A SUBJECT-INDEPENDENT ACOUSTIC-TO-ARTICULATORY INVERSION

Prasanta Kumar Ghosh and Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering,
University of Southern California, Los Angeles, CA 90089

prasantg@usc.edu, shri@sipi.usc.edu

ABSTRACT

Acoustic-to-articulatory inversion is usually done in a subject-dependent manner, i.e., the inversion procedure may not work well if the parallel acoustic and articulatory training data is not available from the subjects in the test set. In this paper, we propose a subject-independent acoustic-to-articulatory inversion procedure; the proposed scheme requires acoustic-articulatory training data only from one subject and uses a generic acoustic model to perform acoustic-to-articulatory inversion for any arbitrary test subject. Experimental results on the MOCHA database show that the subject-independent inversion procedure can achieve an inversion accuracy close to the accuracy of the subject-dependent procedure especially for the lip aperture, tongue tip and tongue body articulatory trajectories. We also investigate various articulatory features to analyze the effectiveness of the proposed inversion procedure.

Index Terms— acoustic-to-articulatory inversion, electromagnetic articulography, tract variables, generalized smoothness criterion

1. INTRODUCTION

Estimation of representations in the articulatory space from representations in the acoustic space is known as acoustic-to-articulatory inversion. There are various representations or models available for both spaces. The characteristics of the mapping between articulatory space and acoustic space depend on the chosen representations. There have been different approaches for acoustic-to-articulatory inversion proposed in the literature such as dynamic programming (DP) [1], mixture density model [2], neural network model [3], generalized smoothness criterion (GSC) [4] etc. All these approaches use training data (i.e., parallel articulatory and acoustic representations) from a subject to learn the mapping between acoustic and articulatory spaces and then perform inversion on the (unseen) acoustic data of the subject. The realizations of the acoustic and articulatory spaces and, hence, their mapping vary across subjects due to the variability in the vocal tract configurations. Therefore, existing approaches for inversion may not work well if sufficient amount of the parallel acoustic and articulatory data from the target test subjects are not available during training.

Acoustic-to-articulatory inversion is potentially useful for deriving articulatory features for supporting speech or speaker recognition [5], where only the acoustic speech signal is available. Since direct articulatory evidence may be scarce (limited amounts of data available for a small number of talkers), in such cases, an efficient acoustic-to-articulatory inversion procedure needs to be developed which is robust to the lack of sufficient or no training data.

In this work, we propose an acoustic-to-articulatory inversion technique which requires acoustic-articulatory training data from only one subject and can be used to perform inversion on any other subject's acoustic signal. The proposed inversion technique works on the principle of representing an acoustic feature with respect to a generic acoustic space, obtained using speech data from a pool of talkers. Thus when a test subject's acoustic data is given for inversion, it is matched with the training subject's acoustic data with respect to the generic acoustic space. This enables us to obtain the articulatory feature trajectory using the articulatory-to-acoustic mapping learned from the data of the exemplary training subject. It should be noted that the range and values of the estimated articulatory trajectory correspond to the training subject and not the test subject. The estimated articulatory trajectories can be interpreted as the articulatory movement expected when the training subject tries to mimic the utterance spoken by the test subject. It is hypothesized that, for a given utterance, the reference articulatory trajectory (corresponding to test subject) and the estimated trajectory (corresponding to the training subject) will have similar shapes and, thus, correlation between these trajectories can be used to measure the accuracy of inversion. We investigate the efficacy of the proposed inversion using experimentally obtained articulatory data. We find that the accuracy of the estimated articulatory trajectory using the proposed approach is close to the accuracy obtained by an existing inversion technique where the parallel articulatory and acoustic data from the test subject is available to perform inversion.

2. DATASET AND PRE-PROCESSING

For the analysis and experiments of this paper, we use the Multichannel Articulatory (MOCHA) database [6] that contains acoustic and corresponding articulatory ElectroMagnetic Articulography (EMA) data from one male and one female talker of British English. The articulatory position data have high frequency noise resulting from EMA measurement error. Also the mean position of the articulators changes from utterance to utterance; hence, the position data needs pre-processing before it can be used for analysis. Following the pre-processing steps outlined in [4], we obtain parallel acoustic and articulatory data at a frame rate of 100 observations per second. Of the 460 utterances available from each speaker, data from 368 utterances (80%) are used for training, 37 utterances (8%) as the development set (dev set), and the remaining 55 utterances (12%) as the test set for the subject-dependent inversion procedure. For the proposed subject-independent inversion, one subject's data is used for training and the other subject's test utterances are used during testing.

3. ACOUSTIC AND ARTICULATORY FEATURES

Following the analysis of mutual information-based acoustic feature selection in [4], we chose the 14 dimensional mel frequency cepstral coefficients (MFCCs) including the zero-th coefficient (computed using 20 msec frame length and 10 msec frame shift) to be the features representing the acoustic space. The articulatory features are computed from the samples of the articulatory trajectories available in EMA data. We use 14 dimensional raw EMA features (i.e., X and Y co-ordinates of upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD), and

†Work supported in part by NIH and ONR-MURI.

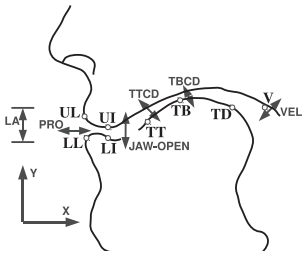

Fig. 1. *An illustrative diagram showing the direct EMA position features and the derived tract variable (TV) features.*

velum (V)). In addition, we use tract variable (TV) features motivated by articulatory phonology [7, 8], which conceptualizes speech as being produced as an ensemble of articulatory gestures. Gestures are defined as the dynamical control regimes for constriction actions in eight different constriction tract variables consisting of five constriction degree variables, lip aperture (LA), tongue body (TBCD), tongue tip(TTCD), velum (VEL), glottis (GLO), and three constriction location variables, lip protrusion (PRO), tongue tip (TTCL), tongue body (TBCL). TV features represent articulatory constrictions and do not depend on the absolute position of the articulators; hence, TV features are expected to be less variable across subjects as opposed to the raw EMA features which depend on the absolute sensor positions. Fig. 1 illustrates the EMA and TV features on the midsagittal plane. The considered TV features are computed in the following way:

$$\begin{aligned}
\text{LA} = |\text{ul_y} - \text{ll_y}|\,, \qquad & \text{PRO} = \frac{\text{ul_x} + \text{ll_x}}{2} \\
\text{JAW-OPEN} &= |\text{bn_y} - \text{li_y}| \ \text{(bn is a ref. sensor)} \\
\text{TTCD} &= \frac{|a_{tt}\text{tt_x} + b_{tt}\text{tt_y} + c_{tt}|}{\sqrt{a_{tt}^2 + b_{tt}^2}} \\
\text{TBCD} &= \frac{|a_{tb}\text{tb_x} + b_{tb}\text{tb_y} + c_{tb}|}{\sqrt{a_{tb}^2 + b_{tb}^2}} \\
\text{VEL} &= [\ \text{v_x} \quad \text{v_y}\ ]\mathbf{V_e}
\end{aligned}$$

a_{tt}, b_{tt}, and c_{tt} are the parameters of the line ($a_{tt}x + b_{tt}y + c_{tt} = 0$) representing the palate near TT region and similarly, a_{tb}, b_{tb}, and c_{tb} are for TB region. Since palate data were not available, these lines are manually fixed for each subject to represent the palate. $\mathbf{V_e}$ is a 2×1 unit norm eigenvector corresponding to the highest eigenvalue of the correlation matrix of the random vector $[\ \text{v\_x} \quad \text{v\_y}\ ]^{\text{T}}$. $[\cdot]^{\text{T}}$ is the transpose operator. Note that the computed TV features are not identical to the tract variables as defined in articulatory phonology; rather LA, PRO, JAW_OPEN, TTCD, TBCD, VEL are some representative features inspired by the tract variable concept.

4. PROPOSED SUBJECT-INDEPENDENT INVERSION

Based on the generalized smoothness criterion (GSC) for the acoustic-to-articulatory inversion introduced in [4], we develop a formulation to estimate the articulatory feature vector sequence for a given acoustic feature vector sequence corresponding to a test speech utterance. However, unlike GSC, in our proposed subject-independent inversion, we assume access to a generic acoustic space generated by the speech signal features (denoted by $\{\mathbf{c}_j; 1 \le j \le R\}$) obtained from various subjects (the TIMIT corpus [9] is chosen for this purpose) in addition to the parallel acoustic and articulatory features ($\{(\mathbf{z}_i, x_i); 1 \le i \le T\}$) from an exemplary training subject. Note that the acoustic data from the test subject for inversion need not be there in this generic acoustic space. We will show that this additional knowledge about the generic acoustic space will play a crucial role in estimating the articulatory features from the acoustics of any arbitrary test subject.

Let the test acoustic feature sequence be denoted by $\mathbf{u}_n,\ 1 \le n \le N$ (n denotes the frame index). Let the time sequence of an articulatory feature that we need to estimate from $\mathbf{u}_n,\ 1 \le n \le N$ be denoted by $x[n],\ 1 \le n \le N$. Since the articulatory feature trajectory is smooth, in general, the best estimate of an articulatory feature sequence is obtained using a smoothness criterion as follows [4]:

$$\begin{aligned}
\{x^\star[n]; 1 \le n \le N\} &= \arg\min_{\{x[n]\}} J\left(x[1], \cdots, x[N]\right) \\
&\triangleq \arg\min_{\{x[n]\}} \left\{ \sum_n \left(y[n]\right)^2 + C \sum_n \sum_l \left(x[n] - \eta_n^l\right)^2 p_n^l \right\}, \qquad (1)
\end{aligned}$$

where J denotes the cost function to be minimized and $y[n] = \sum_{k=1}^{N} x[k]h[n-k]$, where $h[n]$ is the articulatory feature-specific high-pass filter (FIR or IIR). $\left\{\eta_n^l; 1 \le l \le L\right\}$ is the set of L possible values of the articulatory feature at the n^{th} frame. p_n^l denotes the probability that η_n^l is the value of the articulatory position at the n^{th} frame given that $\mathbf{u}_n$ is the test acoustic feature. η_n^l and p_n^l are obtained using $\{(\mathbf{z}_i, x_i)\}$, $\{\mathbf{c}_j\}$, and $\mathbf{u}_n$. L can be, in general, equal to T. C is the trade-off parameter between the first term (smoothness constraint) and the second term (data proximity) in J.

A closed form solution of $x^\star[n]$ can be computed once $h[n]$, η_n^l, and p_n^l are determined [4]. Also, it was shown in [4] that the solution can be obtained recursively over time without any loss in performance. $h[n]$ and its cut-off frequency are designed here following [4]. Finally, the articulator-specific cut-off frequency and the trade-off variable C are optimized for the training subject on a development set (30% of the entire parallel acoustic and articulatory data of the training subject) so that the mean squared error (MSE) between the actual articulator trajectories and the estimated ones is minimized.

The basic principle of determining η_n^l, and $p_n^l,\ 1 \le l \le L$ in [4] was to choose articulatory features from the training data so that the corresponding acoustic features in the training data are close to the test acoustic feature in an Euclidean sense. Since the test and train acoustic features were considered for the same subject, such an acoustic proximity-based approach was appropriate for the subject-dependent acoustic-to-articulatory inversion. However, in the proposed subject-independent inversion framework, the test subject is different from the training subject and their acoustic spaces are different in general. Hence, a Euclidean distance measure $d_E(\mathbf{z}_i, \mathbf{u}_n)$ may not be a reliable metric of acoustic proximity due to inter-subject acoustic variability. Hence, we need to transform the acoustic feature vectors to another space where the closeness measure between two points $d\left(\Phi(\mathbf{z}_i), \Phi(\mathbf{u}_n)\right)$ is robust to such inter-subject variability, where $\Phi(\cdot)$ is the transformation function from the acoustic feature space to the new space and d is an appropriate measure of closeness between two points in the new space. Also, computing the distance between $\mathbf{u}_n$ and $\mathbf{z}_i,\ 1 \le i \le T$ at each frame is computationally expensive because T (the number of parallel acoustic and articulatory features of the training subject) is in the order of 10^5. For example, when a 14-dimensional MFCC vector with zero-th coefficient is considered as acoustic feature and T=5×10^5, the computation of the distances between $\mathbf{u}_n$ and $\mathbf{z}_i,\ 1 \le i \le T$ at each frame (n) requires $14T$ multiplications and $27T$ additions (and takes $\sim$0.4 second in MATLAB software on a desktop computer). Therefore, a computationally efficient closeness measure is desirable. Below we propose a transformation function Φ and a new closeness measure d in the range space of $\Phi(\cdot)$.

Let $\mathcal{A}$ be the acoustic space represented by $\mathbf{c}_j,\ 1 \le j \le R$. We perform a K-means clustering with K number of clusters; let $\mathcal{A}_k$ denote the k^{th} cluster. Note that $\bigcup_{k=1}^{K} \mathcal{A}_k = \mathcal{A}$. The density of the

data points in each cluster is modeled using a M-mixture Gaussian mixture model (GMM), i.e.,

$$p(\mathbf{v}|\mathcal{A}_k) = p(\mathbf{v}|\mathbf{v} \in \mathcal{A}_k) = \sum_{m=1}^{M} w_m \mathcal{N}\left(\mathbf{v}; \mu_m^k, \Sigma_m^k\right), \quad k = 1, ..., K$$

where $\mathbf{v}$ is the acoustic feature vector, μ_m^k and Σ_m^k are the mean vector and the covariance matrix of the m^{th} component of GMM in the k^{th} cluster respectively. μ_m^k and Σ_m^k are estimated using the expectation maximization algorithm [10]. w_m is the weight for the m^{th} component. Given an acoustic feature vector $\mathbf{v}$, $\Phi(\mathbf{v}) \triangleq \frac{1}{Z}\left[p(\mathbf{v}|\mathcal{A}_1) \cdots p(\mathbf{v}|\mathcal{A}_K)\right]^{\mathrm{T}}$, where $Z = \sum_{k=1}^{K} p(\mathbf{v}|\mathcal{A}_k)$, a normalization constant. Thus, $\Phi(\mathbf{v})$ is a K dimensional vector representing the probabilities of $\mathbf{v}$ to get generated from each of the K clusters in the acoustic space[1]. We used K=32 for our experiment (we also tried K=50, 64, but no significant performance benefit was obtained).

To determine η_n^l, one approach could be measuring proximity between $\Phi(\mathbf{u}_n)$ and $\Phi(\mathbf{z}_i)$, $1 \leq i \leq T$ using a distance metric, say, Euclidean distance. Since this can be prohibitive due to large T, we propose a closeness measure based on highest valued element in the $\Phi(\mathbf{v})$ vector. Note that $\Phi(\mathbf{z}_i)$, $1 \leq i \leq T$ can be computed a-priori. Let $\mathcal{B}_r = \{\mathbf{z}_i | r = \arg\max_k p(\mathbf{z}_i|\mathcal{A}_k)\}$, $1 \leq r \leq K$. Also let $\pi_i^r = \max_k \frac{p(\mathbf{z}_i|\mathcal{A}_k)}{\sum_k p(\mathbf{z}_i|\mathcal{A}_k)}$, $\forall \mathbf{z}_i \in \mathcal{B}_r$, $1 \leq r \leq K$. Given $\mathbf{u}_n$, $\Phi(\mathbf{u}_n)$ can be computed and suppose that r_1^{th} element in $\Phi(\mathbf{u}_n)$ is the largest among all elements, i.e., $r_1 = \arg\max_k p(\mathbf{u}_n|\mathcal{A}_k)$. Then we assume that $\mathbf{u}_n$ is acoustically closer to $\{\mathbf{z}_i | \mathbf{z}_i \in \mathcal{B}_{r_1}\}$ compared to $\{\mathbf{z}_i | \mathbf{z}_i \notin \mathcal{B}_{r_1}\}$, i.e., when the probability of generating two acoustic features by the same cluster is higher compared to other clusters in the acoustic space, those two feature vectors are assumed to be acoustically close. This reduces the search space of the acoustic features. We compute $\delta_i = \left| \frac{p(\mathbf{u}_n|\mathcal{A}_{r_1})}{\sum_k p(\mathbf{u}_n|\mathcal{A}_k)} - \pi_i^{r_1} \right|$, $\forall \mathbf{z}_i \in \mathcal{B}_{r_1}$. δ_i, $\forall \mathbf{z}_i \in \mathcal{B}_{r_1}$ are sorted in an ascending order. Note that δ_i is the absolute difference of two scaler values and hence, is computationally faster compared to computing the norm of two vectors. Let $\mathbf{z}_i'$, $1 \leq i \leq L$ be the acoustic features corresponding to the top L sorted δ_i. The articulatory features corresponding to $\mathbf{z}_i'$, $1 \leq i \leq L$ are used as η_n^l. p_n^l are computed using the Euclidean distance between $\Phi(\mathbf{u}_n)$ and $\Phi(\mathbf{z}_i')$, $1 \leq i \leq L$, i.e., $p_n^i = \frac{\Delta_i^{-1}}{\sum_{i=1}^{L} \Delta_i^{-1}}$, where, $\Delta_i = ||\Phi(\mathbf{u}_n) - \Phi(\mathbf{z}_i')||$. Note that the Euclidean distance (Δ_i) is computed only for L acoustic features and in practice, we choose $L << T$ (L=200 in our experiment). Therefore, the computation of η_n^l and p_n^l and, hence, the acoustic-to-articulatory mapping can happen in real time. For example, when 14-dimensional MFCC vector is considered as acoustic feature and T=5×10^5, the computation of η_n^l and p_n^l at each frame requires, on an average, only $\frac{1}{K}T$ additions (and takes ∼0.01 second in MATLAB software on a computer).

5. EXPERIMENTAL RESULTS

Note that there are only two subjects in the MOCHA corpus. We run the proposed inversion procedure to estimate articulatory feature trajectories of each subject in the corpus using the other subject's data for training (i.e. $\{\mathbf{z}_i, x_i\}$). For comparison, we have considered two different types of inversion procedure - 1) the subject-dependent inversion using GSC; this is identical to the one reported in [4], which is referred here as inversion scheme-1 (IS-1), 2) the subject-independent inversion using GSC as in [4] except that the training data is obtained from one subject and the other subject's data is used for testing and evaluation; we refer to this as inversion scheme-2 (IS-2). We refer to the proposed inversion procedure as inversion scheme-3 (IS-3). For consistency across three inversion schemes, we have used 12% of the available utterances (Section 2) for each subject as testset. Note that IS-1 is a subject-dependent inversion scheme and hence, it is expected that the best inversion accuracy will be obtained with IS-1, among the three inversion schemes considered.

We use Pearson's correlation coefficients ρ as a measure of accuracy between the estimated and reference articulatory trajectories. The estimated articulatory trajectory values will correspond to the shape and size of the training subject's articulators and, hence, will not be similar to the test subject's articulators, in general. Thus, we have not used root mean square (RMS) as a measure since it will not reflect the accuracy of inversion. Since the training and test subjects are not identical, it is important to choose appropriate features and the evaluation metric to analyze the quality of inversion. Ranges of the raw EMA feature values for the same articulator may be different between subjects, but the shape of the articulator trajectories are expected to be similar when two subjects utter the same utterance; this similarity will be more apparent for the TV features. For example, if the acoustic signal has a stop consonant /t/, tongue tip will go up to form the constriction against palate and will come down for the release. Thus, it is expected that both the reference and estimated trajectories of tt_y will have a peak corresponding to /t/ (similarly, TTCD trajectories will have a dip). However, the actual trajectory values may not be identical since the reference and estimated trajectory values corresponds to the test subject and training subject respectively. Thus, the quality of inversion can be measured by the similarity or correlation between the reference and the estimated trajectories. Hence, the higher the ρ, the better is the inversion quality.

Below we report the accuracy of inversion using IS-1, IS-2, and IS-3 separately for the cases when the EMA features and the TV features are used as the representations of the articulatory space.

5.1. EMA features

Fig. 2(a) and (b) show averaged ρ using IS-1, IS-2, and IS-3 on the male and female subject's testset, respectively, when EMA features are used to represent the articulatory space. It is clear that IS-1 yields the highest ρ for all EMA features due to its subject-dependent nature of inversion. ρ for IS-3 is greater than that for IS-2 for most of the EMA features indicating the effectiveness of the proposed inversion procedure. Lower ρ for IS-2 is expected since we select η_n^l based on closeness between the acoustic features of the two subjects and acoustic spaces of two subjects are, in general, different. Since, for IS-3, η_n^l are chosen based on the closeness between the acoustic features in a transformed probability space, IS-3 is more robust to inter-subject acoustic variations compared to IS-2.

5.2. TV features

TV features capture various constriction events during speech production. Thus, the TV features, to a certain extent, mitigate inter-subject vocal tract shape and articulatory configuration differences. Depending on the way we compute various TV features (Section 3), some of them such as PRO, VEL may suffer from greater inter-subject variability than LA, TTCD, and TBCD. This becomes clear from Fig. 3, which illustrates a set of randomly selected TV feature trajectories from the Female subject's testset and their estimates using IS-1 and IS-3. For clarity, estimates using IS-2 are not shown. It is evident from Fig. 3 that IS-1 yields more accurate trajectories since it is subject-dependent inversion. It is also clear from the figure that the computed TTCD, TBCD, and LA measures are invariant across subjects and hence the estimates of these feature trajectories using IS-3 are close to the reference trajectories although estimated and reference trajectories corresponds to two different subjects. In contrast, the values of the estimated trajectories of VEL, PRO, and

[1] Any quantity other than $p(\mathbf{v}|\mathcal{A}_k)$ can also be considered, e.g., the posterior probabilities $p(\mathcal{A}_k|\mathbf{v})$. However, in this paper, we have performed all experiments using $p(\mathbf{v}|\mathcal{A}_k)$. We do not get any significant improvement in recognition performance for the chosen corpus by using $p(\mathcal{A}_k|\mathbf{v})$ as against $p(\mathbf{v}|\mathcal{A}_k)$.

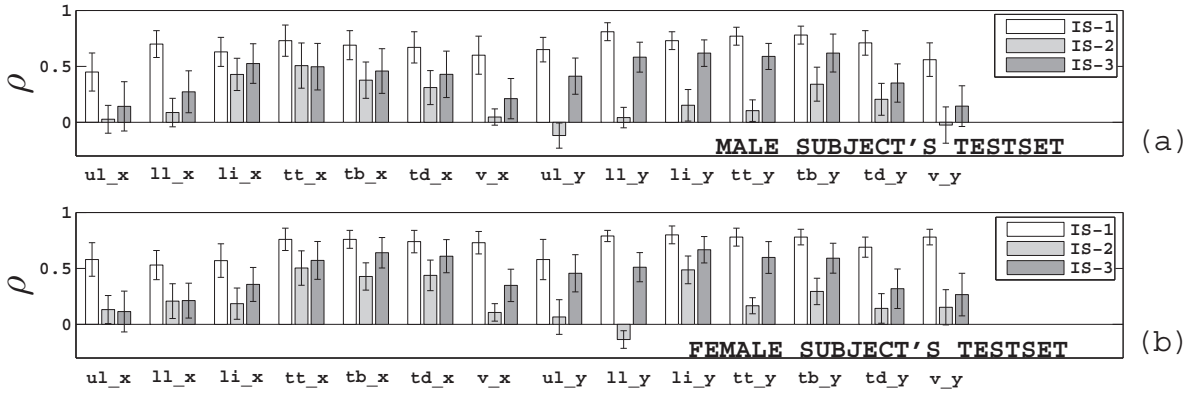

Fig. 2. Bar diagram of the mean ρ obtained using IS-1, IS-2, and IS-3 for various EMA features separately over all male and female subjects' test utterances. Errorbar indicates standard deviations of ρ across respective test utterances.

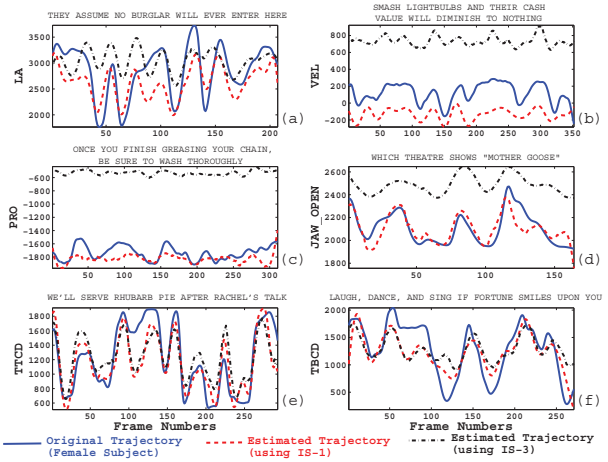

Fig. 3. Illustrative examples of the estimates of the TV feature trajectories using IS-1 and IS-3. These trajectories are randomly selected from the female subject's testset.

JAW_OPEN using IS-3 are different from those of the reference ones. However, there are similarities between the shapes of the trajectories. For a more comprehensive evaluation of the inversion quality for the TV features, we report the averaged ρ obtained using IS-1, IS-2, and IS-3 for both the male and female subjects' testset in Fig. 4. For different inversion schemes, we observe a trend in the ρ values similar to that for the EMA features (Fig. 2) indicating the effectiveness of the proposed inversion procedure in the case of TV features too.

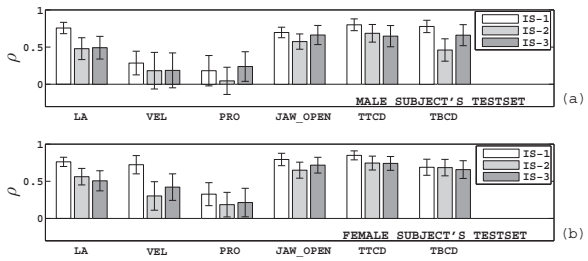

Fig. 4. Bar diagram of the mean ρ obtained using IS-1, IS-2, and IS-3 for various TV features separately over all male and female subjects' test utterances.

6. CONCLUSIONS

It is important to note that the proposed subject-independent inversion scheme that uses a transformed acoustic representation performs better than IS-2 but worse than IS-1 (subject-dependent inversion). It should also be noted that the proposed scheme is more efficient computationally than IS-1 and IS-2 ($\sim$40 times faster). Thus, the proposed inversion procedure (IS-3) is attractive when articulatory features need to be estimated in a subject-independent fashion in a real-time scenario for speech or speaker recognition. We estimate the trajectories for each articulator independently; however, the correlations among different articulator trajectories are well-known. Thus, the accuracy of the proposed inversion scheme can be further improved by exploiting inter-articulator correlation. Finally, subject-independent inversion provides us the opportunity to investigate the similarity between the acoustic-articulatory map across subjects. If the subject-independent inversion performs well on a specific set of test subjects, then their acoustic-articulatory map might be similar to that of the training subject. These considerations are part of our future works.

7. REFERENCES

[1] J. Schroeter and M. M. Sondhi, "Dynamic programming search of articulatory codebooks," *Proceedings ICASSP, Glasgow, UK*, vol. 1, pp. 588–591, 23-26 May 1989.

[2] A. Black T. Toda and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," *Proc. ICSLP, Jeju Island, Korea*, pp. 1129–1132, October 4-8 2004.

[3] K. Richmond, *Estimating articulatory parameters from the acoustic speech signal*, Ph.D. Thesis, The Centre for Speech Technology Research, Edinburgh University, 2002.

[4] P. K. Ghosh and S. S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2162–2172, 2010.

[5] A. Toutios and K. Margaritis, "Acoustic-to-articulatory inversion of speech: a review," *Proceedings of the International 12th TAINN*, 2003.

[6] A. A. Wrench and H. J. William, "A multichannel articulatory database and its application for automatic speech recognition," *5th Seminar on Speech Production: Models and Data, Bavaria*, pp. 305–308, 2000.

[7] C. P. Browman and L. Goldstein, "Towards an articulatory phonology," *Phonology Yearbook*, vol. 3, pp. 219–252, 1986.

[8] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, pp. 201–251, 1989.

[9] DARPA-TIMIT, *Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1*, 1990.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.