ML vs. MAP PARAMETER ESTIMATION OF LINEAR DYNAMIC SYSTEMS FOR ACOUSTIC-TO-ARTICULATORY INVERSION: A COMPARATIVE STUDY

İ. Yücel Özbek and Mübeccel Demirekler

EE Department, Middle East Technical University, Ankara, Turkey

email: iozbek@metu.edu.tr, demirek@metu.edu.tr

ABSTRACT

This work proposes a maximum a posteriori (MAP) based parameter learning algorithm for acoustic-to-articulatory inversion. Inversion method is based on single global linear dynamic system (GLDS) representation of acoustic and articulatory data. MAP based learning algorithm considers a prior distribution for the parameter set as well as the likelihood of the training data. Therefore in this paper, we investigate the selection of prior distributions with hyperparameters for GLDS to improve the performance of articulatory inversion. The performance of the proposed learning algorithm and comparison of it with the maximum likelihood (ML) based learning method are examined on an extensive set of examples. These results show that the performance of the articulatory inversion method based on GLDS is significantly improved via MAP based learning algorithm.

1. INTRODUCTION

Electromagnetic Articulography (EMA) Trajectories provide the movement of certain articulators during a speech utterance. They contain useful information about speech production and this information can be used in a variety of speech applications including speech recognition and synthesis. Therefore, the reliable estimation of articulatory trajectories could improve performance of these applications. Recently, numerous methods are proposed to find reliable estimates of articulatory trajectories. Among them, [1, 2] uses neural network and mixture density network. Other methods given in the literature are GMM regression [3, 4], HMM [5], SVM regression [6, 7] and codebook usage [8]. A combination of acoustic and visual features is used in [9, 10].

Consideration of articulatory inversion as a state estimation problem via state space representation can be seen in [11, 12, 13]. In state space representation, the position of the each articulator is considered as a state of the dynamic system and they are governed by state equation. The observations are the acoustic (and/or visual) data like Mel-frequency cepstral coefficients (MFCC), and the transformation from acoustic to articulatory data is controlled via observation equation. The observation equation is either a nonlinear [12, 13] or a linear affine function [11]. Studies in the literature estimate the parameters of the model by ML criterion [12, 13, 11]. The data used in parameter estimation, i.e. the training data consists of articulatory and acoustic vector pairs.

State space representation of articulatory inversion problem gives a compact formulation so that filtering and smoothing can be applied relatively easily. For this purpose Bayesian recursive estimation (i.e. Kalman filter etc.) can be used. The performance of the system is directly related to appropriate modeling of the state and observation, and the accuracy of the estimation of model parameter set. Appropriate modeling may not be possible due to uncertainties in the system structures. The accuracy of the parameter estimation is related to the adequacy and consistency of the training samples in the training database. If any one of them (modeling and parameter estimation problems) is poor, the performance of the inversion system will degrade.

To avoid these types of problems, in this paper we propose a MAP based learning algorithm for a single GLDS to be used in estimation of articulatory trajectories. MAP based learning algorithm use some prior information about model parameters. Therefore, we also propose a prior density selection method to improve performance of the articulatory inversion. We also compare the performances of both ML and MAP based learning algorithms in various experiments.

The rest of the paper is organized as follows: Sec.2 gives problem formulation of articulatory inversion based on single global linear dynamic system (GLDS). Sec.3 describes learning and inference methods for GLDS. The experimental results are given in Sec.4. Sec.5 presents conclusions and future work plan.

2. ARTICULATORY INVERSION BASED ON GLDS

The acoustic-to-articulatory inversion problem can be converted into the state estimation problem of a single global linear dynamic system (GLDS). The dynamics of articulation and acoustic-to-articulatory transform are characterized via piece-wise affine functions given as follows.

$$x_{k+1} = Fx_k + u + w_k \tag{1}$$

$$z_k = Hx_k + d + v_k \tag{2}$$

where

- $x_k \in \mathbb{R}^{n_x}$ denotes the continuous-valued state vector related to articulatory data with dimension of n_x
- $z_k \in \mathbb{R}^{n_z}$ denotes the observation vector related to the acoustic data with dimension of n_z
- w_k and v_k are Gaussian white noise with corresponding covariances $Q \in \mathbb{R}^{n_x \times n_x}$ and $R \in \mathbb{R}^{n_z \times n_z}$

$$w_k \sim \mathcal{N}(w_k; 0, Q)$$

$$v_k \sim \mathcal{N}(v_k; 0, R)$$

- Initial state x_1 has Gaussian distribution with following parameters

$$x_1 \sim \mathcal{N}(x_1; \bar{x}, \Sigma)$$

- $F \in \mathbb{R}^{n_x \times n_x}$ and $H \in \mathbb{R}^{n_z \times n_x}$ are state transition and observation matrices respectively. $u \in \mathbb{R}^{n_x}$ and $d \in \mathbb{R}^{n_z}$ are corresponding bias vectors.
- The over all parameter set of GLDS is $\Theta = \{\bar{x}, \Sigma, H, d, R, F, u, Q\}$

Assume that we have a training database $D = \{X, Z\}$ that links acoustic observations Z and articulatory observations X. The problem of the acoustic-to-articulatory inversion involves two separate tasks, that we may call "learning" and "inference":

- LEARNING: Learning is the estimation of the model parameters Θ given the training data set D and prior distribution $p(\Theta)$. We examine both maximum likelihood (ML) and a maximum a posterior (MAP) learning methods. That is,

$$\hat{\Theta}^{ML} = \arg\max_{\Theta} p(Z, X|\Theta),$$
$$\hat{\Theta}^{MAP} = \arg\max_{\Theta} p(Z, X|\Theta)p(\Theta).$$

- INFERENCE: The estimation of the articulatory state x_k given acoustic data $z_{1:\tau} = \{z_1, \ldots, z_\tau\}$ and estimated parameter set $\hat{\Theta}$. Estimated state is found via minimum mean square error (MMSE) method as follows.

$$x_{k|\tau} = \mathrm{E}[x_k|z_{1:\tau}]$$

where $\mathrm{E}[\cdot]$ is the expectation operator. If $\tau = k$, the estimation is called filtering; if $\tau = N$ (where N is the length of the observation sequence), the estimation is called fixed-interval smoothing.

The following two sections explore, in detail, the problems of learning Θ from measured X, Z, and of inferring X from measured Z.

3. LEARNING AND INFERENCE

3.1 Learning

For the estimation of the parameter vector Θ, suppose that training database D contains L training sequences that contains acoustic observations $Z = \{z^l_{1:N_l}\}^L_{l=1}$ and articulatory observations, $X = \{x^l_{1:N_l}\}^L_{l=1}$. Suppose that each of the lth sequence contains N_l vectors, that is $x^l_{1:N_l} = \{x^l_1, \ldots, x^l_{N_l}\}$ and $z^l_{1:N_l} = \{z^l_1, \ldots, z^l_{N_l}\}$.

3.1.1 *Maximum likelihood (ML) Based Learning*

In the maximum likelihood learning criterion, the parameter set Θ can be estimated via maximizing the logarithm of the joint likelihood function $L(\Theta) = p(Z, X|\Theta)$ using training data set D.

$$\hat{\Theta}^{ML} = \arg\max_{\Theta} \ln L(\Theta) \tag{3}$$

Under independent observation sequences assumption, $L(\Theta)$ can be written as follows

$$\begin{aligned} L(\Theta) &= \prod_{l=1}^{L} p(x^l_{1:N_l}, z^l_{1:N_l}|\Theta) \\ &= \prod_{l=1}^{L} \left(p(x^l_1|\Theta) \prod_{k=1}^{N_l} p(z^l_k|x^l_k, \Theta) \prod_{k=2}^{N_l} p(x^l_k|x^l_{k-1}, \Theta) \right) \end{aligned} \tag{4}$$

Table 1: ML Based Parameter Estimation for GLDS

Define the following summations:

$$N \triangleq \textstyle\sum_{l=1}^{L} N_l - 1, \qquad \bar{x}_c \triangleq \frac{1}{N} \sum_{l=1}^{L} \sum_{k=1}^{N_l-1} x_k.$$

$$\bar{x}_p \triangleq \textstyle\frac{1}{N} \sum_{l=1}^{L} \sum_{k=2}^{N_l} x_{k-1}, \quad \bar{z}_c \triangleq \frac{1}{N+L} \sum_{l=1}^{L} \sum_{k=1}^{N_l} z_k.$$

ML based estimated parameters:

$$\hat{\bar{x}} = \textstyle\frac{1}{L} \sum_{k=1}^{L} x^l_1, \quad \hat{\Sigma} = \frac{1}{L} \sum_{l=1}^{L} (x^l_1 - \hat{\bar{x}})(x^l_1 - \hat{\bar{x}})^T$$

$$\begin{aligned} \hat{F} = &\left(\textstyle\sum_{l=1}^{L} \sum_{k=2}^{N_l} (x^l_k - \bar{x}_c)(x^l_{k-1} - \bar{x}_p)^T \right) \\ &\times \left(\textstyle\sum_{l=1}^{L} \sum_{k=2}^{N_l} (x^l_{k-1} - \bar{x}_p)(x^l_{k-1} - \bar{x}_p)^T \right)^{-1} \end{aligned}$$

$$\hat{u} = \bar{x}_c - \hat{F}\bar{x}_p$$

$$\hat{Q} = \textstyle\frac{1}{N} \sum_{l=1}^{L} \sum_{k=2}^{N_l} (x^l_k - \hat{F}x^l_{k-1} - u)(x^l_k - \hat{F}x^l_{k-1} - u)^T$$

$$\begin{aligned} \hat{H} = &\left(\textstyle\sum_{l=1}^{L} \sum_{k=1}^{N_l} (z^l_k - \bar{z}_c)(x^l_k - \bar{x}_c)^T \right) \\ &\times \left(\textstyle\sum_{l=1}^{L} \sum_{k=1}^{N_l} (x^l_k - \bar{x}_c)(x^l_k - \bar{x}_c)^T \right)^{-1} \end{aligned}$$

$$\hat{d} = \bar{z}_c - \hat{H}\bar{x}_c$$

$$\hat{R} = \textstyle\frac{1}{N+L} \sum_{l=1}^{L} \sum_{k=1}^{N_l} (z^l_k - x^l_k - \hat{d})(z^l_k - x^l_k - \hat{d})^T$$

Taking the logarithm of $L(\Theta)$ and substituting in (3) gives

$$\begin{aligned} \hat{\Theta}^{ML} = \arg\max_{\Theta} &\left(\sum_{l=1}^{L} \sum_{k=1}^{N_l} \ln p(z^l_k|x^l_k, H, d, R) \right. \\ &\left. + \sum_{l=1}^{L} \ln p(x^l_1|\bar{x}, \Sigma) + \sum_{l=1}^{L} \sum_{k=2}^{N_l} p(x^l_k|x^l_{k-1}, F, u, Q) \right) \end{aligned} \tag{5}$$

where,

$$p(x^l_1|\bar{x}, \Sigma) \triangleq \mathscr{N}(x^l_1; \bar{x}, \Sigma) \tag{6}$$

$$p(z^l_k|x^l_k, H, d, R) \triangleq \mathscr{N}(z^l_k; Hx^l_k + d, R) \tag{7}$$

$$p(x^l_k|x^l_{k-1}, F, u, Q) \triangleq \mathscr{N}(x^l_k; Fx^l_{k-1} + u, Q) \tag{8}$$

Derivatives of (5) for each unknown parameter, and roots of the equations that are obtained by setting the derivatives equal to zero are listed in Table 1. These roots are the estimation formulae of the unknown parameters.

3.1.2 *Maximum a Posteriori (MAP) Based Learning*

Maximum likelihood estimation of a GLDS tends to over-fit the training data, leading to degraded test-set performance. In order to improve generalizability of the learned parameters, we propose a regularized learning algorithm based on MAP (maximum *a posteriori*) learning. Specifically, we propose to impose a prior distribution $p(u, F, Q)$ that encourages the regression matrix, F, to take values slightly smaller (therefore slightly more generalizable [14]) than its maximum-likelihood values. In the maximum a posteriori learning criterion, the parameter set Θ is estimated based on training data set $D = \{X, Z\}$ and prior distribution $p(\Theta)$, therefore

In the maximum a posteriori learning criterion, the parameter set Θ is estimated based on the training data set D and prior distribution $p(\Theta)$, therefore

$$\hat{\Theta}^{MAP} = \arg\max_{\Theta} \ln L(\Theta) + \ln p(\Theta) \tag{9}$$

where, $L(\Theta)$ is the likelihood function defined in (4) and $p(\Theta)$ is the prior distribution for the model parameter set Θ, which can be defined as follows. In this work, the model parameter set is divided into two subsets $\Theta = \{\Theta_1, \Theta_2\}$, where $\Theta_1 = \{\bar{x}, \Sigma, H, d, R\}$ and $\Theta_2 = \{\mathbf{F}, Q\}$. where, $\mathbf{F} \triangleq [u, F]$ is the augmented parameter. Under the prior independence assumption, the joint prior density can be written as follows

$$p(\Theta) = p(\Theta_1)p(\Theta_2) \tag{10}$$

In this work prior density $p(\Theta_1)$ is assumed to be noninformative uniform prior, i.e. $p(\Theta_1) = constant$. Under this assumption, (9) reduces to the

$$\hat{\Theta}^{MAP} = \arg\max_{\Theta} \ln L(\Theta) + \ln p(\mathbf{F}, Q) \tag{11}$$

The joint prior distribution for $p(\mathbf{F}, Q)$ can be written as

$$p(\mathbf{F}, Q) = p(\mathbf{F}|Q)p(Q) \tag{12}$$

Now, we need to specify the prior distribution for $p(\mathbf{F}|Q)$ and $p(Q)$. For this purpose, the conjugate prior distributions are chosen. A prior distribution is said to be a conjugate prior distribution for a given model if the resulting posterior distribution is from the same family as the prior. The prior distribution $p(\mathbf{F}|Q)$ is the matrix normal distribution [15, 16] defined as

$$\begin{aligned} p(\mathbf{F}|Q) &\triangleq \mathcal{N}(\mathbf{F}; 0, Q, \Omega) \\ &\propto |\Omega^{-1}|^{\frac{n_x}{2}} |Q^{-1}|^{\frac{n_x+1}{2}} \exp\left(-\frac{1}{2} \operatorname{tr} \Omega^{-1} \mathbf{F}^T Q^{-1} \mathbf{F}\right) \end{aligned} \tag{13}$$

where, 0 is the mean of the matrix normal distribution. Ω and Q are two corresponding covariances. The prior distribution $p(Q)$ is the inverse Wishart distribution [15, 16] defined as follows

$$\begin{aligned} p(Q) &\triangleq \mathcal{W}^{-1}(Q; \Psi, v) \\ &\propto |Q^{-1}|^{\frac{v+n_x+1}{2}} \exp\left(-\frac{1}{2} \operatorname{tr} Q^{-1} \Psi\right) \end{aligned} \tag{14}$$

where,v and Ψ are the degrees of freedom and scale matrix for inverse Wishart distribution. Combining (13) and (14), the joint prior density $p(\mathbf{F}, Q)$ becomes

$$\begin{aligned} p(\mathbf{F}, Q) &\propto |\Omega^{-1}|^{\frac{n_x}{2}} |Q^{-1}|^{\frac{v+2n_x+2}{2}} \\ &\times \exp\left(-\frac{1}{2} \operatorname{tr}\left(\Omega^{-1} \mathbf{F}^T Q^{-1} \mathbf{F} + Q^{-1} \Psi\right)\right) \end{aligned} \tag{15}$$

Substituting (15) into (11) and rearranging, the maximization criterion of Θ_2 can be written as follows.

$$\begin{aligned} \hat{\Theta}^{MAP} = \arg\max_{\Theta} &\left\{ \ln L(\Theta) + \ln |\Omega^{-1}|^{\frac{n_x}{2}} ||Q^{-1}|^{\frac{v+2n_x+2}{2}} \right. \\ &\times \left. \exp\left(-\frac{1}{2} \operatorname{tr}\left(\Omega^{-1} \mathbf{F}^T Q^{-1} \mathbf{F} + Q^{-1} \Psi\right)\right) \right\} \end{aligned} \tag{16}$$

Table 2: MAP Based Parameter Estimation for GLDS

MAP based estimated parameters:

$$\hat{\mathbf{F}} \triangleq \left[\hat{u}, \hat{F}\right], \mathbf{x_k}^{l} \triangleq \left[1, (x_k^l)^T\right]^T, \Upsilon \triangleq \hat{\mathbf{F}} \Omega^{-1} \hat{\mathbf{F}}^T + \Psi$$

$$\begin{aligned} \hat{\mathbf{F}} = &\left(\textstyle\sum_{l=1}^{L} \sum_{k=2}^{N_l} x_k^l \mathbf{x}_{k-1}^l\right) \\ &\times \left(\textstyle\sum_{l=1}^{L} \sum_{k=2}^{N_l} \mathbf{x}_{k-1}^l \mathbf{x}_{k-1}^l + \Omega^{-1}\right)^{-1} \end{aligned}$$

$$\hat{Q} = \frac{\sum_{l=1}^{L} \sum_{k=2}^{N_l} (x_k^l - \hat{\mathbf{F}} \mathbf{x}_{k-1}^l)(x_k^l - \hat{\mathbf{F}} \mathbf{x}_{k-1}^l)^T + \Upsilon}{\sum_{l=1}^{L} \sum_{k=2}^{N_l} + v + 2n_x + 2}$$

Taking derivatives of (16) for each unknown parameter, and setting derivatives equal to zero estimation formulas can be obtained. These formulas are given in Table 2[1].

3.2 Inference

After the parameter learning stage, the filtered state $\hat{x}_{k|k}$ can be estimated by Kalman filter (KF) in a recursive manner. Smoothing is also a standard procedure in Kalman filtering. In this work we have obtained smoothed estimates by applying KF in forward and backward direction. Smoothed states $\hat{x}_{k|N}$ are obtained as a combination of forward and backward estimates. Kalman filtering and smoothing algorithms are described in [17] in detail.

4. EXPERIMENTS

4.1 Experimental Conditions

In this work, we use the MOCHA database [18]. The acoustic data and EMA trajectories of one female talker (fsew0) are used; these data include 460 sentences. Audio features (Mel-frequency cepstral coefficients (MFCC)) were computed using a 36 ms window with 18 ms shift. The articulatory data are EMA trajectories, which are the X and Y coordinates of the lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum and velum. EMA trajectories are normalized by the methods suggested in [1] and downsampled to match the 18 ms shift rate. All the model parameters of GLDS are tested using 10-fold cross-validation. For each fold, nine tenths of the data (414 sentences) are used for training and one tenth (46 sentences) for testing. Cross-validation performance measures (RMS error and correlation coefficient) are computed as the average of all ten folds.

4.2 Hyperparameter Assessment

The parameters of the prior distributions are called hyperparameters and they must be estimated to solve articulatory inversion problem. The hyperparameter set of the proposed model is $\theta_h = \{\Omega, \Psi, \alpha, v\}$. In this work, we choose the fol-

[1] Since the estimation formulae of the rest of the parameters are same as given in Table 1, we do not repeat them in Table 2

lowing parameters for joint prior distribution.

$$\Omega^{-1} = \alpha \sum_{l=1}^{L} \sum_{k=2}^{N_l} \mathbf{x}_{k-1}^{l} \mathbf{x}_{k-1}^{lT}$$

$$\Psi = \frac{1}{v} I_{n_x \times n_x}$$

where, $\mathbf{x_k}^l$ is the augmented articulatory state trajectories defined as $\mathbf{x_k}^l \triangleq \left[1,(x_k^l)^T\right]^T$ In this way, prior distribution becomes an invariant prior distribution [15]. Degree of freedom of the inverse Wishart distribution v is also fixed and is equal to the total number of observations. That is,

$$v = \sum_{l=1}^{L} N_l - 1.$$

Therefore, the only unknown hyperparameter is $\{\alpha\}$. The parameter α is estimated via trial and error. In this work, we test the values of $\alpha \in S = \{0.1, 0.3, 0.5\}$.

4.3 Performance Measures

The performance of the algorithms is measured using three performance measures, namely, RMS error, normalized RMS error and correlation coefficient, all of which are described in [1, 4, 10].

- RMS error:

$$E_{RMS}^{i} \triangleq \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_k^i - \hat{x}_k^i)^2}, \quad i = 1,\ldots,m$$

where x_k^i and $\hat{x}_k^i$ are true and estimated position, respectively, of the ith articulator in the kth frame.

- Normalized RMS error:

$$E_{NRMS}^{i} \triangleq \frac{E_{RMS}^{i}}{\sigma_i}, \quad i = 1,\ldots,m$$

where σ_i is the standard deviation of ith articulator x^i.

- Correlation coefficient:

$$\rho_{x,\hat{x}}^{i} \triangleq \frac{\sum_{k=1}^{N}(x_k^i - \bar{x}_k^i)(\hat{x}_k^i - \bar{\hat{x}}_k^i)}{\sqrt{\sum_{k=1}^{N}(x_k^i - \bar{x}_k^i)^2}\sqrt{\sum_{k=1}^{N}(\hat{x}_k^i - \bar{\hat{x}}_k^i)^2}}$$

for $i = 1,\ldots,m$ where $\bar{x}^i$ and $\bar{\hat{x}}^i$ are the average position of true and estimated ith articulator respectively.

4.4 Experimental Results

Experimental results of the proposed method are given in this sub-section. The comparison of the learning methods based on ML and MAP criteria for a single GLDS in terms of RMS error and correlation coefficient can be seen in Fig.1. Examination of the figure shows that the MAP based learning method significantly improves the performance of the articulatory inversion. The performance of the proposed algorithm is tested for various α values and it is observed that $\alpha = 0.3$ gives the best performance for MAP based learning algorithm. The RMS error and the correlation coefficient between the true (measured) and the estimated articulatory trajectories for filtering mode are about 2.34 mm and 0.52 for ML learning method while the corresponding results for MAP ($\alpha = 0.3$) based learning method are about 1.93 mm and 0.59 respectively. That means MAP based learning algorithm in filtering mode reduces RMS error about 17.5% (from 2.34 mm to 1.93 mm) and improve the correlation coefficients about 13.4% (from 0.52 to 0.59). When we consider inferences based on smoothed estimate, The RMS error and correlation coefficient for ML based learning method are about 2.3 mm and 0.59, corresponding results for MAP ($\alpha = 0.3$) based learning method are 1.85 mm and 0.65. That means MAP based learning algorithm in smoothing reduces RMS error about 18% (from 2.26 mm to 1.85 mm) and improve the correlation coefficients about 10% (from 0.59 to 0.65). The second observation from Fig.1 is that the smoothing is highly improves the performance compared to filtering. A similar result is reported in [3, 4] for articulatory inversion based on GMM and in [1] for articulatory inversion based on (TMDN). In our work smoothing reduces the RMSE from 2.34 mm to 2.26 mm (a 3.4% relative improvement) and improves the correlation coefficient about from 0.52 to 0.59 (a 13.4% relative improvement) when ML based learning is used. Similarly, smoothing reduces RMSE about 4.1% (from 1.93 mm to 1.85 mm) and improves correlation coefficient about 10.1% (from 0.59 to 0.65) for the MAP ($\alpha = 0.3$) based learning method.

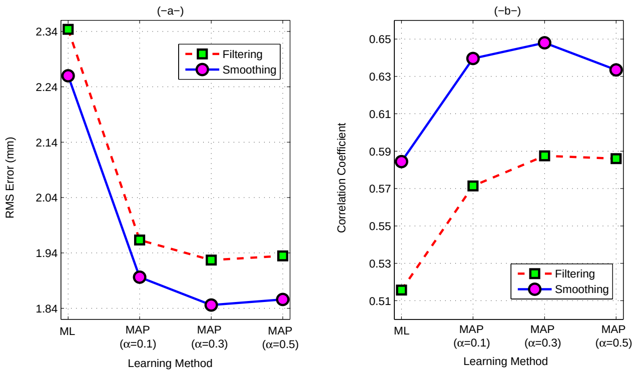

Figure 1: RMS error (-a-) and correlation coefficient (-b-) between true (measured) and estimated articulatory trajectories for ML and MAP (with various α values) learning, and corresponding filtered and smoothed estimation results.

Fig.2 provides more details regarding the utility of MAP based learning method in articulatory inversion. The abscissa distinguishes different articulators. As an example, in Fig.2, normalized RMS error for Y axis of upper lip (uly) reduced from 1.13 to 0.85. That means, there is a 24.7% relative error reduction. In general, this figure denotes that MAP based learning algorithm reduces normalized RMS about (14-28%). Fig.3 illustrates an example of the estimated (based on ML and MAP ($\alpha = 0.3$) and the true x-coordinates of articulatory trajectories for lower incisor. The utterance is taken form MOCHA database.

5. CONCLUSION AND FUTURE WORK

In this paper, we have examined parameter estimation methods that are based on ML and MAP criteria for acoustic-to-articulatory inversion which is done by using a single global linear dynamic system (GLDS). The main aim of this work

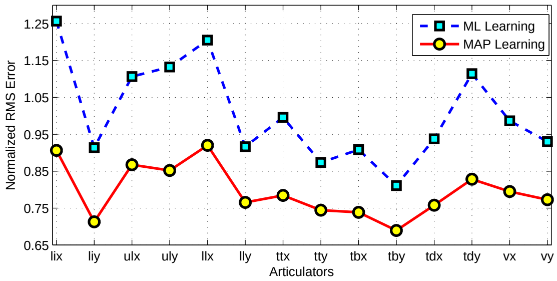

Figure 2: Normalized RMS errors for each articulator for ML and MAP ($\alpha = 0.3$) learning methods. The abbreviations li, ul,ll, tt,tb, td and v denote lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum and velum, respectively. The suffixes x and y of the articulator abbreviations show the corresponding X and Y coordinates respectively.

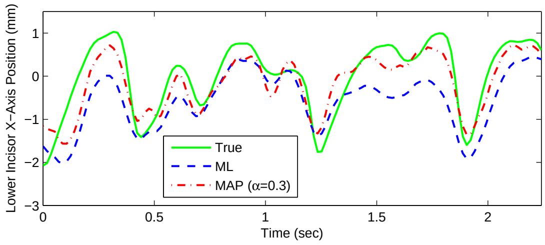

Figure 3: An example of estimated (based on ML and MAP method) and true articulatory trajectories of x-coordinate for lower incisor.(taken from MOCHA database)

is to show that MAP based parameter estimation method significantly improves the performance of the articulatory inversion. Experiments have been conducted on the MOCHA databases. In the literature, [11] uses a single GLDS in articulatory inversion with similar experimental setup. They estimate the model parameter set via ML criterion and their best RMSE and correlation coefficient results between the true (measured) and estimated articulatory trajectories are 2.15 mm and 0.59 respectively. According to the experimental results given in Sec. 4.4, our best results are obtained via MAP based learning algorithm by using smoothing method. The RMS error and correlation coefficient are about 1.85 mm and 0.65 respectively, which is significantly better than the results of [11]. The main reason of the improvement is using MAP based learning instead of ML and also smoothing instead of filtering.

Our future work plan is to generalize the single GLDS to a multiple models dynamic system which is known as jump Markov linear system (JMLS) or switching linear dynamic system (SLDS) and use MAP based learning. The preliminary experimental results showed that MAP based learning algorithm is also improves the performance of articulatory inversion when multiple models are used.

REFERENCES

[1] K. Richmond, "Estimating articulatory parameters from the speech signal," Ph.D. dissertation, The Center for Speech Technology Research, Edinburgh, KU, 2002.

[2] C.Qin and M. Carreira-Perpin, "A comparison of acoustic features for articulatory inversion," in *Proc. INTERSPEECH*, 2007.

[3] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, pp. 215–227, 2008.

[4] İ. Y. Özbek, M. Hasegawa-Johnson, and M. Demirekler, "Formant trajectories for acoustic-to-articulatory inversion," in *Proc. INTERSPEECH*, 2009.

[5] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production models," *IEEE Trans. Sp. Au. Proces*, vol. 12, no. 2, p. 175185, March 2004.

[6] A. Toutios and K. Margaritis, "Contribution to statistical acoustic-to-EMA mapping," in *EUSIPCO*, 2008.

[7] V. Mitra, İ. Y. Özbek, H. Nam, X. Zhou, and C. Espy-Wilson, "From acoustic to vocal tract time functions," in *ICASSP*, 2009.

[8] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 444–460, 2005.

[9] H. Kjellström and O. Engwall, "Audiovisual-to-articulatory inversion," *Speech Communication*, vol. 51, no. 3, pp. 195–209, 2009.

[10] A. Katsamanis, G. Papandreou, and P. Maragos, "Face active appearance modeling and speech acoustic information to recover articulation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 17, no. 3, pp. 411–422, 2009.

[11] A. Katsamanis, G. Ananthakrishnan, G. Papandreou, P. Maragos, and O. Engwall, "Audiovisual speech inversion by switching dynamical modeling governed by a hidden markov process," in *Proc. EUSIPCO*, 2008.

[12] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *5'th Seminar on Speech Production: Models and Data, Kloster Seeon,Germany*, 2000, pp. 237–240.

[13] L. J. Lee, P. Fieguth, and L. Deng, "A functional articulatory dynamic model for speech production," in *Proc. ICASSP*, 2001, pp. 797–800.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1996.

[15] T. P. Minka, "Bayesian linear regression," 3594 Security Ticket Control, Tech. Rep., 1999.

[16] D. B. Rowe, *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*. New York: CRC Press Company, 2003.

[17] D. Simon, *Optimal State Estimation: Kalman, $\mathscr{H}_\infty$, and Nonlinear Approaches*. Wiley, 2006.

[18] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *In Proc. 5th Seminar on Speech Production*, 2000, [Online]. Available: http://www.cstr.ed.ac.uk/artic.