Analysis of inter-articulator correlation in acoustic-to-articulatory inversion using generalized smoothness criterion

Prasanta Kumar Ghosh and Shrikanth S. Narayanan

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

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

10.21437/Interspeech.2011-686

Abstract

The movements of the different speech articulators are known to be correlated to various degrees during speech production. In this paper, we investigate whether the inter-articulator correlation is preserved among the articulators estimated through acoustic-to-articulatory inversion using the generalized smoothness criterion (GSC). GSC estimates each articulator separately without explicitly using any correlation information between the articulators. Theoretical analysis of inter-articulator correlation in GSC reveals that the correlation between any two estimated articulators may not be identical to that between the corresponding measured articulatory trajectories; however, based on smoothness constraints provided by the real articulatory data, we found that, in practice, the correlation among articulators is approximately preserved in GSC based inversion. To validate the theoretical analysis on inter-articulator correlation, we propose a modified version of GSC where correlations among articulators are explicitly imposed. We found that there is no significant benefit in inversion using such modified GSC, which further strengthens the conclusions drawn from the theoretical analysis of inter-articulator correlation.

Index Terms: acoustic-to-articulatory inversion, inter-articulation correlation, 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. In this paper, we consider the problem of estimating articulatory position vector sequence (or trajectory) from a given MFCC (Mel Frequency Cepstral Coefficients of short time acoustic speech signal frames) vector sequence. Exemplary articulatory vectors can correspond to the positions (i.e., X and Y values in mid-sagittal plane) of the upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD), and the velum (V), as considered in this paper.

Previous studies have provided evidence that the mapping between the acoustic and articulatory spaces is not unique [1]. Therefore, many inversion techniques [2, 3, 4, 5, 6, 7, 8, 9] have been proposed where issues of non-uniqueness are mitigated by constraining the articulator trajectories to be smooth – e.g., low-pass filtering the estimated articulator trajectories [7, 8, 9]. We recently introduced a technique for incorporating the smoothness constraint inside the optimization problem using generalized smoothness criterion (GSC) [2]. In contrast to other inversion techniques, GSC estimates each articulator trajectory in an independent fashion using the acoustic feature sequence from a test utterance. However, it is well known that many of the measured articulators' movements can be correlated with one another [10]. For example, the upper and lower lips' movements are correlated since they move together to create lip opening and closure. Similarly since TT, TB, TD are three locations on the same physical tongue organ and, hence, their movements are expected to be correlated too. It is however not clear whether the GSC formulation preserves the inter-articulator correlation since it estimates each articulator independently.

In this work, we perform a theoretical analysis on the correlations among articulators estimated using GSC and derive their relations to the correlations among measured articulatory trajectories. Based on the analysis of inter-articulator correlation in GSC we show that, theoretically, there is no guarantee that GSC preserves the inter-articulator correlation as observed in the training data. However, when the theoretical analysis is examined with respect to real articulatory data, we found that the differences between the correlations among estimated articulators and those among measured articulators are not significant. Thus, it turns out that, in practice, the correlation among articulators estimated by GSC are approximately similar to those among measured articulators.

To further validate this inter-articulator correlation property of GSC, we develop a way within the GSC framework by which inter-articulator correlation can be explicitly imposed in the acoustic-to-articulatory inversion such that the empirical correlation coefficient between any two estimated articulatory trajectories will be identical to that between the respective articulatory variables (measured) in training data. The accuracy of the inversion obtained by exploiting the inter-articulator correlation is compared experimentally against the accuracy obtained by treating each articulator independently during inversion using GSC [2]. Based on the comparison, we observe that there is no significant benefit in explicitly imposing correlation in the inversion using GSC, which further justifies the validity of the theoretical analysis.

2. Dataset and pre-processing

For the analysis and experiments of this paper, we use the Multichannel Articulatory (MOCHA) database [11] containing acoustic and corresponding articulatory ElectroMagnetic Articulography (EMA) data from one male and one female subject. 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 as outlined by Ghosh et al. [2], we obtain parallel acoustic and articulatory data at a frame rate of 100 observations per second. Of the 460 utterances available from each subject, 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.

†Work supported by NIH and NSF.

Copyright © 2011 ISCA

3. Generalized smoothness criterion (GSC) for articulatory inversion

In this section, we briefly describe the principle of GSC [2] for inversion. Let $\{(\mathbf{z}_i,\ \mathbf{x}_i)\,; 1 \le i \le T\}$ represent the parallel acoustic feature vector and articulatory position vector pairs in the training set, where $\mathbf{z}_i$ and $\mathbf{x}_i$ represent the i^{th} acoustic and articulatory vector respectively. $\mathbf{x}_n$=$[x_n^1\ x_n^2\ \cdots\ x_n^{14}]^{\mathsf{T}}$, where the 14 elements correspond to X and Y co-ordinates of seven articulators considered in this work. $[\cdot]^{\mathsf{T}}$ denotes the transpose operator. Now suppose, for the acoustic-to-articulatory inversion, a (test) speech utterance is given and the acoustic feature vectors computed for this utterance are denoted by $\{\mathbf{u}_n; 1 \le n \le N\}$. The GSC is used to estimate the j^{th} articulatory position trajectory $\left\{x_n^j; 1 \le n \le N\right\}$ by solving the following optimization problem [2]:

$$\arg\min_{\{x_n^j\}} \left[\sum_n \left(y^j[n]\right)^2 + C_j \sum_n \sum_l \left(x_n^j - \eta_n^{l,j}\right)^2 p_n^l\right], \quad (1)$$

where, $y_n^j = \sum_{k=1}^N x_k^j h_{n-k}^j$ and h_n^j is an articulator specific high-pass filter with cut-off frequency f_c^j. Thus the first term on the right hand side of Eq. (1) is used to minimize the high frequency components in x_n^j so that the articulator position trajectory is smooth. $\left\{\eta_n^{l,j}; 1 \le l \le L\right\}$ are the L possible values of the j^{th} articulatory position at the n^{th} frame of the test speech utterance and p_n^l are their probabilities [2]. C_j is the trade off parameter between two terms in the objective function (Eq. (1)).

The solution of Eq. (1) for the j^{th} articulator can be written as

$$\mathbf{x}^{j\star} = \left(\mathbf{R}^j + C_j\mathbf{I}\right)^{-1} \mathbf{d}^j, \quad (2)$$

where, $\mathbf{x}^{j\star} = [x_1^{j\star}\ ...\ x_N^{j\star}]^{\mathsf{T}}$ is the optimal articulatory trajectory. $\mathbf{R}^j = \left\{R_{pq}^j\right\} = \left\{R^j(p-q)\right\} \triangleq \sum_n h^j[n-p]h^j[n-q]$. $\mathbf{I}$ is the identity matrix and $\mathbf{d}^j = \left[C_j \sum_l \eta_1^{l,j} p_1^l, \cdots, C_j \sum_l \eta_N^{l,j} p_N^l\right]^{\mathsf{T}}$. Note that the solution (Eq. (2)) can be obtained recursively with frame index n without any loss in accuracy [2].

4. Correlation among estimated articulator trajectories

We theoretically investigate how the correlations among estimated articulatory trajectories in GSC differ from those among the measured articulatory trajectories.

From Eq. (2), we can write (for $i \ne j$) the estimated articulatory trajectory in the following way:

$$\begin{cases} x_n^{j\star} = \sum_{m_1=1}^N a_{m_1} C_j \sum_{l=1}^L \eta_{m_1}^{l,j} p_{m_1}^l \\ x_n^{i\star} = \sum_{m_2=1}^N b_{m_2} C_i \sum_{l=1}^L \eta_{m_2}^{l,i} p_{m_2}^l \end{cases} \quad (3)$$

where, $\{a_{m_1},\ 1 \le\ m_1 \le N\}$ and $\{b_{m_2},\ 1 \le\ m_2 \le N\}$ are the n^{th} rows of $\left(\mathbf{R}^j + C_j\mathbf{I}\right)^{-1}$ and $\left(\mathbf{R}^i + C_i\mathbf{I}\right)^{-1}$ respectively. Let us assume η_n^l, i.e., $\left[\eta_n^{l,1}, \cdots, \eta_n^{l,14}\right]$, is *i.i.d.* random vector $\forall l, n$. Also let the j^{th} and i^{th} articulators, i.e., $\eta_n^{l,j}$ & $\eta_n^{l,i}$, have correlation co-efficients ρ_{ji}. In other words,

$$\begin{aligned} \mathbb{E}\left[\eta_n^{l,j}\right] = \mu_j, \qquad & \mathbb{V}\left[\eta_n^{l,j}\right] = \mathbb{E}\left[\left(\eta_n^{l,j} - \mu_j\right)^2\right] = \sigma_j^2, \\ \mathbb{COV}\left(\eta_{n_1}^{l_1,j},\ \eta_{n_2}^{l_2,i}\right) &= \mathbb{E}\left[\left(\eta_{n_1}^{l_1,j} - \mu_j\right)\left(\eta_{n_2}^{l_2,i} - \mu_i\right)\right] \\ &= \rho_{ji}\sigma_i\sigma_j\delta_{n_1,n_2}\delta_{l_1,l_2}, \end{aligned} \quad (4)$$

where, $\mathbb{E}[\cdot]$, $\mathbb{V}[\cdot]$, and $\mathbb{COV}[\cdot,\cdot]$ denote the mean, variance, and covariance of the random variables. $\delta_{m,n}$ is the kronecker delta, i.e., $\delta_{m,n} = 1$, when $m = n$ and $\delta_{m,n} = 0$, when $m \ne n$. Therefore, the mean of $x_n^{j\star}$ and $x_n^{i\star}$ are as follows:

$$\begin{aligned} \mathbb{E}\left[x_n^{j\star}\right] &= \sum_{m_1=1}^N a_{m_1} C_j \sum_{l=1}^L \mathbb{E}\left[\eta_{m_1}^{l,j}\right] p_{m_1}^l \\ &= \mu_j C_j \sum_{m_1=1}^N a_{m_1} \sum_{l=1}^L p_{m_1}^l \\ &= \mu_j C_j \sum_{m_1=1}^N a_{m_1} \left(\sum_{l=1}^L p_{m_1}^l = 1\right) \quad (5) \\ \text{Similarly, } \mathbb{E}\left[x_n^{j\star}\right] &= \mu_i C_i \sum_{m_2=1}^N b_{m_2} \quad (6) \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{V}\left[x_n^{j\star}\right] &= \mathbb{E}\left[\left(x_n^{j\star} - \mathbb{E}\left[x_n^{j\star}\right]\right)^2\right] \\ &= C_j^2 \left(\sum_{m_{11}=1}^N \sum_{m_{12}=1}^N \sum_{l_1=1}^L \sum_{l_2=1}^L a_{m_{11}} a_{m_{12}} p_{m_{11}}^{l_1} p_{m_{12}}^{l_2} \right. \\ &\qquad \left. \mathbb{E}\left[\left(\eta_{m_{11}}^{l_1,j} - \mu_j\right)\left(\eta_{m_{12}}^{l_2,j} - \mu_j\right)\right]\right) \\ &= C_j^2 \left(\sum_{m_{11}=1}^N \sum_{m_{12}=1}^N \sum_{l_1=1}^L \sum_{l_2=1}^L a_{m_{11}} a_{m_{12}} p_{m_{11}}^{l_1} p_{m_{12}}^{l_2} \right. \\ &\qquad \left. \sigma_j^2 \delta_{m_{11},m_{12}} \delta_{l_1,l_2}\right) \ \text{(using Eq. (4))} \\ &= C_j^2 \sigma_j^2 \sum_{m_1=1}^N \sum_{l=1}^L \left(a_{m_1} p_{m_1}^l\right)^2 \quad (7) \\ \&\ \mathbb{V}\left[x_n^{i\star}\right] &= C_i^2 \sigma_i^2 \sum_{m_2=1}^N \sum_{l=1}^L \left(b_{m_2} p_{m_2}^l\right)^2 \quad (8) \end{aligned}$$

And, the covariance between the j^{th} and i^{th} estimated articulators is

$$\begin{aligned} \mathbb{COV}\left(x_n^{j\star}, x_n^{i\star}\right) &= \mathbb{E}\left[\left(x_n^{j\star} - \mathbb{E}\left[x_n^{j\star}\right]\right)\left(x_n^{i\star} - \mathbb{E}\left[x_n^{i\star}\right]\right)\right] \\ &= C_i C_j \left(\sum_{m_1=1}^N \sum_{m_2=1}^N \sum_{l_1=1}^L \sum_{l_2=1}^L a_{m_1} b_{m_2} p_{m_1}^{l_1} p_{m_2}^{l_2} \right. \\ &\qquad \left. \mathbb{E}\left[\left(\eta_{m_1}^{l_1,j} - \mu_j\right)\left(\eta_{m_2}^{l_2,i} - \mu_i\right)\right]\right) \\ &= C_i C_j \left(\sum_{m_1=1}^N \sum_{m_2=1}^N \sum_{l_1=1}^L \sum_{l_2=1}^L a_{m_1} b_{m_2} p_{m_1}^{l_1} p_{m_2}^{l_2} \rho_{ji} \right. \\ &\qquad \left. \sigma_i\sigma_j\delta_{m_1,m_2}\delta_{l_1,l_2}\right) \\ &= C_i C_j \rho_{ji}\sigma_i\sigma_j \left(\sum_{m=1}^N \sum_{l=1}^L a_m b_m \left(p_m^l\right)^2\right) \end{aligned}$$

Hence, the correlation coefficient between the estimated j^{th} and i^{th} articulators, i.e., $x_n^{j\star}$ & $x_n^{i\star}$ is

$$\begin{aligned} \rho_{ji}^\star &= \frac{\mathbb{COV}\left(x_n^{j\star}, x_n^{i\star}\right)}{\sqrt{\mathbb{V}\left[x_n^{j\star}\right]}\sqrt{\mathbb{V}\left[x_n^{i\star}\right]}} \\ &= \frac{\sum_{m=1}^N \sum_{l=1}^L a_m b_m \left(p_m^l\right)^2}{\sqrt{\sum_{m=1}^N \sum_{l=1}^L \left(a_m p_m^l\right)^2}\sqrt{\sum_{m=1}^N \sum_{l=1}^L \left(b_m p_m^l\right)^2}} \rho_{ji} \quad (9) \end{aligned}$$

Thus, the correlation $\rho^{\star}_{ji}$ between the j^{th} and i^{th} articulator trajectories estimated using GSC is not necessarily identical to the inter-articulator correlation ρ_{ji} in the training set. It is important to note that when the cut-off frequencies of the j^{th} and i^{th} articulator-specific filters are nearly identical, their impulse responses (i.e., h^j_n and h^i_n) and hence the correlation matrices $\mathbf{R}^j$ and $\mathbf{R}^i$ are approximately same. In addition, if the trade-off parameters of the respective articulators, i.e., C_j and C_i, are similar, then $a_m \approx b_m$, $1 \leq m \leq N$ because $\{a_{m_1},\ 1 \leq\ m_1 \leq N\}$ and $\{b_{m_2},\ 1 \leq\ m_2 \leq N\}$ are the n^{th} rows of $\left(\mathbf{R}^j + C_j\mathbf{I}\right)^{-1}$ and $\left(\mathbf{R}^i + C_i\mathbf{I}\right)^{-1}$ respectively. Under such circumstances (i.e., $a_m \approx b_m$), it is easy to see that $\rho^{\star}_{ji} \approx \rho_{ji}$; this means that when $a_m \approx b_m$, the correlation is approximately preserved among the articulators estimated by GSC. However, as shown by Ghosh et al. [2], neither f^j_C nor C_j is identical across different articulators and, hence, we compute $\rho^{\star}_{ji}/\rho_{ji}$ (from Eq. (9)) for each test utterance. Fig. 1 demonstrates the average values of $\rho^{\star}_{ji}/\rho_{ji}$ for each pair of articulatory variables (j-th and i-th, $1 \leq j, i \leq 14$) separately for each subject in the MOCHA database. Note that standard deviation (SD) of $\rho^{\star}_{ji}/\rho_{ji}$ over all test utterances is minimal (maximum SD among all pair of articulatory variables is 5.54×10^{-5}).

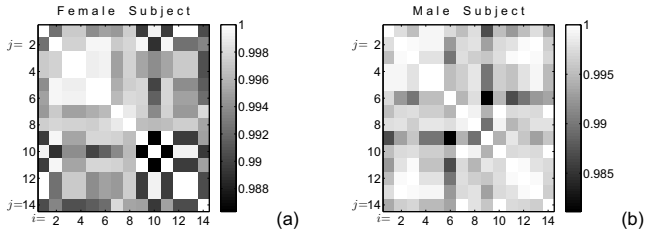

Figure 1: $\rho^{\star}_{ji}/\rho_{ji}$ for all pairs of articulators for (a) female (b) male subjects in the MOCHA database. i and j vary over articulatory variable index 1,...,14.

It is clear from Fig. 1 that $\rho^{\star}_{ji}/\rho_{ji}$ is very close to 1 for all pairs of articulators and hence $\rho^{\star}_{ji}$ is approximately same as ρ_{ji} for each subject in the MOCHA database. Therefore, in practice GSC approximatley preserves the inter-articulator correlation although theoretically the correlation among estimated articulators is not identical to that among measured articulatory variables.

To further validate this conclusion based on this theoretical analysis, we develop a modified version of GSC framework where inter-articulator correlation is explicitly imposed among estimated articulators during inversion. The goal is to examine where there is any significant benefit in inversion by explicitly preserving the inter-articulator correlation and thereby examine the validity of the theoretical analysis above.

5. Modified GSC to preserve inter-articulator correlation

From Eq. (9) it is worth noting that if the variables in the training set ($\eta^{l,j}_n$ & $\eta^{l,i}_n$) were uncorrelated (i.e., $\rho_{ji} = 0$), then the estimated trajectories ($x^{j\star}_n$ and $x^{i\star}_n$) would also be uncorrelated (i.e., $\rho^{\star}_{ji}$=0). This observation motivates us to transform the articulatory position variables, $\left\{x^j_n;\ j = 1, ..., 14\right\}$, into another set of variables, $\left\{\tilde{x}^j_n;\ j = 1, ..., 14\right\}$, where $\tilde{x}^j_n$ and $\tilde{x}^k_n$ are uncorrelated $\forall\ j,\ k$. The GSC can be used in the transformed variable domain for inversion and, after inversion, the correlation between variables can be imposed by transforming them back to the original articulatory position variable domain. Whitening is one of the approaches where a random vector ($\mathbf{x}_n$, in our case) is linearly transformed to make its components uncorrelated [12]. We transform $\mathbf{x}_n$ to obtain $\tilde{\mathbf{x}}_n$.

5.1. Transformation of the articulatory position vector

Let $\mu_{\mathbf{x}}$ and $\mathbf{K_{xx}}$, respectively, be the mean vector and the covariance matrix of the random vector $\mathbf{x}_n$. Let the eigen decomposition of $\mathbf{K_{xx}}$=$V\Lambda V^{\mathrm{T}}$, where V is the orthogonal eigenvector matrix ($V^{\mathrm{T}}V = I$) and Λ is the diagonal eigen value matrix. The following linear transformation whitens $\mathbf{x}_n$ to the random vector $\tilde{\mathbf{x}}_n$; the components of $\tilde{\mathbf{x}}_n$ are uncorrelated as it is easy to show that $\tilde{\mathbf{x}}_n$ has a diagonal covariance matrix $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$.

$$\tilde{\mathbf{x}}_n = V^{\mathrm{T}}\mathbf{x}_n \tag{10}$$

where, $\mu_{\tilde{\mathbf{x}}} = \mathrm{E}\left(\tilde{\mathbf{x}}_n\right) = V^{\mathrm{T}}\mu_{\mathbf{x}}, \quad \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = V^{\mathrm{T}}\mathbf{K_{xx}}V = \Lambda.$

Note that the component of $\tilde{\mathbf{x}}_n$ does not correspond to any physically meaningful articulatory parameters any more. However, $\mathbf{x}_n$ can be recovered from $\tilde{\mathbf{x}}_n$ by $\mathbf{x}_n = \left(V^{\mathrm{T}}\right)^{-1}\tilde{\mathbf{x}}_n = V\tilde{\mathbf{x}}_n$.

5.2. Frequency analysis of the transformed variables

In the GSC formulation [2], the cut-off frequencies of the articulator specific high-pass filters h^j_n are determined based on the analysis of the frequency content of all articulatory position variables $\left\{x^j_n;\ j = 1, ..., 14\right\}$. Similarly, to estimate $\tilde{x}^j_n$ using GSC, we need to analyze the frequency content of the transformed variables $\left\{\tilde{x}^j_n;\ j = 1, ..., 14\right\}$ to determine the cut-off frequency of the high-pass filters in the transformed variable domain. We calculate the frequency f^j_c below which α% of the total energy of the transformed variable trajectory is contained. This is done for each utterance in the training set and the mean f^j_c (along with standard deviation (SD)) over all utterances is calculated for α=90, 95. It is found that the range of f^j_c for most of the transformed variables $\tilde{x}^j_n$ is similar to what was observed in the frequency analysis of the articulator position variables in x^j_n [2]. This is expected since $\tilde{x}^j_n$ is a linear combination of the articulator position variables $\mathbf{x}_n$. This analysis will help us choose the cut-off frequencies while designing filters $h^j_n,\ \forall j$ in GSC.

5.3. Inversion using transformed articulatory features

Since we do not know the true covariance matrix $\mathbf{K_{xx}}$ of $\mathbf{x}_n$, we estimate $\mathbf{K_{xx}}$ using the realizations of $\mathbf{x}_n$ in the training set as follows:

$$\mathbf{K_{xx}} = \frac{1}{T-1}\sum_{n=1}^{T}\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)^{\mathrm{T}} \tag{11}$$

where $\bar{\mathbf{x}} = \frac{1}{T}\sum_{n=1}^{T}\mathbf{x}_n$. Using eigen-decomposition of estimated $\mathbf{K_{xx}}$, we obtain the eigenvector matrix V, which is used to transform $\mathbf{x}_n$ vectors of training, dev and test set to $\tilde{\mathbf{x}}_n$ (using Eq. (10)), where variables are uncorrelated. Parallel acoustic vectors and transformed articulatory vectors $\tilde{\mathbf{x}}_n$ of the training set are used to estimate $\eta^{l,j}_n$ and p^l_n in a way similar to that described by Ghosh et al. [2]. The GSC (Eq. (1)) is used to estimate the trajectories of $\tilde{x}^{j\star}_n,\ \forall\ j = 1, ..., 14$ separately using the acoustic data of the test set. Finally, $x^{j\star}_n,\ \forall\ j = 1, ..., 14$ are obtained by transforming $\tilde{\mathbf{x}}^{\star}_n$ back using $\mathbf{x}^{\star}_n = [x^{1\star}_n\ x^{2\star}_n\ \cdots\ x^{14\star}_n]^{\mathrm{T}} = V\tilde{\mathbf{x}}^{\star}_n = V[\tilde{x}^{1\star}_n\ \tilde{x}^{2\star}_n\ \cdots\ \tilde{x}^{14\star}_n]^{\mathrm{T}}$. By this reverse transformation, we correlate different variables so that correlation among them is similar to the inter-articulator correlation observed in the training data. Note that the transformation matrix (V) is learned on the training set. Therefore, it is assumed that correlations between different articulator positions in the test set are similar to those in the training set.

5.4. Articulatory inversion results using modified GSC

The proposed approach of utilizing inter-articulator correlation for acoustic-to-articulatory inversion is evaluated separately for

the male and female subjects data in the MOCHA-TIMIT corpus [11]. The accuracy of the inversion is evaluated separately on the test set for both subjects in terms of both root mean squared (RMS) error and Pearson correlation co-efficient [13] between the actual articulatory position in the test set and the position estimated by GSC. The RMS error $\mathcal{E}$ reflects the average closeness between actual and estimated articulator trajectories. The correlation ρ indicates how similar the actual and estimated articulator trajectories are.

The dev set is used to tune the cut-off frequency f_c^j of filter h_n^j and the trade-off parameter C_j. For our experiment we considered L=200. Increasing L further did not improve the result. For designing articulator specific high-pass filters h_n^j, we considered an IIR high pass filter with cut-off frequency f_c^j and stop-band ripple 40 dB down compared to the pass-band ripple similar to that used by Ghosh et al [2]. From Section 5.2, we observed that most of the energy of the transformed variables is below 9-10 Hz and, hence, we choose values of f_c^j from the following set $\left\{f_c^j\right\} = \left\{1.5 + \frac{(k-1)}{19}\,(7.5)\,; k = 1, \cdots, 20\right\}$. Similarly, the set of values for C_j was chosen from the set {.001, .005, .01, 0.05, .1, .5, 1, 5, 10, 50, 100}. The f_c^j and C_j combination which yielded the minimum value of the averaged $\mathcal{E}$ (averaged over all utterances of the dev set) are selected separately for each subject. These best choices of f_c^j and C_j are used to perform inversion on the test set using modified GSC. The $\mathcal{E}$ and ρ are computed between the actual trajectories and the estimated trajectories for every utterance in the test set. The mean $\mathcal{E}$ and ρ (averaged over all utterances in the test set) along with the standard deviation (SD) are shown in Fig. 2 for both the female and male subjects. For comparison, Fig. 2 also shows $\mathcal{E}$ and ρ when GSC is directly used in the articulator position variable domain [2].

From Fig. 2, it is clear that for most of the cases, the accuracy of estimates is similar or higher (i.e., lower $\mathcal{E}$ or higher ρ) when inter-articulator correlation is utilized using transformation of variables, but there are cases when GSC on transformed domain either increases the mean $\mathcal{E}$ or decreases the mean ρ (e.g. ul_x, ll_x, ul_y for male subject and tt_x for female subject). However, considering the SD of $\mathcal{E}$ and ρ, the changes in the accuracy in terms of $\mathcal{E}$ and ρ are insignificant. Thus the gain in performance because of explicitly using inter-articulator correlation by transformation of variable approach is not significant. This supports the conclusions based on the theoretical analysis (Sec. 4) that the correlation among articulators are approximately preserved in GSC and, hence, there is no further benefit by imposing inter-articulator correlation explicitly.

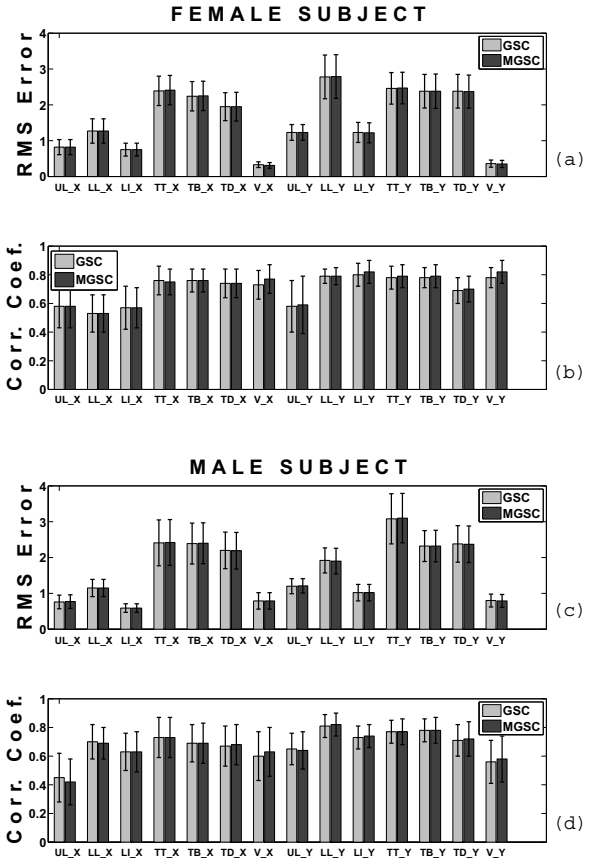

Figure 2: (a)-(b): The average RMS Error ($\mathcal{E}$) and Corr. Coef. (ρ) of inversion accuracy using GSC [2] and modified GSC (MGSC) on the test set for the female subject; (c)-(d): (a)-(b) repeated for the male subject. Error bars indicate SD.

6. Conclusions

The analysis of inter-articulator correlation in this paper reveals that acoustic-to-articulatory inversion using GSC approximately preserves inter-articulator correlations although articulatory inversion in GSC is performed for each articulatory trajectory separately. Using both theoretical and experimental analysis, we observe that the smoothness constraints for different articulators are similar and this could be the reason for GSC to approximately preserve correlation in articulatory inversion in practice.

7. References

[1] C. Qin and M. A. Carreira-Perpinan, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," *Proc. Interspeech*, pp. 74–77, 2007.

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

[3] V. N. Sorokin, A. Leonov, and A. V. Trushkin, "Estimation of stability and accuracy of inverse problem solution for the vocal tract," *Speech Communication*, vol. 30, pp. 55–74, 2000.

[4] J. Schroeter and M. M. Sondhi, "Dynamic programming search of articulatory codebooks," *Proceedings ICASSP, Glasgow, UK*, pp. 588–591, 1989.

[5] M. G. Rahim, W. B. Kleijn, J. Schroeter, and C. C. Goodyear, "Acoustic-to-articulatory parameter mapping using an assembly of neural networks," *Proc. ICASSP*, pp. 485–488, 1991.

[6] 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. Am.*, vol. 100, no. 3, pp. 1819–1834, 1996.

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

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

[9] ——, "Mixture density networks, human articulatory data and acoustic-to-articulatory inversion of continuous speech," *Proceedings Workshop on Innovation in Speech Processing WISP*, 2001.

[10] V. D. Singampalli and P. J. B. Jackson, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695–710, 2009.

[11] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," *Proc. ICSLP, Beijing, China*, pp. 145–148, 2000.

[12] H. Stark and J. W. Woods, *Probability and Random Processes with Applications to Signal Processing*, 3rd ed. Prentice Hall, August 3, 2001.

[13] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. Chapman & Hall, 1974.