, S_{ba},…,S_{ja} was randomly sampled to match the amount of data in S_a. The difference between System3 and System2 is that System3 has some of the target speaker's data in the training set.

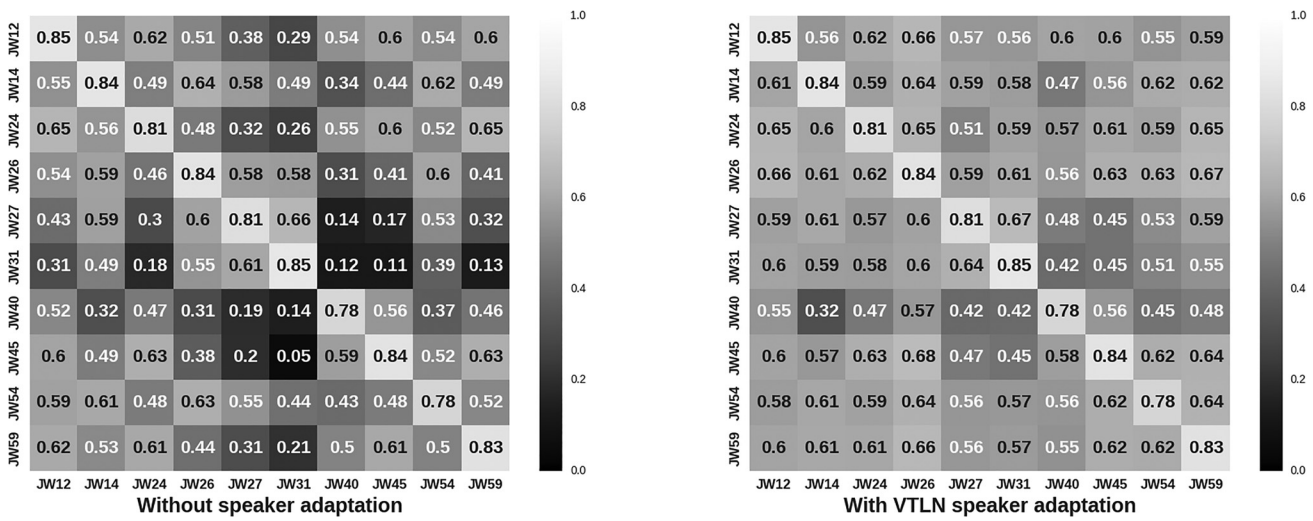


FIG. 10. Visualization of the cross speaker test correlations. Correlation of 1 corresponds to white and 0 corresponds to black.

TABLE III. Correlation results of SD, Sys1, Sys2, and Sys3 for all speakers.

Speech inversion system	Average amount of training data (min)	Spk 1 JW12 M	Spk 2 JW14 F	Spk 3 JW24 M	Spk 4 JW26 F	Spk 5 JW27 F	Spk 6 JW31 F	Spk 7 JW40 M	Spk 8 JW45 M	Spk 9 JW54 F	Spk 10 JW59 M	Average
SD	5.68	0.848	0.843	0.811	0.839	0.811	0.853	0.781	0.838	0.776	0.831	0.823
Sys1	5.68	0.669	0.659	0.608	0.631	0.556	0.507	0.560	0.615	0.635	0.642	0.608
Sys2	5.68	0.714	0.656	0.630	0.708	0.627	0.635	0.648	0.668	0.656	0.697	0.664
Sys3	5.68	0.738	0.699	0.715	0.738	0.660	0.708	0.583	0.685	0.687	0.717	0.693

In total, 40 speech inversion systems were trained. In the above described systems, the amount of training data for each system was kept the same in order to have a fair comparison with the SD system. However, the transformed data available for each target speaker was about ten times more because of the other nine speakers' data put together. We created versions of Systems 1, 2, and 3 using all the transformed data. We call these systems *Sys1_alldata*, *Sys2_alldata*, and *Sys3_alldata*.

C. Results of leave-one-speaker-out experiments

For each speaker, a test set containing 10% of the speaker's data was created and which was kept separate from all the speech inversion training and VTLN procedure. Each of the systems SD, System1, 2, and 3, were evaluated on each speaker's test set. The PPMC was computed between the actual and estimated TVs. Table III shows the correlation results of all the speech inversion systems across all speakers. The numbers show correlation values averaged across all six TVs. The correlation for LP tract variable is the least and that for TBCL is the highest. The performance of *Sys1* is very poor compared to SD because the training dataset for this system consists of a small number of utterances from multiple speakers. Transforming the data from the other nine speakers to the target speaker's acoustic space using the proposed VTLN approach provides an average of 9.2% relative improvement in correlation over *Sys1*. The amount of improvement in correlation varies across all speakers. Some speakers like JW14 and JW24 show marginal or no improvement in the performance, whereas for JW31, we see a large 25% relative improvement. In order to see the influence of speaker specific training data on the performance, we created *Sys3*, which contained a part of the target speaker's training set data. The overall amount of training data for *Sys3* was kept the same as the amount of training data available for each target speaker. This provided an average of 4.4% improvement in correlation compared to *Sys2*. However, the correlations of *Sys3* were still relatively 15% below the average correlation of the SD systems.

Table IV shows the correlation results for the speech inversion systems trained with all the available data from the other nine speakers. These are the systems *Sys1_alldata*, *Sys2_alldata*, and *Sys3_alldata* as described in Sec. VII. We observe that the results are much better than those in Table III. The performance gain obtained by performing the VTLN adaptation is around 5.9% relative to the correlation results of *Sys1_alldata*. It is interesting to observe that adding all the training data of the target speaker, as done in the training of *Sys3_alldata* provides a system that performs nearly as well as the speaker dependent SD systems. This demonstrates that adding VTLN adapted data from multiple speakers does not degrade the performance relative to the speaker dependent systems.

Based on the results shown in Tables III and IV, we can conclude that the amount of training data plays a great role in the accuracy of the speech inversion system. Even if the data is from multiple speakers, more data is always good. The VTLN speaker adaptation normalizes multiple speakers' acoustic data to match a target speaker. VTLN provides an average of 9.2% relative improvement of correlation (Sys1 to Sys2) on the speech inversion system trained on the 9 speakers' dataset. Adding a small amount of the target speaker's data in the training set improves the correlation further by 4.4% over Sys2. In spite of performing VTLN, the correlation performance of Sys2 trained on the transformed data is 16% poorer relative to the SD systems. The systems trained with all data show that having more training data from multiple speakers can make the speech inversion system better. The accuracy of Sys1_alldata is 15.9% relatively poorer than SD due to the mismatch between the acoustic spaces of the training speakers and the test speakers. With the VTLN based transformation of the training data, the accuracy improves by 5.9% relative to Sys1_alldata. This means our proposed adaptation technique helps reduce the mismatch between the acoustic spaces. Adding all of the target speakers' training data along with the transformed data of the other nine speakers' is almost as good as the speaker

TABLE IV. Correlation results of SD, Sys1_alldata, Sys2_alldata, and Sys3_alldata for all speakers. The row in boldface corresponding to Sys2_alldata is the speaker independent scenario where no speech data is available for the target speaker.

Speech inversion system	Average amount of training data (min)	Spk 1 JW12 M	Spk 2 JW14 F	Spk 3 JW24 M	Spk 4 JW26 F	Spk 5 JW27 F	Spk 6 JW31 F	Spk 7 JW40 M	Spk 8 JW45 M	Spk 9 JW54 F	Spk 10 JW59 M	Average
SD	5.68	0.848	0.843	0.811	0.839	0.811	0.853	0.781	0.838	0.776	0.831	0.823
Sys1_alldata	51.13	0.712	0.731	0.703	0.716	0.676	0.611	0.652	0.706	0.691	0.718	0.692
Sys2_alldata	**51.13**	**0.755**	**0.748**	**0.736**	**0.773**	**0.710**	**0.698**	**0.709**	**0.730**	**0.714**	**0.753**	0.733
Sys3_alldata	56.81	0.819	0.803	0.793	0.830	0.776	0.809	0.790	0.806	0.782	0.817	0.802

TABLE V. Cross corpus evaluation of systems trained on EMA-IEEE and tested on MOCHA database. Numbers show average correlations for cross corpus tests with and without VTLN adaptation.

Test →	fsew0			msak0		
Train ↓	No adapt	With adapt	% Change	No adapt	With adapt	% Change
F01	0.563	0.559	−0.71%	0.446	0.526	18.03%
F02	0.519	0.528	1.80%	0.355	0.434	22.31%
F03	0.469	0.496	5.88%	0.280	0.413	47.57%
F04	0.540	0.544	0.80%	0.420	0.533	27.04%
M01	0.464	0.477	2.69%	0.463	0.503	8.66%
M02	0.414	0.468	13.08%	0.509	0.492	−3.36%
M03	0.531	0.494	−6.98%	0.440	0.504	14.42%
M04	0.409	0.430	5.12%	0.515	0.524	1.90%

dependent performance. The performance of Sys3_alldata is 2.5% lower relative to the SD systems.

The XRMB dataset that we use in our experiments contains 46 speakers. Instead of limiting our speaker adaptation experiments to just the ten speaker subset, we extended the experiment outlined in Sec. IV by performing VTLN based speaker adaptation. In our experiments in Sec. IV, we had used five speakers in the test set of the speaker independent speech inversion, while the train and development splits contained 36 and 5 speakers, respectively. We speaker adapted the 41 speakers in the train and development set to each of the five test set speakers using the VTLN adaptation as outlined in Sec. V B. We then trained a 5-hidden-layer DNN for each of the speaker adapted set of the training data. Each test speaker was then evaluated using the corresponding DNN trained on the speaker adapted training data. We obtained an average correlation of 0.791 on the test set. Note that this result is comparable to the result shown in Table I. We observe that the VTLN adaptation provides a 1.15% relative improvement in correlation compared to the results without speaker adaptation. Perhaps the model trained on 4 h of data containing a diverse set of 36 speakers makes the model robust to speaker variation, and hence the VTLN adaptation only provides a small improvement.

VIII. CROSS CORPUS EXPERIMENTS

In order to evaluate the efficacy of the VTLN speaker adaptation in a cross corpus setting, we performed cross corpus experiments using the EMA-IEEE and MOCHA TIMIT datasets described in Sec. II. The MOCHA-TIMIT dataset contains two British English speakers: *"fseq0,"* and *"msak0."* The EMA-IEEE dataset contains eight American English speakers (four males *"M01–M04"* and four females *"F01–F04"*). We used only the normal speaking rate utterances from the EMA-IEEE dataset for our experiments. The audio in both the datasets were recorded at the same sampling rate of 16 kHz. Although the datasets were collected with different instruments by different researchers, the articulatory data were converted to nine TVs as outlined in Sec. II B 1. Since both datasets were transformed to TVs using the same procedure, they are appropriate for performing cross corpus experiments.

For each speaker in both datasets, we trained 64 component GMMs using the same procedure described in Sec. V. An SD speech inversion system was trained for each speaker in both datasets. The architecture of the speech inversion system was the same as described in Sec. III. For this experiment, we fixed the architecture of the neural networks as 5 hidden layer feedforward networks with 512 nodes in each layer. The same network architecture was used for all the speaker dependent speech inversion systems. For all the experiments, 80% of the speaker's data was used for training, while 10% each was used for cross validation and testing.

The SD systems from the MOCHA database were evaluated on the test utterances from the speakers in the EMA-IEEE dataset and vice versa. VTLN based speaker adaptation as described in Sec. V B was performed on each speaker to adapt towards the target speaker in the other database. The adaptation of a test speaker was performed with only the acoustic features from the speaker's test set. Since the adaptation is unsupervised, it does not require the ground truth TVs for adaptation. The cross corpus evaluations were performed again on the speaker adapted features. Pearson correlation between the actual and estimated TVs was used as the evaluation metric.

Tables V and VI show the results of the cross corpus evaluation. The correlations are considerably low compared to the matched corpus speaker independent systems. This is expected because of the mismatched speakers, accent, and corpus. We observed that the VTLN adaptation significantly improved performance for some speaker pairs like *F03-msak0* in Table V, where it improved the correlation by 47.57%, and *fsew0-M02* in Table VI, where the correlation improved by 49.85%. The performance across gender (average correlation = 0.438) without adaptation was worse than the performance for matched gender cases (average correlation = 0.525). The average improvement in correlation after speaker adaptation was 15.0% for mismatched gender test, while it was just 1.31% for matched gender test.

TABLE VI. Cross corpus evaluation of systems trained on MOCHA and tested on EMA-IEEE database. Numbers show average correlations for cross corpus tests with and without VTLN adaptation.

Train ↓ | Test →		F01	F02	F03	F04	M01	M02	M03	M04
fsew0	No adapt	0.571	0.491	0.516	0.612	0.465	0.386	0.475	0.422
	With adapt	0.556	0.490	0.481	0.612	0.514	0.578	0.505	0.455
	% Change	−2.61%	−0.15%	−6.79%	0.00%	10.44%	49.85%	6.27%	7.76%
msak0	No adapt	0.524	0.439	0.390	0.594	0.542	0.580	0.550	0.515
	With adapt	0.524	0.463	0.483	0.639	0.538	0.601	0.560	0.495
	% Change	0.00%	5.48%	23.78%	7.54%	−0.57%	3.60%	1.87%	−3.76%

IX. DISCUSSION & CONCLUSION

In this paper, we proposed a VTLN based speaker adaptation technique for acoustic-to-articulatory-speech inversion. We also performed several speech inversion experiments on three different datasets.

We first developed a DNN based speaker independent speech inversion system (Sec. IV) on the multi-speaker XRMB dataset. We explored different acoustic feature representations (MFCC, PLP, MELFB) and also tuned the network parameters of the DNN. We found a 5-hidden-layer DNN with 512 nodes in each layer, trained with contextualized MFCC features (17 frame splicing) as input was best suited for the speaker independent speech inversion system. We obtained a correlation of 0.782 (Table I) on a held out set of five speakers from the XRMB dataset. For the rest of our experiments in the paper, we fixed the speech inversion system architecture (as shown in Fig. 3, except the number of hidden layers in the neural network) and the input feature as contextualized MFCCs.

We then performed cross-speaker experiments on a randomly selected subset of ten speakers from the XRMB dataset (Sec. VI). We trained speaker dependent speech inversion systems for the ten selected speakers and performed mismatched speaker experiments across the ten speakers. The results showed that the VTLN based speaker adaptation improves the average correlation by 52.16% relative to the case when no adaptation was performed in mismatched gender trials, whereas the improvement due to VTLN adaptation in the matched gender trials was 3.14% (Fig. 9).

In Sec. VII, we performed leave-one-speaker-out experiments on the random subset of ten speakers selected for the cross-speaker experiments. Using the VTLN adaptation, we created speaker transformed datasets for each of the ten speakers by converting the remaining nine speaker's data using VTLN adaptation. We then performed a series of leave-one-speaker-out experiments using the speaker transformed datasets. We found that the VTLN provides an average of 9.2% relative improvement in correlation in the leave-one-speaker-out experiments. We also observed that the amount of training data plays a great role in the accuracy of the speech inversion system.

Finally, we performed cross-corpus speech inversion experiments (Sec. VIII) across the MOCHA-TIMIT and EMA-IEEE datasets. The results of the cross corpus speech inversion experiment highlight the gender dependence of the speech inversion system. We observe that the average correlation on mismatched gender (and in this case, mismatched corpus and accent) test was 16.6% poorer relative to the matched gender (but mismatched corpus and accent) test. After performing our VTLN based adaptation, the relative gap in performance between the mismatched and matched gender tests reduced to 6.04%. It is also interesting to note that the VTLN adaptation had a very small effect on the matched gender performance. This is probably because the unsupervised GMM speaker acoustic spaces overlap well within each gender category. In other words, the speaker acoustic spaces of male speakers overlap with each other and the female acoustic spaces overlap with each other. The male and female acoustic spaces are more separated due to the difference between the male and female vocal tract lengths. Thus, the VTLN adaptation makes more impact on the mismatched gender test than the matched gender case.

In this paper, we have examined the variability of speech and articulations across speakers and developed a speaker adaptation approach to normalize the speaker differences. The experiments in this paper show that data from multiple speakers can be normalized and combined to create better speaker independent speech inversion systems. This approach can be extended to combine data from different articulatory datasets to create a single improved speech inversion system. The VTLN approach of transforming the training data from multiple speakers to create multiple speaker adapted versions can be used as data augmentation for training speech inversion systems. In the future, we plan to explore this method to augment the limited amount of articulatory data available to train bigger, and more accurate speech inversion systems. The cross corpus experiment offers mismatch in corpus as well as accents. In the future, we plan to perform further experiments to isolate the effects of the accent mismatch. We also plan to further explore the VTLN based unsupervised adaptation to improve the performance of matched gender scenario. In all experiments, we have assumed that the TV representation is approximately speaker invariant. We plan to study this assumption and estimate the cross speaker variability of TVs for matching utterances. Exploring methods to perform speaker adaptation in the TV domain is part of our ongoing research. We believe that a combination of acoustic and articulatory speaker normalization would further improve the performance of speaker independent speech inversion systems.

ACKNOWLEDGMENT

We thank the National Science Foundation for supporting this research with the Grant Nos. IIS1162525, IIS1162046, and BCS1436600.

Afshan, A., and Ghosh, P. K. (**2015**). "Improved subject-independent acoustic-to-articulatory inversion," Speech Commun. **66**, 1–16.

Aryal, S., and Gutierrez-Osuna, R. (**2014**). "Accent conversion through cross-speaker articulatory synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 4–9, Florence, Italy, pp. 7694–7698.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (**1978**). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am. **63**(5), 1535–1555.

Bengio, Y., and Lecun, Y. (**2007**). "Scaling learning algorithms towards AI," in *Large-Scale Kernel Machines*, edited by J. W. L. Bottou, O. Chapelle, and D. DeCoste (MIT Press, Cambridge, MA), pp. 321–360.

Browman, C. P., and Goldstein, L. (**1992**). "Articulatory Phonology: An Overview *," Phonetica **49**, 155–180.

Cavin, M. (**2015**). "The use of ultrasound biofeedback for improving English/r," Working Papers Ling. Circle **25**(1), 32–41.

Eide, E., and Gish, H. (**1996**). "A parametric approach to vocal tract length normalization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, May 9, Atlanta, GA, pp. 346–348.

Ghosh, P. K., and Narayanan, S. (**2010**). "A generalized smoothness criterion for acoustic-to-articulatory inversion," J. Acoust. Soc. Am. **128**(4), 2162–2172.

Ghosh, P. K., and Narayanan, S. S. (**2011**). "A subject-independent acoustic-to-articulatory inversion," in *Proceedings of the IEEE International Conference*

on Acoustics, Speech and Signal Processing, May 22–27, Prague, Czech Republic, pp. 4624–4627.
Helfer, B. S., Quatieri, T. F., Williamson, J. R., Mehta, D. D., Horwitz, R., and Yu, B. (**2013**). "Classification of depression state based on articulatory precision," in *Proceedings of INTERSPEECH*, August 25–29, Lyon France.
Hiroya, S., and Honda, M. (**2004**). "Estimation of Articulatory Movements From Speech Acoustics Using an HMM-Based Speech Production Model," IEEE Trans. Speech Audio Process. **12**(2), 175–185.
Hogden, J. (**1996**). "Improving on Hidden Markov Models: An articulatory constrained, maximum likelihood approach to speech recognition and speech coding," Technical Report, Los Alamos National Laboratory, Los Alamos, NM.
Hueber, T., Girin, L., Alameda-Pineda, X., and Bailly, G. (**2015**). "Speaker-adaptive acoustic-articulatory inversion using cascaded Gaussian mixture regression," IEEE Trans. Audio Speech Lang. Process. **23**(12), 2246–2259.
Ji, A. (**2014**). "Speaker independent acoustic-to-articulatory," Ph.D. thesis, Marquette University, Milwaukee, WI.
King, S., and Taylor, P. (**2000**). "Detection of phonological features in continuous speech using neural networks," Comput. Speech Lang. **14**(4), 333–353.
Kirchhoff, K. (**1999**). "Robust speech recognition using articulatory information," Bielefeld University, Bielefeld, Germany.
Kirchhoff, K., Fink, G. A., and Sagerer, G. (**2002**). "Combining acoustic and articulatory feature information for robust speech recognition," Speech Commun. **37**(3-4), 303–319.
Krstulović, S. (**2001**). "Speech analysis with production constraints," Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, France.
Laprie, Y., and Mathieu, B. (**1998**). "A variational approach for estimating vocal tract shapes from the/nspeech signal," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98*, May 15, Seattle, WA, pp. 929–932.
McGowan, R. S. (**1994**). "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," Speech Commun. **14**(1), 19–48.
Mitra, V. (**2010**). "Articulatory information for robust speech recognition," Ph.D. thesis, University of Maryland, College Park, MD.
Mitra, V., Nam, H., Espy-Wilson, C. Y., Saltzman, E., and Goldstein, L. (**2010**). "Retrieving tract variables from acoustics: A comparison of different machine learning strategies," IEEE J. Selected Topics Signal Process. **4**(6), 1027–1045.
Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (**2012**). "Recognizing articulatory gestures from speech for robust speech recognition," J. Acoust. Soc. Am. **131**(3), 2270–2287.
Mitra, V., Shriberg, E., McLaren, M., Kathol, A., Richey, C., Vergyri, D., and Graciarena, M. (**2014a**). "The SRI AVEC-2014 Evaluation System," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge—AVEC '14*, November 7, Orlando, FL, pp. 93–101.
Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C. Y., and Saltzman, E. (**2014b**). "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*, May 4–9, Florence, Italy, pp. 3017–3021.
Mokhtari, P., Clermont, F., and Tanaka, K. (**2000**). "Toward an acoustic-articulatory model of inter-speaker variability," in *Proceedings of the Sixth International Conference on Spoken Language Processing*, October 16–20, Beijing, China, pp. 158–161.
Mokhtari, P., Kitamura, T., Takemoto, H., and Honda, K. (**2007**). "Principal components of vocal-tract area functions and inversion of vowels by linear regression of Cepstrum coefficients," J. Phon. **35**(1), 20–39.
Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (**2004**). "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**(4), 1771–1776.
Ouni, S., and Laprie, Y. (**2005**). "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," J. Acoust. Soc. Am. **118**(1), 444–460.
Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., and Maas, E. (**2014**). "Ultrasound visual feedback treatment and practice variability for residual speech sound errors," J. Speech Lang. Hear. Res. **57**(6), 2102–2115.
Qin, C., and Carreira-Perpiñán, M. Á. (**2007**). "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," in *Proceedings of INTERSPEECH*, August 27–31, Antwerp, Belgium, pp. 74–77.
Richmond, K. (**2006**). "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, September 17–21, Pittsburg, PA, pp. 577–580.
Richmond, K., King, S., and Taylor, P. (**2003**). "Modelling the uncertainty in recovering articulation from acoustics," Comput. Speech Lang. **17**(2–3), 153–172.
Rothauser, E., Chapman, W., and Guttman, N. (**1969**). "IEEE Recommnded Pratice for Speech Quality Measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246.
Sadjadi, S. O., Slaney, M., and Heck, L. (**2013**). "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," IEEE Speech and Language Processing Technical Committee Newsletter, pp. 1–4.
Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecol. Psychol. **1**(4), 333–382.
Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (**1987**). "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," Brain Lang. **31**, 26–35.
Tiede, M., Espy-Wilson, C. Y., Goldenberg, D., Mitra, V., Nam, H., and Sivaraman, G. (**2017**). "Quantifying kinematic aspects of reduction in a contrasting rate production task," J. Acoust. Soc. Am. **141**(5), 3580–3580.
Toda, T., Black, A., and Tokuda, K. (**2004**). "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *Proceedings of ICSLP*, Jeju Island, Korea, pp. 1129–1132.
Westbury, J. (**1994**). X-ray Microbeam Speech Production Database User's Handbook, Version 1 (Univ. of Wisconsin, Madison).
Wrench, A. A. (**2000**). "A Multichannel Articulatory Database and its Application for Automatic Speech Recognition," in *Proceedings of the 5th Seminar of Speech Production*, May 1–4, Bavaria, Germany, pp. 305–308.
Young, S. J., Evenmann, G., Gales, M. J. F., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (**2009**). *The HTK Book (for version 3.4)*, 3rd ed. (Cambridge University Press, Cambridge, UK).
Zhan, P., Zhan, P., and Waibel, A. (**1997**). "Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition," CMU Comput. Sci. Tech. Rep. **87**, 97–148.