Parallel Reference Speaker Weighting for Kinematic-Independent Acoustic-to-Articulatory Inversion

An Ji, Michael T. Johnson, *Senior Member, IEEE*, and Jeffrey J. Berry

***Abstract*—Acoustic-to-articulatory inversion, the estimation of articulatory kinematics from an acoustic waveform, is a challenging but important problem. Accurate estimation of articulatory movements has the potential for significant impact on our understanding of speech production, on our capacity to assess and treat pathologies in a clinical setting, and on speech technologies such as computer aided pronunciation assessment and audio-video synthesis. However, because of the complex and speaker-specific relationship between articulation and acoustics, existing approaches for inversion do not generalize well across speakers. As acquiring speaker-specific kinematic data for training is not feasible in many practical applications, this remains an important and open problem. This paper proposes a novel approach to acoustic-to-articulatory inversion, Parallel Reference Speaker Weighting (PRSW), which requires no kinematic data for the target speaker and a small amount of acoustic adaptation data. PRSW hypothesizes that acoustic and kinematic similarities are correlated and uses speaker-adapted articulatory models derived from acoustically derived weights. The system was assessed using a 20-speaker data set of synchronous acoustic and Electromagnetic Articulography (EMA) kinematic data. Results demonstrate that by restricting the reference group to a subset consisting of speakers with strong individual speaker-dependent inversion performance, the PRSW method is able to attain kinematic-independent acoustic-to-articulatory inversion performance nearly matching that of the speaker-dependent model, with an average correlation of 0.62 versus 0.63. This indicates that given a sufficiently complete and appropriately selected reference speaker set for adaptation, it is possible to create effective articulatory models without kinematic training data.**

***Index Terms*—Acoustic-to-articulatory inversion, electromagnetic articulography.**

I. Introduction

HUMAN speech is generated through the simultaneous movement of multiple articulators, including the tongue, jaw and lips, controlled together through the speech production process. Accurate knowledge of articulatory movements represents high-level information about speech production that can be of great value to many speech technologies and clinical applications. However, the reversal of the speech production process to estimate articulatory movements from speech acoustics is a challenging task requiring speaker-specific models. This task, referred to as acoustic-to-articulatory inversion, can help us understand and model the underlying latent variables involved with speech production at a deeper level.

Example applications of articulatory inversion include the integration of articulatory information with acoustic features to improve the performance of automatic speech recognition systems [1], [2]. In addition, articulatory information can improve the quality of synthesized speech [3] or synthesized video, such as facial animation in films and video-games [4]. Visualizing the position of the articulators derived from acoustic signal is extremely useful in speech pathology assessment and treatment systems, as well as in Computer Aided Language Learning (CALL) and Computer Aided Pronunciation Training (CAPT) systems, where a reliable inverse mapping can enable more accurate pronunciation assessment and correspondingly more specific corrective feedback [5]–[9].

Current methods for acoustic-to-articulatory inversion are based on a variety of mapping techniques including codebook search, neural networks, Kalman filtering, GMMs, and HMMs [10]–[16]. Nearly all these methods use parallel articulatory and acoustic training data from a single subject to learn the mapping between acoustic and articulatory spaces, and then perform inversion on the acoustic data of the same subject. Variability in physical vocal tract configurations, as well as speaker-specific relationship between acoustics and articulation, cause the mapping from acoustic to articulatory space to vary significantly across subjects. Because of this, most existing approaches for inversion work do not generalize to new speakers without kinematic data.

However, the acquisition and measurement of kinematic data is much more problematic than acoustic data, with higher equipment costs and significantly greater invasiveness and inconvenience to speakers. As such, most practical applications of articulatory inversion would not allow for the collection of detailed kinematic data on an individual user basis. This would include, for example, automatic speech recognition methods that wish to benefit from articulatory-derived features, or accent modification applications for CALL that target identification of specific articulatory causes underlying pronunciation errors.

Manuscript received November 24, 2015; revised April 26, 2016 and June 22, 2016; accepted June 25, 2016. Date of publication July 07, 2016; date of current version August 02, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mohamed Afify.

A. Ji was with the Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI 53233 USA. She is now with Ford Motor Company, Dearborn, MI 48126 USA (e-mail: ji_an30@hotmail.com).

M. T. Johnson is with the Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI 53233 USA (e-mail: mike.johnson@marquette.edu).

J. J. Berry is with the Department of Speech Pathology and Audiology, Marquette University, Milwaukee, WI 53233 SA (e-mail: jeffrey.berry@marquette.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2588340

Recent research has begun to address this problem through what is sometimes termed speaker-independent inversion, which we refer to as kinematic-independent, defined as acoustic-to-articulatory inversion on a new speaker for whom there is no kinematic data and a relatively small amount of labelled acoustic adaptation data. Dusan and Deng have applied vocal tract length normalization to their speech inversion model [13], with some improvement. Hiroya and Honda have also introduced a speaker adaptation technique based on an HMM-based speech production model [17] which models a linear relationship between the speech spectrum and articulatory features in each state. In addition, Hueber *et al.* have combined voice conversion and acoustic-to-articulatory inversion into a single GMM-based mapping framework [18], with good initial results using two speakers' data.

In this paper, a robust kinematic-independent inversion method called Parallel Reference Speaker Weighting (PRSW) is proposed. Specifically, a reference speaker weighting (RSW) adaptation approach [19] is modified to a parallel structure in which synchronized speaker-dependent acoustic and articulatory models on a set of reference speakers are adapted in parallel. Under the hypothesis that acoustic and articulatory similarities are correlated, the PRSW method learns adaptation weights in the acoustic model space and then applies them in the articulatory model space. This creates a speaker-specific inversion mapping that can estimate articulatory trajectories for new speakers who have no kinematic training data.

The article is organized as follows: Section II introduces the HMM-based acoustic-to-articulatory inversion approach. Section III presents theoretical aspects of the proposed PRSW adaptation techniques. Experimental set up and practical implementation are described in Section IV. Results and discussion are given in Section V, with conclusion in Section VI.

II. Methods

A. HMM Based Inversion Framework

Due to the ill-posed nature of the inversion problem, it can be beneficial to connect the articulatory and acoustic domains through a phoneme or state level representation, instead of seeking a direct mapping. In this work, we use a synchronized Hidden Markov Model (HMM) framework, similar to that of Zhang and Renals [16], to tie the two domains at the level of individual phoneme states in a sequential model. The diagram of such a parallel acoustic-articulatory model is illustrated in Fig. 1. In this approach, two separate HMMs are built, one in the acoustic observation space and one in the articulatory observation space, with these models explicitly tied through state sequence synchronization. Parallel acoustic and articulatory data are used to train the acoustic and articulatory HMMs separately. Within each state, a GMM is used for modeling the statistical distribution of the feature vectors in each domain, although the approach is generalizable and could be extended to other state observation models such as deep neural networks [20]. The number of mixture components differs between the acoustic and articulatory model domains, because the acoustic features have a more complex distribution than the trajectory patterns

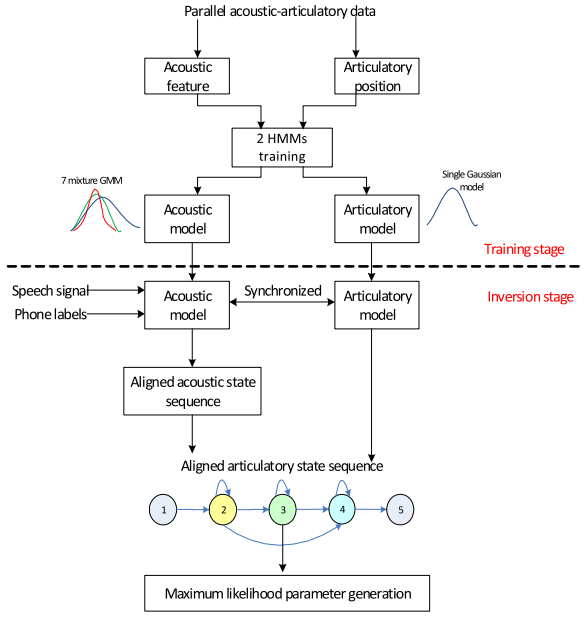


Fig. 1. Diagram of the HMM-based articulatory-to-acoustic inversion system.

of individual articulatory features [16]. In the present work the acoustic HMM uses a GMM with seven mixture components, the number of mixture components being empirically tuned to balance the amount of training data, with inversion results showing a slight decline for a higher number of mixture components after that point. In the inversion stage, the test speech signal is input to the acoustic HMM to compute an optimal HMM state sequence using the Viterbi algorithm, and the correspondingly aligned articulatory HMMs can be used to recover the articulatory trajectory. The articulatory HMM generates a smoothed position trajectory, using the articulatory means combined with a dynamic smooth window of the articulatory distribution as in [21], based on the maximum likelihood parameter generation algorithm.

1) Training: The acoustic and articulatory HMMs are trained separately. The acoustic HMM is trained first using the maximum likelihood Expectation Maximization algorithm, after which the trained acoustic models are used to calculate state level alignments to the frame-by-frame data. These alignments are then used with the articulatory data to directly calculate the articulatory HMM state means and variances.

2) Forced Alignment: In the inversion stage, the speech signal and phone labels are input into the acoustic HMM, and a state sequence is produced through forced alignment with the Viterbi algorithm. The articulatory states matching the corresponding acoustic states are concatenated into an articulatory state sequence.

3) Maximum Likelihood Parameter Generation Using Dynamic Features: Once the articulatory state alignment is generated, the recovery algorithm needs to estimate a smooth and slow changing articulatory trajectory from the HMM state sequence. As described in [21], the observation data sequence $\mathbf{O}$ is estimated by maximizing $P(\mathbf{O}|\mathbf{Q},\lambda)$ with respect to $\mathbf{O}$ for a fixed state sequence, $\mathbf{Q} = [\, q_1 \;\; q_2 \;\; \cdots \;\; q_T \,]$ where λ represents

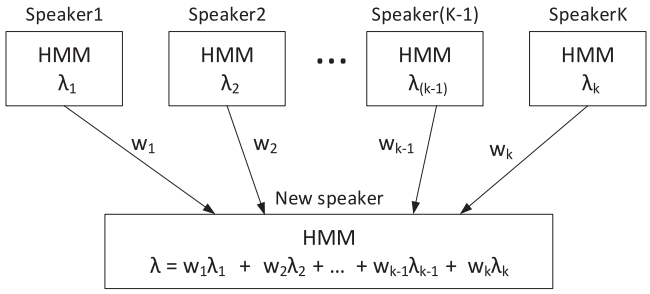


Fig. 2. Reference speaker weighting.

the model parameters. The result is that a smoothly varying ML estimate of the articulatory trajectory can be recovered from the Gaussian PDF state sequences via

$$\mathbf{C} = \mathbf{D}^{-1}\mathbf{\Sigma}\left(\mathbf{D}^{T}\right)^{-1}\mathbf{D}\mathbf{\Sigma}^{-1}\mathbf{U}^{T} \tag{1}$$

where $\mathbf{D}$ is a fixed static-to-dynamic computation matrix [21], $\mathbf{\Sigma}$ sigma is a block-diagonal matrix consisting of concatenated state covariance matrix, and $\mathbf{U}$ is a matrix of concatenated mean vectors.

B. Reference Speaker Weighting

The proposed approach is based on RSW adaptation [19]. RSW is a rapid speaker adaptation approach originally designed for small amounts of adaption data, typically 5-10 seconds of speech. The idea is based on model combination, and the small number of parameters allows this to work effectively even with limited adaptation data. Under the assumption that pronunciation patterns vary in similar ways across speakers, pronunciation variants of phonemes that are unseen in the adaptation data can still be correctly adjusted. Creating an RSW adaptation framework requires a set of individual speaker-dependent models across a reasonably diverse reference speaker set as a starting point for estimating the parameters of a new speaker.

The basic idea of this method is shown in Fig. 2. A new speaker's model can be estimated from a weighted combination of reference speakers. Each reference speaker is represented by a supervector, which is constructed by concatenating the mean vectors of all acoustic model parameters.

RSW estimates the model of a new speaker from the span of the K reference speaker models. Speaker-dependent models are trained using HMMs. Supervectors are used to represent the HMM model parameters. Using the reference speakers' supervectors, a set of maximum likelihood weights is estimated to match the new speaker's adaptation data by using the expectation maximization algorithm. The new speaker's model can then be constructed from a linear combination of reference speakers' model using these weights.

Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K\}$ be the set of reference speaker supervectors, defined as the concatenation of the Gaussian means from all state models in sequence. The RSW estimate of a new speaker's supervector is

$$\mathbf{s} \approx \mathbf{s}^{\mathrm{rsw}} = \sum_{k=1}^{K} \mathbf{w}_k \mathbf{y}_k = \mathbf{YW} \tag{2}$$

and the mean vector of the r th Gaussian is

$$\mu_r^{(\mathrm{rsw})} = \sum_{k=1}^{K} \mathbf{w}_k \mathbf{y}_{\mathrm{mr}} = \mathbf{Y}_r \mathbf{W} \tag{3}$$

where m is the mixture index, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K]^T$ is the weight vector and r is the number of Gaussian mixtures. Given adaptation data $\mathbf{O} = [\,\mathbf{o}_1\ \mathbf{o}_2\ \cdots\ \mathbf{o}_T\,]$, the Maximum Likelihood estimate of w can be found by solving a system of K linear equations,

$$\mathbf{w} = \left[\sum_{r=1}^{R}\left(\sum_{t=1}^{T}\gamma_t(r)\right)\mathbf{Y}_r^T\mathbf{C}_r^{-1}\mathbf{Y}_r\right]^{-1} \times \left[\sum_{r=1}^{R}\mathbf{Y}_r^T\mathbf{C}_r^{-1}\left(\sum_{t=1}^{T}\gamma_t(r)\,\mathbf{o}_t\right)\right] \tag{4}$$

where $\gamma_t(r)$ is the posterior probability of observing O_t in the r_{th} Gaussian, and C_r is the covariance matrix of the r th Gaussian.

RSW uses the model parameters of selected speakers to create a composite model for new unseen speakers. RSW is closely related to another fast speaker adaptation method, Eigenvoice, which uses principal component analysis to find a set of orthogonal basis vectors to create reference vectors. Both of these methods require the model of a new speaker to lie on the span of some reference vectors. In our acoustic-to-articulatory inversion application, RSW is chosen because we have one-to-one matched acoustic and articulatory models for individual speakers, which allows us to use the information from the acoustic space to adapt the model in articulatory space.

C. Parallel Reference Speaker Weighting

The proposed PRSW approach is based on the fundamental assumption that similarity between speakers in the acoustic domain is highly correlated to similarity between speakers in the articulatory domain. In the PRSW approach, a new speaker with no kinematic data is compared to a set of reference speakers in the acoustic domain, weights for these speakers are estimated using a maximum likelihood approach, and then weighted composite models are created from these weights in both the acoustic and articulatory domains.

In PRSW, the speaker combination that generates the new speaker in acoustic space is assumed to be consistent with those in the articulatory space. The new speaker's articulatory realization can be recovered from the reference speakers' articulatory model by using acoustically derived weights. In the inversion stage, identical weights are used in the articulatory space. Let $\mathbf{A} = [\,\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_K\ ]$ be the set of reference speaker articulatory super vectors. Then the RSW estimate of the new speaker's articulatory supervector is

$$\mathbf{A}_{\mathrm{unknown}} \approx \sum_{k=1}^{K} w_k \mathbf{a}_k = \mathbf{AW}, \tag{5}$$

where $\mathbf{W}$ is the same weight matrix derived from acoustics through equation (3). The new speaker's articulatory movements can be estimated from the adapted model by using the

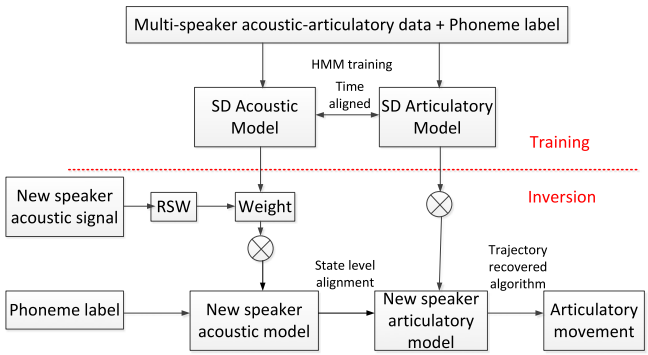


Fig. 3. Parallel RSW.

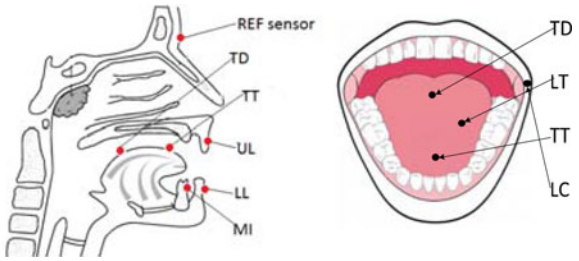


Fig. 4. EMA-MAE sensor placement.

maximum likelihood parameter generation algorithm described above. Fig. 3 illustrates this method for constructing an acoustic-articulator inversion model using the new PRSW approach.

To implement PRSW, parallel acoustic and articulatory HMMs are trained for each reference speaker. RSW is used to adapt a new acoustic model for the unknown speaker. For new speakers without articulatory data, weights are calculated from acoustic adaptation data and these weights are used to generate both acoustic and articulatory models.

III. Experimental Study

A. Data Set

The Marquette EMA-MAE corpus [22] includes synchronous acoustic and three-dimensional kinematic data collected via Electromagnetic Articulography (EMA) at 400 Hz for 40 speakers, 20 native English speakers and 20 native Mandarin speakers speaking English. In the present work we use only the 20 native English speakers' data. Acoustic records were obtained using a cardioid pattern directional condenser microphone positioned approximately 1 meter from participants. The corpus includes approximately 45 minutes (198 ± 5 utterances) of synchronized acoustic and kinematic data for each speaker, including word, sentence, and paragraph level speech samples. Each sensor records three-dimensional position as well as two-dimensional orientation representing orientation of the sensor's transverse plane. For each speaker, the data have been separated into training (90%) and test (10%) sets.

As shown in Fig. 4, articulatory sensors included the jaw (MI) (labial surface of lower front incisors), lower lip (LL), upper lip (UL), tongue dorsum (TD), and tongue tip (TT), all placed in the midsagittal plane. In addition, there were two lateral sensors, one (LC) at the left corner of the mouth to help indicate lip rounding and one (LT) in the left central midpoint of the tongue body to help indicate lateral tongue curvature.

Each subject's data includes a bite-plate record that records the position of two reference sensors fixed in a dental impression. The placement of these two sensors, in addition to the primary reference sensor at the subject's forehead, allows for a geometric bite-plate calibration that orients the sensor Cartesian coordinate system so that the x-y plane represents the midsagittal plane and the x-z plane represents the maxillary occlusal plane. In addition, a palate trace is recorded that maps each subject's hard palate.

TABLE I
Articulatory Features

	Description
VT1	Tongue dorsum normalized horizontal position
VT2	Tongue dorsum vertical height to hard palate
VT3	Tongue body normalized horizontal position
VT4	Tongue body vertical height to hard palate
VT5	Tongue apex normalized horizontal position
VT6	Tongue apex vertical height to hard palate
VT7	Normalized horizontal lip protrusion
VT8	Normalized vertical lip separation

B. Acoustic and Articulatory Features

Acoustic features include a standard set of 12 Mel frequency cepstral coefficients (MFCCs) plus energy, along with delta and delta-delta coefficients, giving a 39-dimensional acoustic feature vector. Hamming windowed frames of length 25 ms were used, with a step size of 10 ms. Articulatory features are calculated from EMA kinematic sensor data that are scale-normalized, with vertical features as differential distances to the hard palate data for the subject. These features are shown in Table I.

Articulatory feature normalization is implemented using the distance between the central incisors and the midsagittal point between the first molars from each speaker's bite plate record as a dividing constant for horizontal positions, giving relative rather than absolute information about the tongue's position relative to the vocal tract. The horizontal (x-axis) variables VT1, 3, 5, and 7, are all calculated directly from sensor position divided by this normalization constant. The vertical (y-axis) variables VT2, 4, and 6 are computed as the vertical distance between the sensor position and the palate, representing vocal tract height at the sensor positions, including two midsagittal positions and one lateral position. Palate height at the sensor's location is determined using a thin-plate spline representation [23] of the subject's palate trace record. Lip protrusion VT7 is taken directly from the sensor x position without any normalization, and vertical lip separation VT8 is calculated as

$$VT8 = \frac{(UL_y - LL_y) - (UL_y - LL_y)_{\text{closed position}}}{(UL_y - LL_y)_{\max}} \quad (6)$$

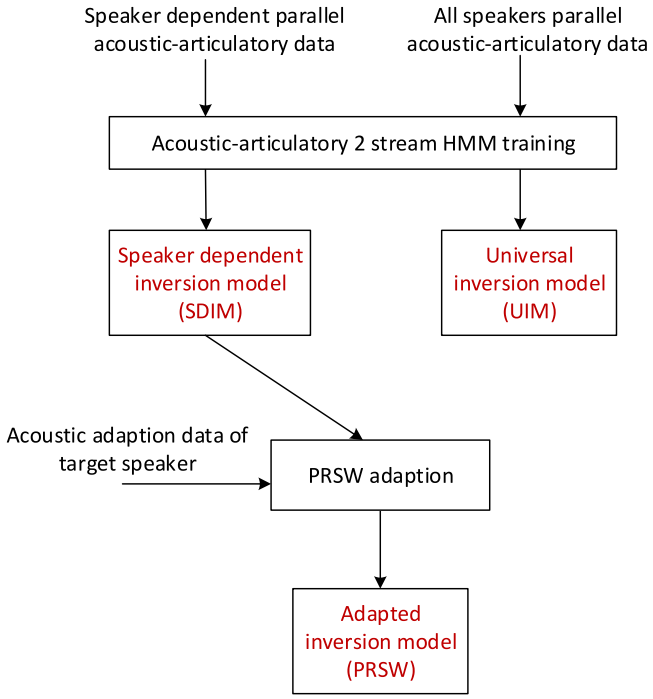


Fig. 5. Experimental paradigm.

which is lip separation scaled to a [0,1] working space.

C. Evaluation

Typical metrics for performance evaluation of acoustic-to-articulatory inversion include the root-mean-square (RMS) error as well as correlation between the actual and estimated articulatory position. However, for a kinematic-independent framework, two studies [18], [24] have shown that average RMS error is not suitable for evaluating the cross-speaker acoustic-to-articulatory inversion due to differences in scaling and dynamic range caused by a lack of kinematic data. Without articulatory data for the test speaker the estimated articulatory outputs represent the correct movement patterns but not necessarily the new speaker's articulatory mean and variance, which are impacted by both physiological differences and sensor placement differences across subjects.

Thus the correlation metric, which is a measure of overall similarity between the reference and the estimated trajectories, is used as the primary evaluation criterion for cross-speaker inversion results. Correlation is given directly by

$$r = \frac{\sum_{i=1}^{m}\left(f(x_i) - \overline{f(x_i)}\right)(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{m}\left(f(x_i) - \overline{f(x_i)}\right)^2}\sqrt{\sum_{i=1}^{m}(y_i - \overline{y_i})^2}} \tag{7}$$

where $\overline{f(x_i)}$ and $\overline{y_i}$ are the means of the estimated and actual articulatory values, respectively.

As a secondary evaluation metric, we use normalized RMS error to remove the mean bias of articulatory trajectories. Denoting the actual values of the articulatory measure as y and the corresponding values of the estimated output as $f(x)$, the normalized RMS error over the whole test set is calculated as:

$$E = \frac{\sqrt{\frac{1}{M}\sum_{i=1}^{m}(f(x_i) - y_i)^2}}{std(y)} \tag{8}$$

where m is the number of examples in the test set, y_i is the true articulatory variable value, $f(x_i)$ is the inversion output, and $std(y)$ is the standard deviation of the articulatory variable across the full test set. Note that this metric is still negatively impacted by any bias or dynamic range differences between the estimates and the true kinematic data.

A good articulatory inversion system is expected to obtain low RMS error and high correlation with respect to real articulatory data. In prior work, several different EMA datasets have been used across various different methodologies, which makes it difficult to compare results or have a strong frame of reference for expected performance. However, MOCHA-TIMIT has been the most widely used EMA dataset. The lowest RMS error reported is from Richmond's trajectory mixture density networks [12] which is 0.99 mm with correlation 0.79 on the MNGU0 speaker data. For the kinematic-independent inversion, Ghosh and Narayanan [24] achieved average correlation 0.4 with two speakers.

D. Experimental Scheme

Several experimental comparisons were conducted to evaluate the proposed PRSW method. An overview of the experimental paradigm is illustrated in Fig. 5. In building comparative models, individual speaker-dependent inversion models (SDIMs) were created for all 20 speakers, as well as a single universal inversion model (UIM) that was built from the combined data, each based on the method shown in Fig. 1. Following this, a new acoustically adapted and kinematic data independent model was created for each speaker, by considering the 19 remaining speakers as the reference speaker set, and using the PRSW approach to build a new inversion model from these, using the new speaker's acoustic adaptation data but no kinematic data.

The underlying goal is for the adapted PRSW model to achieve inversion performance significantly better than the UIM in which all speakers' parallel acoustic-articulatory data are used and approaching that of the speaker-dependent model in which every individual speaker's parallel acoustic-articulatory data are used. The universal model is not quite a true lower bound on performance, since training data included the test speaker's acoustic and articulatory data—it was preferred to have a single UIM rather than 20 different speaker-independent models each with one speaker removed, since this has only a small impact on performance and is much simpler to implement.

Experimental scenarios include the following:

1) Baseline experiment comparing the SDIM, UIM and PRSW inversion performance across 20 native English speakers.
2) A reduced reference speaker implementation, using several speaker selection methods. (Using the baseline result findings that the quality of the inversion models of the reference speakers impacted adaptation performance, as shown in the next section.)

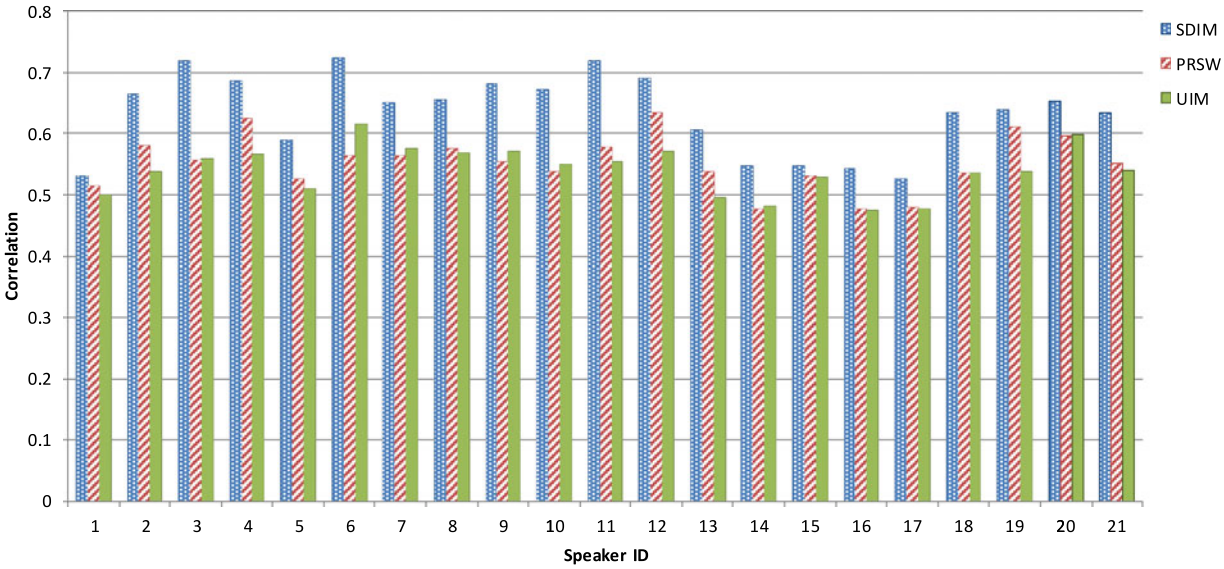


Fig. 6. Baseline correlation results. Mean correlations are SDIM = 0.63, PRSW = 0.55, UIM = 0.54.

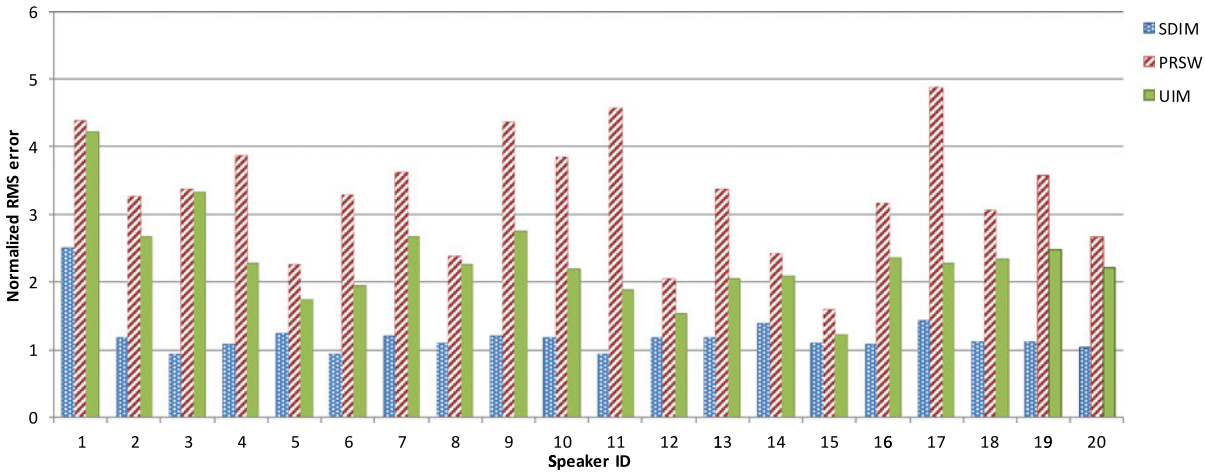


Fig. 7. Baseline RMS results. Mean normalized RMS error are SDIM = 1.19, PRSW = 3.26, UIM = 2.3.

3) An evaluation of the relationship between amount of acoustic adaptation data and inversion performance.

IV. Results and Analysis

A. Baseline Adaption Result

Fig. 6 shows the inversion performance for all 20 speakers as measured by correlation of the estimated and actual articulatory trajectories. From the correlation results we see that 13 out of 20 speakers support the initial hypothesis (SDIM > PRSW > UIM); however, seven speakers have results that show a different pattern (SDIM > UIM > PRSW), with the PRSW method giving relatively poor results. Looking closely at the correlation, the inversion performance of the speaker-dependent models varies widely across the 20 speakers (from highest 0.72 to lowest 0.52). The universal model has a relatively consistent inversion performance for every individual speaker (around 0.54).

Fig. 7 shows the average normalized RMS error for each speaker, and Fig. 8 illustrates an example of these reconstruction results, showing the true trajectory (blue line) along with inversion output from SDIM (red line), UIM (black line), and the PRSW model (green line) for the articulatory feature VT8. Although PRSW matches the overall trajectory shape well, there is an offset and a change in dynamic range that causes a larger RMS error, which illustrates why normalized RMS error is not considered the best measure for evaluating kinematic-independent inversion systems. This differential is caused by physical variation in subjects that cannot be estimated or compensated for without kinematic data, even though the trajectory shape can be accurately estimated. In fact, the offset and dynamic range adjustment act as a beneficial normalization in many applications that acts to reduce speaker variance while still accurately tracking articulatory patterns.

B. Variation Across Speakers

There is a large variation in the baseline speaker-dependent inversion performance across the 20 speakers. This variation can be further investigated by analyzing the articulatory

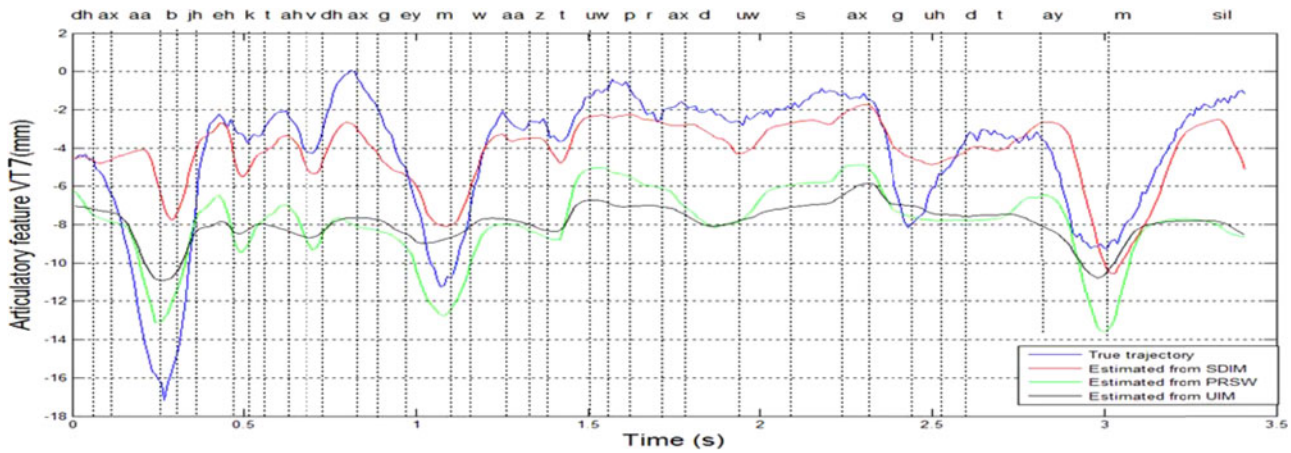


Fig. 8. Example of reconstructed articulatory features from the three different models.

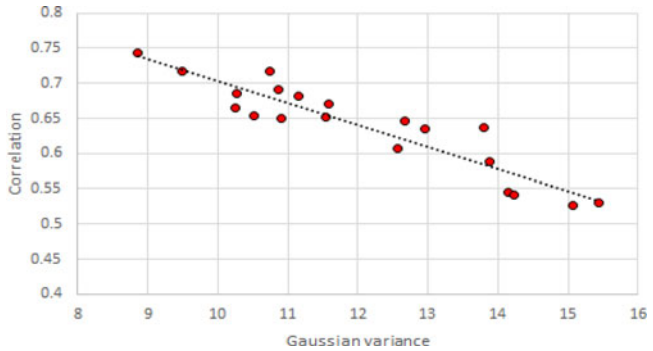


Fig. 9. Scatter plot of articulatory model variance versus correlation of speaker-dependent models.

feature model parameters for each speaker. The mapping from acoustic-to-articulatory space is through state alignment, so the more consistent the articulatory feature values are for identical phoneme sequences, the better the expected performance of the inversion system. The Gaussian variance in the articulatory HMM states is a good measure of this consistency.

The scatter plot in Fig. 9 shows a linear relationship between the consistency of articulatory features and the inversion performance as measured by correlation. In this figure, each red dot represents an individual speaker. A higher variance indicates that the speaker has a less consistent articulatory pattern, which is correlated with the inversion model having less accurate estimates of articulatory feature patterns. Speakers with lower variance articulatory models have better performing inversion models.

These results highlight a concern with the proposed PRSW method, which is that not all reference speakers give high-quality speaker-dependent acoustic-to-articulatory inversion models. This may be a result of less consistent articulatory or pronunciation patterns or other similar factors, but suggest the idea that a reduced set of reference speakers that includes only those with good individual inversion performance would lead to a more robust and generalizable kinematic-independent system. In principal, a large set of speakers with parallel acoustic and kinematic data, across a diversity of speaking and dialect patterns, could be used to create an initial set of reference models. After this, a subset of speakers, still representing a diversity of speaking patterns but having high-quality inversion performance, could be used to create the PRSW reference set for new speakers.

In the next section, two different reference speaker selection strategies will be explored: one based on limiting the total number of reference speakers based on acoustic similarity (weight thresholding) and the other based on globally limiting the reference speaker set based on speaker-dependent inversion performance (M-best pre-selection). Both of the adaptation approaches use each speaker's full set of data if enrolled as reference speakers.

C. Selection of Reference Speakers

Normally, the quality of an adapted acoustic model is dependent on the selection of reference speakers. The influence of selection approaches has been investigated in previous studies for acoustic models [25], [26] but not for articulatory models. In this section, two different reference selection strategies for the proposed acoustic-to-articulatory inversion system have been implemented and analyzed.

1) Weight Thresholding: In the proposed weight thresholding approach for reference speaker selection, the full set of reference speakers is used with the PRSW approach to identify maximum likelihood speaker weights, and then the speaker weights are thresholded against a fixed value α. Only speakers exceeding the minimum weight level are included for building the target speaker inversion model, with weights re-normalized to sum to unity. Since RSW weights can be regarded as a similarity measurement, this speaker selection can be viewed as a nearest neighbor implementation requiring a minimum similarity for inclusion. Note that this method does not eliminate speakers based on their individual speaker-dependent inversion performance, but rather based on minimum target speaker similarity.

In order to investigate the effect of different thresholds, the threshold α is incremented in small steps (0.01), with a maximum value of 0.09, the maximum threshold for this data

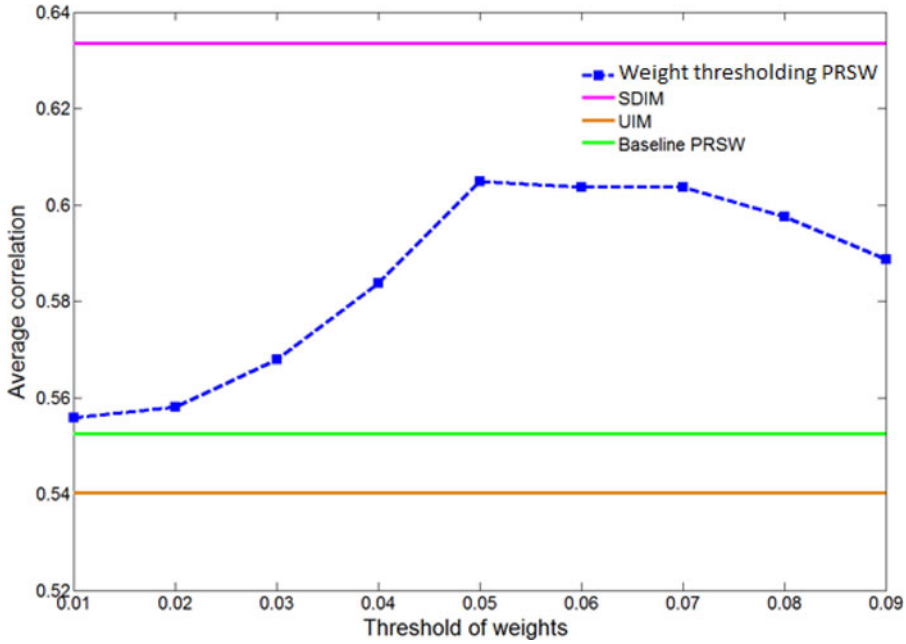


Fig. 10. Correlation as a function of threshold, for PRSW using weight thresholding selection.

set that ensures there is at least one speaker in the reference speaker set.

Fig. 10 shows the plot of threshold as average performance across 20 speakers. Results show that the average performance for this approach is always higher than the baseline PRSW results. With the initial threshold of 0.01 the performance is close to that of the baseline PRSW. As the threshold increases, the performance continues to improve until a threshold of 0.05, and then decreases slowly. The high performance suggests that reducing the number of speaker models being combined to create the new test speaker has a positive overall impact on articulatory consistency. Although in this case a threshold of 0.05 is the best, the optimal weight threshold would vary as a function of both the number of references speakers in a particular data set and their relative similarity to the target speaker. Once a specific reference speaker set is established, the optimal weight threshold can be determined on a set of development data and should then give consistent results for new speakers.

2) M-Best Global Pre-Selection: In the proposed global pre-selection approach, the M speakers with the best speaker-dependent inversion performance are selected globally as reference speakers, with the other speakers eliminated from consideration. Because of the observed large variance across speakers in terms of the quality of speaker-dependent results, speaker-dependent inversion performance can be regarded as a measure of model consistency. The hypothesis is that the more consistent the reference speakers, the higher the upper limit on inversion results of the adapted model.

In this global pre-selection method, the core reference speaker set is the same for each test speaker, including exactly the M-best reference speakers according to speaker-dependent model correlation performance. When the test speaker is in the M best list, the next best speaker is included instead, so that the reference set is maintained at M consistently across all 20 speakers. This means that the reference speaker sets are not fully identical, but always have at least 19 speakers in common. In this experiment, M is increased from 1 to 19.

Fig. 11 shows the plot of the average performance as a function of M across 20 speakers. With the weight thresholding method, the overall performance dominates that of the baseline PRSW regardless of reference speaker set. In the initial case $M = 1$, a single reference speaker acts as a surrogate model for the target speaker. As the number of reference speakers increases, the average performance increases until reaching a peak at $M = 7$, then decreases significantly. For this dataset, $M = 7$ results in speakers having an SDIM correlation greater than 0.67 being selected as reference speakers. As with the weight thresholding approach, the optimal parameter M is also a function of the original number of speakers, and more importantly of the quality of those speaker models as measured by the speaker-dependent inversion performance.

Table II shows full results for all 20 speakers for SDIM, UIM, baseline, and both thresholding approaches, over this same test dataset. Although there is a large variation in performance across individual speakers, indicating the complexity of acoustic-to-articulatory inversion in general, results show that reducing the reference speaker set improves performance in all 20 cases, and results in a final PRSW system that outperforms the UIM system for every test speaker and approaches the accuracy of the speaker-dependent systems.

Qualitative examination of the individual reference speakers selected through the acoustic and global selection methods reveals similar speaker selection results. The accuracy of the adapted model depends both on the similarity in the acoustic space and on the consistency of reference speakers' articulatory patterns, but the latter is especially important. Together these

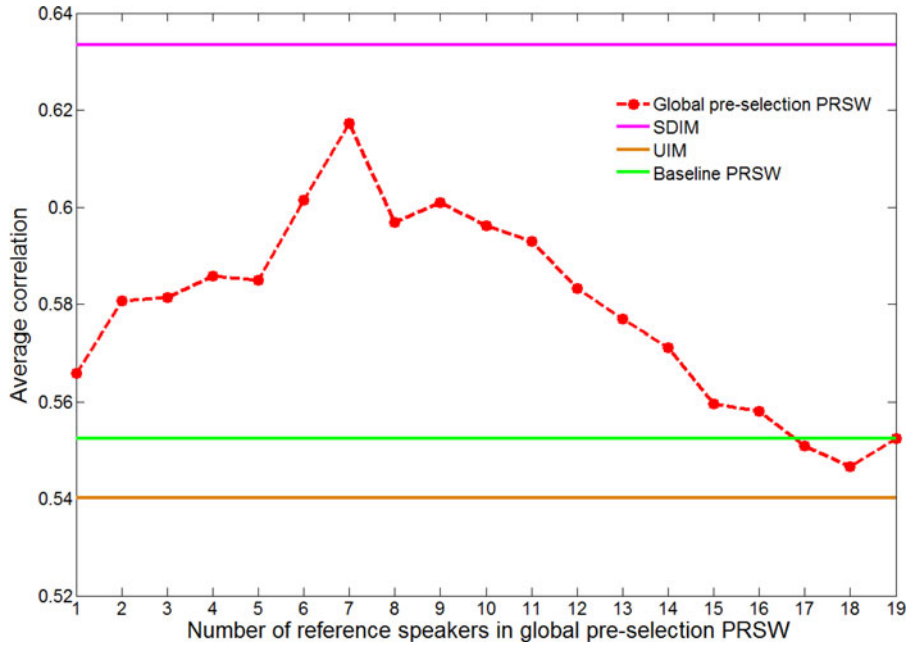


Fig. 11. Correlation as a function of the number of reference speakers, for PRSW M-best global pre-selection.

TABLE II
RESULTS FOR ALL 20 SPEAKERS ACROSS ALL METHODS

Speaker	UIM	PRSW, all	PRSW, weight α = 0.05 (M)	PRSW, global (global, M = 7)	SDIM
1	0.50	0.51	0.55 (8)	**0.56**	0.53
2	0.54	0.58	0.62 (6)	0.66	**0.67**
3	0.56	0.56	0.63 (6)	0.62	**0.72**
4	0.57	0.62	0.63 (6)	0.64	**0.69**
5	0.51	0.53	0.61 (6)	**0.62**	0.59
6	0.62	0.56	0.67 (7)	0.68	**0.72**
7	0.57	0.56	0.63 (7)	0.64	**0.65**
8	0.57	0.57	0.61 (5)	0.66	**0.65**
9	0.57	0.55	0.63 (6)	0.64	**0.68**
10	0.55	0.54	0.62 (8)	0.63	**0.67**
11	0.55	0.58	0.63 (6)	0.65	**0.72**
12	0.57	0.63	0.66 (6)	0.66	**0.69**
13	0.50	0.54	0.61 (5)	**0.61**	**0.61**
14	0.48	0.48	0.54 (8)	**0.57**	0.55
15	0.53	0.53	0.53 (7)	**0.55**	**0.55**
16	0.48	0.48	0.51 (7)	**0.54**	**0.54**
17	0.48	0.48	0.51 (7)	**0.53**	**0.53**
18	0.54	0.54	0.60 (8)	0.62	**0.64**
19	0.54	0.61	0.62 (7)	0.62	**0.64**
20	0.60	0.60	0.67 (7)	**0.65**	**0.65**
Average	0.54	0.55	0.60	0.62	**0.63**

The PRSW inversion model approaches of the accuracy the speaker-dependent (SDIM) inversion models, even though the SDIM model is based on kinematic data of the target speaker, while the PRSW method uses only acoustic adaptation data.

two affect the performance of the adapted model. The results shown here strongly indicate that one of the biggest factors in high quality kinematic-independent acoustic-to-articulatory inversion is a diverse set of reference speakers with consistent articulatory patterns.

It should be noted that the best choice of reference speakers is dependent on the application domain and on the reference speakers themselves, including both the consistency of their acoustic-articulatory patterns and their diversity in terms of representing a broad set of speakers for adaption. This emphasizes the need for a strong set of reference speakers so that it is possible to obtain a good base set of references from which to adapt new models.

D. Performance as a Function of the Quantity of Acoustic Adaptation Data

The PRSW experiments in the previous sections use the full set data from the target speaker to do adaptation, including 198 utterances representing approximately 28 minutes of speaking time. Normally RSW performs effectively with limited adaptation data. One question is whether PRSW still has this property under our proposed inversion framework, and what quantity of adaptation data is sufficient to obtain a good adapted articulatory model. In this section, the impact of the adaption data quantity on inversion performance is investigated. In the following experiments, the utterance set has been divided into 10 approximately equally sized subsets, each of which is roughly 20 utterances and 3 minutes of speaking time.

Fig. 12 shows the inversion performance versus the quantity of adaptation data for one speaker. The baseline PRSW method converges at about 70% of the adaptation data, while the reduced speaker-set PRSW methods converge at about 30–40% of the data, or 60–80 utterances. Although Fig. 12 shows the inversion performance for one specific speaker, this experiment has also been implemented for every speaker individually. Results show these same characteristics with slightly different convergence points for each speaker, ranging from 20 to 80 utterances (3 to 12 minutes of adaptation data). Although this is not as small an adaptation set as some rapid adaptation techniques, it would allow for implementation in the context of applications like

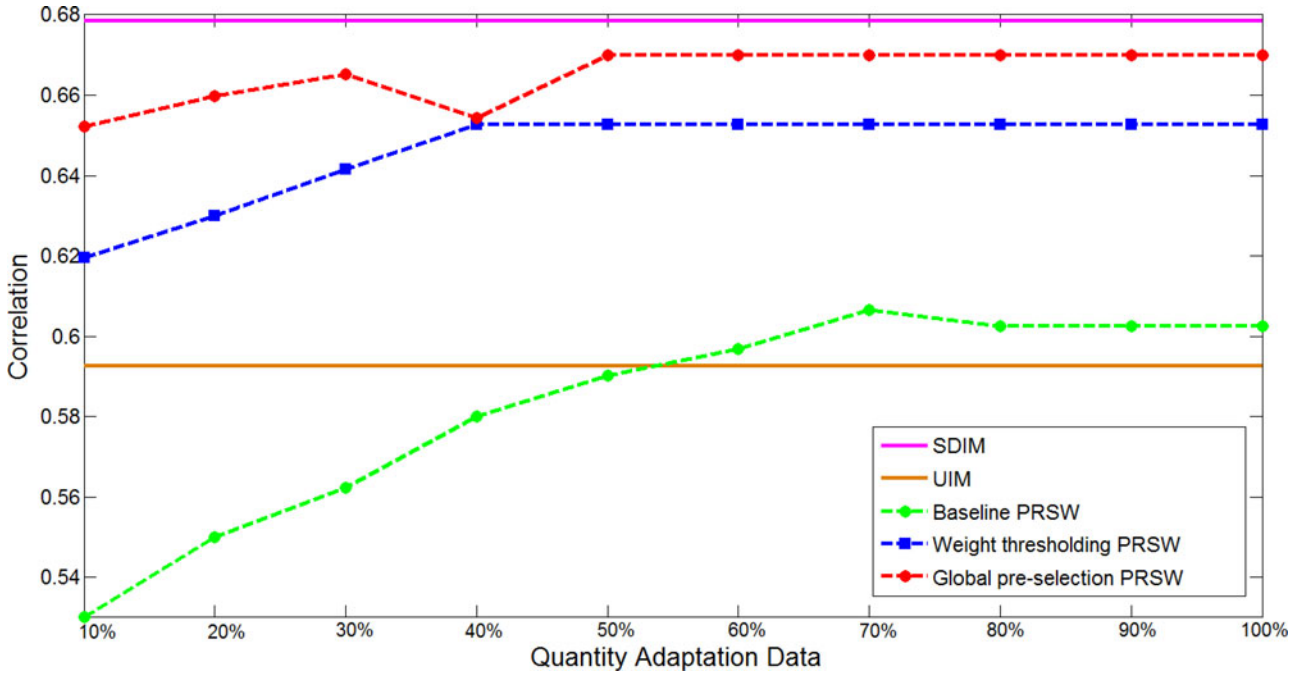


Fig. 12. Inversion performance versus total quantity of adaptation data for a single speaker. (Each adaptation subset represents approximately 3 additional minutes of data.)

pronunciation training with a relatively small amount of acoustic training materials.

V. Conclusion

This paper has presented an acoustic-to-articulatory inversion system that requires no kinematic data for the target speaker and a relatively small amount of acoustic adaptation data. The proposed PRSW framework adapts articulatory models using weights computed in the acoustic space. Initial baseline experiments showed variable performance as measured by correlation between estimated and actual trajectories, with indications that this variability is related to the selection of reference speakers. Based on this idea, two speaker selection methods have been considered, one based on thresholding the number of reference speakers based on acoustic model similarity to the target speaker, and another that is based on globally reducing the reference speaker set using speaker-dependent inversion performance. Experimental results show that both of these selection methods work better than the baseline system. The final system using a globally reduced speaker set resulted in an average correlation score of 0.62, nearly as high as the speaker-dependent result of 0.63 built using kinematic training data, and significantly higher than the universal model result of 0.54. This indicates that the proposed PRSW is able to adapt a good articulatory model for the target speaker without any kinematic data as long as the reference speaker set is carefully selected for acoustic and articulatory consistency. The comparable performance between the kinematic-independent and original speaker-dependent system supports the hypothesis that acoustic similarity can be used as a proxy for articulatory similarity in model building. Given a strong reference speaker set, the proposed PRSW adaptation is an effective approach for acoustic-to-articulatory inversion in the absence of kinematic training data.

References

[1] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *J. Acoustical Soc. Amer.*, vol. 111, pp. 1086–1111, 2002.

[2] V. Mitra, H. Nam, Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Retrieving tract variables from acoustics: A comparison of different machine learning strategies," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 6, pp. 1027–1045, Dec. 2010.

[3] Z. Ling, K. Richmond, J. Yamagishi, and R. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1171–1185, Aug. 2009.

[4] G. Hofer and K. Richmond, "Comparison of HMM and TMDN methods for lip synchronization," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 454–457.

[5] J. S. Levitt and W. F. Katz, "Augmented visual feedback in second language learning: Training japanese post-alveolar flaps to american english speakers," presented at the Proc. Meetings Acoustics, New Orelans, LA, USA, 2008.

[6] O. Engwall, "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher," *Comput. Assisted Lang. Learn.*, vol. 25, pp. 37–64, 2012.

[7] A. Suemitsu, T. Ito, and M. Tiede, "An EMA-based articulatory feedback approach to facilitate L2 speech production learning," *J. Acoust. Soc. Amer.*, vol. 133, 2013, Art. no. 3336.

[8] W. F. Katz, T. F. Campbell, J. Wang, E. Farrar, J. C. Eubanks, and A. Balasubramanian, "Opti-speech: A real-time, 3D visual feedback system for speech training," in *Proc. Interspeech*, 2014.

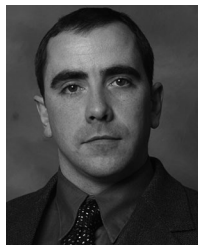
[9] W. F. Katz and S. Mehta, "Visual feedback of tongue movement for novel speech sound learning," *Frontiers Human Neurosci.*, vol. 9, 2015.

[10] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoust. Soc. Amer.*, vol. 100, pp. 1819–1834, 1996.

[11] T. Kaburagi and M. Honda, "An ultrasonic method for monitoring tongue shape and the position of a fixed-point on the tongue surface," *J. Acoustical Soc. Amer.*, vol. 95, pp. 2268–2270, 1994.
[12] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, The Centre for Speech Technol. Res., Edinburgh Univ., Edinburgh, U.K., 2002.
[13] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proc. 5th Seminar Speech Prod., Models Data*, 2000, pp. 237–240.
[14] T. Toda, A. Black, and K. Tokuda, "Acoustic-articulatory inversion mapping with Gaussian mixture model," in *Proc. Int. Conf. Spoken Lang. Process.*, Jeju Island, Korea, 2004, pp. 1129–1132.
[15] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 175–185, Mar. 2004.
[16] L. Zhang and S. Renals, "Acoustic-articulatory modelling with the trajectory HMM," *IEEE Signal Process. Lett.*, vol. 15, pp. 245–248, 2008.
[17] S. Hiroya and M. Honda, "Speaker adaptation method for acoustic-to-articulatory inversion using HMM based speech production model," *IEICE Trans. Inf. Syst*, vol. 87, pp. 1071–1078, 2004.
[18] T. Hueber, G. Bailly, P. Badin, and F. Elisei, "Speaker adaptation of an acoustic-articulatory inversion model using cascaded Gaussian mixture regressions," in *Proc. Interspeech*, Lyon, France, 2013, pp. 2753–2757.
[19] T. J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Commun.*, vol. 31, pp. 15–33, 2000.
[20] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Feb. 2012.
[21] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, Istanbul, Turkey, 2000, pp. 1315–1318.
[22] A. Ji, J. J. Berry, and M. T. Johnson, "The electromagnetic articulography mandarin accented english (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 7719–7723.
[23] Y. Yunusova, M. Baljko, G. Pintilie, K. Rudy, P. Faloutsos, and J. Daskalogiannakis, "Acquisition of the 3D surface of the palate by in-vivo digitization with wave," *Speech Commun.*, vol. 54, pp. 923–931, 2012.
[24] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 4624–4627.
[25] C. Huang, T. Chen, and E. Chang, "Speaker selection training for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Orlando, FL, USA, 2002, pp. 609–612.
[26] R. Kuhn, C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigen voice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

An Ji (S'14) received the B. S. and M. S. degrees in biomedical engineering from Chongqing University, Chongqing, China, in 2006 and 2009, respectively, and the Ph.D. degree in electrical and computer engineering from Marquette University, Milwaukee, WI, USA. She is currently working as a Speech Scientist at Ford Motor Company, Dearborn, MI, USA. Her research interests include robust speech recognition, natural language understanding, and voice interface design.

Michael T. Johnson (M'93–SM'02) received the B.S. degree in computer science engineering and the B.S. degree in engineering with electrical concentration from LeTourneau University, Longview, TX, USA, in 1989 and 1990, respectively, the M.S.E.E. degree from the University of Texas, San Antonio, TX, USA, in 1994, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2000. He was a Design Engineer and an Engineering Manager from 1990 to 1996, and is currently a Professor in the Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI, USA. His primary research area is speech and signal processing, and interests include machine learning, bioacoustics, and nonlinear signal processing.



Jeffrey J. Berry received the B.S. and M.S. degrees in communicative disorders from the University of Wisconsin, Madison, WI, USA, in 1994 and 2000, respectively, and the Ph.D. degree in communicative disorders with a concentration on speech motor control from the University of Wisconsin, Madison, WI, USA, in 2003. He was a National Institutes of Health Post-Doctoral Fellow at the University of Wisconsin Waisman Center from 2003 to 2005 and received the Certificate of Clinical Competence from the American Speech-Language-Hearing Association in 2006. He is currently an Associate Professor in the Department of Speech Pathology and Audiology, Marquette University, Milwaukee, WI, USA. His primary research examines how articulatory-acoustic relations are involved in speech sensorimotor control.