INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Acoustic-to-Articulatory Inversion Mapping based on Latent Trajectory Gaussian Mixture Model

Patrick Lumban Tobing[1], *Tomoki Toda*[2], *Hirokazu Kameoka*[3], *Satoshi Nakamura*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]Information Technology Center, Nagoya University, Japan
[3]NTT Communication Science Laboratories, NTT Corporation, Japan

patrick.lumbantobing.pf3@is.naist.jp, tomoki@icts.nagoya-u.ac.jp,
kameoka.hirokazu@lab.ntt.co.jp, s-nakamura@is.naist.jp

Abstract

A maximum likelihood parameter trajectory estimation based on a Gaussian mixture model (GMM) has been successfully implemented for acoustic-to-articulatory inversion mapping. In the conventional method, GMM parameters are optimized by maximizing a likelihood function for joint static and dynamic features of acoustic-articulatory data, and then, the articulatory parameter trajectories are estimated for given the acoustic data by maximizing a likelihood function for only the static features, imposing a constraint between static and dynamic features to consider the inter-frame correlation. Due to the inconsistency of the training and mapping criterion, the trained GMM is not optimum for the mapping process. This inconsistency problem is addressed within a trajectory training framework, but it becomes more difficult to optimize some parameters, e.g., covariance matrices and mixture component sequences. In this paper, we propose an inversion mapping method based on a latent trajectory GMM (LT-GMM) as yet another way to overcome the inconsistency issue. The proposed method makes it possible to use a well-formulated algorithm, such as EM algorithm, to optimize the LT-GMM parameters, which is not feasible in the traditional trajectory training. Experimental results demonstrate that the proposed method yields higher accuracy in the inversion mapping compared to the conventional GMM-based method.

Index Terms: Gaussian mixture model, inversion mapping, latent trajectory model, EM algorithm, inter-frame correlation

1. Introduction

Speech can be characterized not only by the acoustic spectrum of the vocal tract but also by the more slowly varying parameters, such as articulatory movements [1]. Indeed, utilization of underlying articulatory movements from speech sounds has been studied in many applications, e.g. speech analysis/synthesis [2, 3], speech coding [4], speech recognition [5, 6], speech pathology [7, 8], speech visualization [9, 10], etc. Hence, the need of a robust system to convert acoustic data into articulatory data, so called inversion mapping, grows rapidly.

Lately, the advancement in recording devices that enable a simultaneous recording procedure of acoustic-articulatory data has been inducing many works on statistical methods of acoustic-to-articulatory inversion mapping. Codebook based inversion mapping has been proposed in [11, 12]. Moreover, in [12], it has been reported that the accuracy of inversion mapping is significantly improved by introducing dynamic constraints. Neural network based inversion mapping has been proposed in [13], stating the importance of multiple mixtures of probability density of articulatory parameters. Hidden Markov model (HMM) based inversion mapping has been proposed in [14], incorporating dynamic constraints and also linguistic contexts.

In this paper, we focus on the study of Gaussian mixture model (GMM) based inversion mapping [15], capable of representing multiple mixtures of probability density and incorporating dynamic constraints without textual input. Albeit, in terms of dynamic constraints, this method bears an inconsistency issue. In its training phase, the model parameters are optimized with respect to the likelihood of joint static and dynamic features of acoustic-articulatory data. While in the mapping phase, given a sequence of acoustic parameters, a sequence of articulatory parameters is estimated with respect to the conditional likelihood of their static features, where the inter-frame correlation is considered by imposing a constraint between static and dynamic features. Because of the inconsistency of optimized feature space between the training and mapping phases, the trained GMM is not optimum for the mapping phase.

To overcome this inconsistency problem, a trajectory training method has been proposed, known as the trajectory HMM [16]. To train the trajectory model, optimization is performed with respect to the likelihood of the static features of the training data. This likelihood is obtained by imposing a constraint between static and dynamic features, making it possible to incorporate inter-frame correlation in the training phase. It has been reported that this method can be further extended to incorporate additional constraints, e.g. global variance [17] and modulation spectrum [18]. However, in this standard trajectory training method, it is somewhat difficult to optimize several model parameters, e.g. 1) it is not straightforward to optimize the covariance matrices due to the difficulties of obtaining their analytical ML estimates and 2) it is basically difficult to include all possible mixture component sequences or even to decide the best mixture component sequence to maximize the likelihood.

In this paper, as an alternative approach for solving the inconsistency issue, inspired by the latent trajectory HMM [19], we propose a novel inversion mapping method based on latent trajectory GMM (LT-GMM). In the proposed method, a time sequence of joint static and dynamic features is treated as a latent variable, while a time sequence of static features is treated as an observed variable. A likelihood function of the observed variable is obtained through marginalizing out the latent variable by imposing a soft constraint between static and dynamic features. Based on such likelihood function, parameters are easily optimized by using the EM algorithm. Experimental results indicate that the proposed LT-GMM gives higher accuracy for the inversion mapping compared with the conventional GMM.

http://dx.doi.org/10.21437/Interspeech.2016-1196

2. Conventional GMM-based acoustic-to-articulatory inversion mapping

Let us assume a time sequence of D_x-dimensional static acoustic feature vectors as $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots \boldsymbol{x}_T^\top]^\top$ and that of D_y-dimensional static articulatory feature vectors as $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top]^\top$. At frame t, a $2D_x$-dimensional acoustic feature vector is denoted as $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$, a $2D_y$-dimensional articulatory feature vector is denoted as $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$, where $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$ denote the dynamic feature vector of acoustic parameters and that of articulatory parameters, respectively, and their joint feature vector is denoted as $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$. Then, time sequences of acoustic feature vectors, articulatory feature vectors, and their joint feature vectors are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$, $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$, and $\boldsymbol{Z} = [\boldsymbol{Z}_1^\top, \cdots, \boldsymbol{Z}_T^\top]^\top$, respectively.

2.1. Training phase

The joint probability density of the acoustic and articulatory feature vectors is modeled by a GMM as follows:

$$P(\boldsymbol{Z}|\boldsymbol{\lambda}^{(Z)}) = \prod_{t=1}^{T} \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma}_m^{(Z)}), \quad (1)$$

where the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The mixture component index is m. The total number of mixture components is M. The weight of the mth mixture component is α_m. The set of GMM parameters is $\boldsymbol{\lambda}^{(Z)}$, consisting of weights, mean vectors and covariance matrices of all mixture components. The mean vector $\boldsymbol{\mu}_m^{(Z)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(Z)}$ of the mth mixture component are written as

$$\boldsymbol{\mu}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \quad (2)$$

where the mean vectors of the acoustic parameters and that of the articulatory parameters for the mth mixture component are denoted as $\boldsymbol{\mu}_m^{(X)}$ and $\boldsymbol{\mu}_m^{(Y)}$, respectively. The covariance matrices of the acoustic parameters and that of the articulatory parameters for the mth mixture component are denoted as $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$, respectively. The cross covariance matrices of the acoustic and the articulatory parameters for the mth mixture component are denoted as $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$. These model parameters are optimized with EM algorithm [20].

2.2. Mapping phase

In the conventional GMM-based mapping, given an acoustic parameter sequence $\boldsymbol{X}$, the estimated articulatory parameter sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}), \text{ s.t. } \boldsymbol{Y} = \boldsymbol{W}_y \boldsymbol{y}, \quad (3)$$

where $\boldsymbol{W}_y$ is a transformation matrix to expand a sequence of static feature vectors into its sequence of joint static and dynamic feature vectors. The ML estimate can be determined with EM algorithm [15].

In this paper, an approximation of the above conditional probability density is employed by using a single mixture component sequence $\boldsymbol{m} = \{m_1, \cdots, m_T\}$. First, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined as follows:

$$\hat{\boldsymbol{m}} = \arg\max_{\boldsymbol{m}} \prod_{t=1}^{T} P(m_t|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}). \quad (4)$$

Then, the estimated articulatory parameter sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(Z)}), \text{ s.t. } \boldsymbol{Y} = \boldsymbol{W}_y \boldsymbol{y}, \quad (5)$$

where the ML estimate can be analytically determined [21].

Note that, in this mapping phase, the inter-frame correlation is taken into consideration, thanks to the usage of the constraint between static and dynamic features, i. e. $\boldsymbol{Y} = \boldsymbol{W}_y \boldsymbol{y}$. On the other hand, this constraint is neglected while optimizing the GMM parameters in the training phase.

3. Proposed LT-GMM-based acoustic-to-articulatory inversion mapping

Let the observed variable be a time sequence of joint static feature vectors $\boldsymbol{z} = [\boldsymbol{z}^\top, \cdots, \boldsymbol{z}_T^\top]^\top$, where $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$, and the latent variable be a time sequence of joint static and dynamic feature vectors $\boldsymbol{Z}$. The observed and latent variables are related through a soft constraint as follows:

$$\boldsymbol{Z} \simeq \boldsymbol{W}\boldsymbol{z} = [\boldsymbol{W}_x, \boldsymbol{W}_y][\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top. \quad (6)$$

Employing the above soft constraint, the conditional probability density of the latent variable $\boldsymbol{Z}$ given the observed variable $\boldsymbol{z}$ is written as

$$P(\boldsymbol{Z}|\boldsymbol{z}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{Z}; \boldsymbol{W}\boldsymbol{z}, \boldsymbol{\Sigma}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{z})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{z})\right\}, \quad (7)$$

where its covariance matrix ($\boldsymbol{\Sigma} = \mathrm{diag}\,[\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y]$) represents the variations of margin errors, compensating the soft constraint. By completing the square in the exponent of the above pdf, the conditional probability density of the observed variable $\boldsymbol{z}$ given the latent variable $\boldsymbol{Z}$ is written as

$$P(\boldsymbol{z}|\boldsymbol{Z}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\Lambda}^{-1}\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Z}, \boldsymbol{\Lambda}^{-1}), \quad (8)$$

where

$$\boldsymbol{\Lambda} = \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}. \quad (9)$$

3.1. Training phase

The joint probability density of the acoustic and articulatory feature vector sequences is modeled by an LT-GMM as follows:

$$P(\boldsymbol{z}|\boldsymbol{\lambda}^{(z)}) = \int P(\boldsymbol{z}|\boldsymbol{Z}, \boldsymbol{\Sigma}) P(\boldsymbol{Z}|\boldsymbol{\lambda}^{(Z)}) \mathrm{d}\boldsymbol{Z}, \quad (10)$$

where $\boldsymbol{\lambda}^{(z)}$ is the set of LT-GMM parameters, consisting of the covariance matrix $\boldsymbol{\Sigma}$ and model parameters $\boldsymbol{\lambda}^{(Z)}$. Note that the covariance matrix $\boldsymbol{\Sigma}$ depends on only dimension of $\boldsymbol{Z}_t$, i. e. independent of both time frames and mixture components. The model parameters can be optimized with variational EM algorithm [19].

In this paper, an approximation of the above joint pdf is employed by using the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$. The approximated joint pdf is then written as follows:

$$P(\boldsymbol{z}|\hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(z)}) = \int P(\boldsymbol{z}|\boldsymbol{Z}, \boldsymbol{\Sigma}) P(\boldsymbol{Z}|\hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(Z)}) \mathrm{d}\boldsymbol{Z} = \int \mathcal{N}\left(\begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{Z} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(z|Z)} \\ \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Z)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(zz)} & \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(zZ)} \\ \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(Zz)} & \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(ZZ)} \end{bmatrix}\right) \mathrm{d}\boldsymbol{Z}, \quad (11)$$

where the mean vectors are written as

$$\boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(z|Z)} = \boldsymbol{\Lambda}^{-1}\boldsymbol{W}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Z)}, \tag{12}$$

$$\boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Z)} = \left[\boldsymbol{\mu}_{\hat{m}_1}^{(Z)^{\top}}, \cdots, \boldsymbol{\mu}_{\hat{m}_T}^{(Z)^{\top}}\right]^{\top}, \tag{13}$$

and the covariance matrices are written as

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(zz)} = \boldsymbol{\Lambda}^{-1}\boldsymbol{W}^{\top}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(ZZ)}\boldsymbol{\Sigma}^{-1})\boldsymbol{W}\boldsymbol{\Lambda}^{-1}, \tag{14}$$

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(zZ)} = \boldsymbol{\Lambda}^{-1}\boldsymbol{W}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(ZZ)}, \tag{15}$$

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(Zz)} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(ZZ)}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}, \tag{16}$$

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(ZZ)} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(Z)} = \mathrm{diag}\left[\boldsymbol{\Sigma}_{\hat{m}_1}^{(Z)}, \cdots, \boldsymbol{\Sigma}_{\hat{m}_T}^{(Z)}\right]. \tag{17}$$

Using the likelihood in Eq. (10), we maximize an auxiliary function defined as

$$Q(\boldsymbol{\lambda}^{(Z)}, \hat{\boldsymbol{\lambda}}^{(Z)}) = \int P(\boldsymbol{Z}|\boldsymbol{z}, \boldsymbol{\lambda}^{(z)}) \log\left[P(\boldsymbol{z}|\boldsymbol{Z}, \boldsymbol{\Sigma})P(\boldsymbol{Z}|\hat{\boldsymbol{\lambda}}^{(Z)})\right]\mathrm{d}\boldsymbol{Z}. \tag{18}$$

Considering both the terms that depend on only the model parameters $\boldsymbol{\lambda}^{(Z)}$ and the use of the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$, the auxiliary function can be written as

$$\begin{aligned} Q(\boldsymbol{\lambda}^{(Z)}, \hat{\boldsymbol{\lambda}}^{(Z)}) &\propto \int P(\boldsymbol{Z}|\boldsymbol{z}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(z)}) \log P(\boldsymbol{Z}|\hat{\boldsymbol{m}}, \hat{\boldsymbol{\lambda}}^{(Z)})\mathrm{d}\boldsymbol{Z} \\ &= -\frac{1}{2}\sum_{t=1}^{T} \log\left|\hat{\boldsymbol{\Sigma}}_{\hat{m}_t}^{(Z)}\right| + \mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_{\hat{m}_t}^{(Z)^{-1}}\boldsymbol{R}_t\right) - \boldsymbol{r}_t^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{m}_t}^{(Z)^{-1}}\hat{\boldsymbol{\mu}}_{\hat{m}_t}^{(Z)} \\ &\quad - \hat{\boldsymbol{\mu}}_{\hat{m}_t}^{(Z)^{\top}}\hat{\boldsymbol{\Sigma}}_{\hat{m}_t}^{(Z)^{-1}}\boldsymbol{r}_t + \hat{\boldsymbol{\mu}}_{\hat{m}_t}^{(Z)^{\top}}\hat{\boldsymbol{\Sigma}}_{\hat{m}_t}^{(Z)^{-1}}\hat{\boldsymbol{\mu}}_{\hat{m}_t}^{(Z)}. \end{aligned} \tag{19}$$

In the expectation step (E-step), the expected values of the latent variable, i.e. $\boldsymbol{r} = [\boldsymbol{r}_1^{\top}, \cdots, \boldsymbol{r}_T^{\top}]^{\top}$ and $\boldsymbol{R} = \mathrm{diag}\,[\boldsymbol{R}_1, \cdots, \boldsymbol{R}_T]$, are estimated as follows:

$$\begin{aligned} \boldsymbol{r} &= \mathbb{E}[\boldsymbol{Z}|\boldsymbol{z}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(z)}] \\ &= \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Z)} + \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(Zz)}\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(zz)^{-1}}\left(\boldsymbol{z} - \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(z|Z)}\right), \end{aligned} \tag{20}$$

$$\begin{aligned} \boldsymbol{R} &= \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^{\top}|\boldsymbol{z}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(z)}] \\ &= \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(ZZ)} - \underset{2D}{\mathrm{diag}}\left(\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(Zz)}\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(zz)^{-1}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(zZ)} + \boldsymbol{r}\boldsymbol{r}^{\top}\right), \end{aligned} \tag{21}$$

where diag_{2D} yields a block diagonal matrix with each block having the size of $2D = 2(D_x + D_y)$.

In the maximization step (M-step), by maximizing the auxiliary function with respect to the updated mean vector $\hat{\boldsymbol{\mu}}_m^{(Z)}$ for the mth mixture component, its ML estimate is given by

$$\hat{\boldsymbol{\mu}}_m^{(Z)} = \frac{1}{\gamma_m}\sum_{t=1}^{T}\delta(\hat{m}_t = m)\boldsymbol{r}_t, \tag{22}$$

and by maximizing the auxiliary function with respect to the updated covariance matrix $\hat{\boldsymbol{\Sigma}}_m^{(Z)}$ for the mth mixture component, its ML estimate is given by

$$\hat{\boldsymbol{\Sigma}}_m^{(Z)} = \frac{1}{\gamma_m}\sum_{t=1}^{T}\delta(\hat{m}_t = m)\boldsymbol{R}_t - \hat{\boldsymbol{\mu}}_m^{(Z)}\hat{\boldsymbol{\mu}}_m^{(Z)^{\top}}, \tag{23}$$

where the total number of frames γ_m belonging to the mth mixture component is given by

$$\gamma_m = \sum_{t=1}^{T}\delta(\hat{m}_t = m), \tag{24}$$

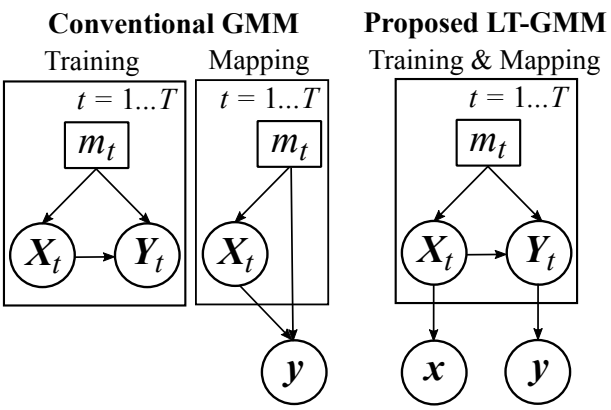

Figure 1: *Simplified graph representations of the training and the mapping phases for the conventional GMM and the proposed latent trajectory GMM (LT-GMM).*

and

$$\delta(\hat{m}_t = m) = \begin{cases} 1, & \text{if true} \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

3.2. Mapping phase

In the proposed LT-GMM-based mapping, given an acoustic feature sequence $\boldsymbol{x}$, the estimated articulatory feature sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda}^{(z)}), \tag{26}$$

where the conditional pdf is written as

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda}^{(z)}) = \int P(\boldsymbol{y}|\boldsymbol{Y}, \boldsymbol{\Sigma})\int P(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{x}, \boldsymbol{\lambda}^{(z)})\mathrm{d}\boldsymbol{X}\mathrm{d}\boldsymbol{Y}. \tag{27}$$

This ML estimate can be determined with variational EM algorithm.

In this paper, an approximation of the above conditional pdf is employed by first determining the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$. Then, the estimated articulatory feature sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{y} P(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(z)}), \tag{28}$$

where the approximated conditional pdf is written as

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(z)}) &= \int P(\boldsymbol{x}|\boldsymbol{Y}, \boldsymbol{\Sigma})\int P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(Z)}) \\ &\qquad P(\boldsymbol{X}|\boldsymbol{x}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(z)})\mathrm{d}\boldsymbol{X}\mathrm{d}\boldsymbol{Y} \\ &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(y|x)}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(y|x)}), \end{aligned} \tag{29}$$

where

$$\boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(y|x)} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(y|x)}\boldsymbol{W}_y^{\top}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Y|x)}, \tag{30}$$

$$\boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Y|x)} = \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Y)} + \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(YX)}\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(XX)^{-1}}(\mathbb{E}[\boldsymbol{X}|\boldsymbol{x}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(z)}] - \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(X)}), \tag{31}$$

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(y|x)} = (\boldsymbol{W}_y^{\top}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{W}_y)^{-1}. \tag{32}$$

Thus, it is apparent that the ML estimate of the articulatory feature sequence $\hat{\boldsymbol{y}}$ is the mean vector of the above conditional pdf, i.e. $\boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(y|x)}$, which can be analytically determined.

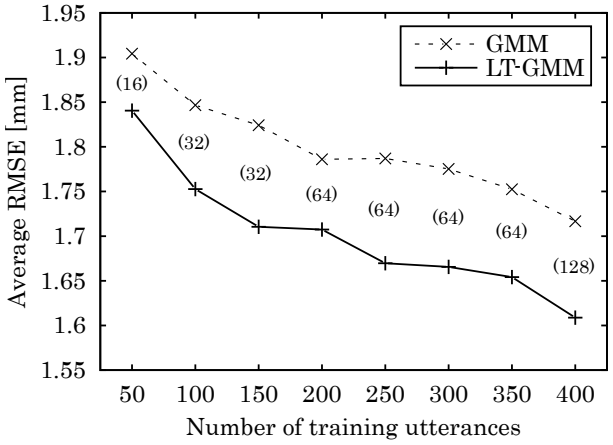

Figure 2: *Average root-mean-square error of estimated articulatory data. Numbers in parentheses are optimum number of mixture components.*

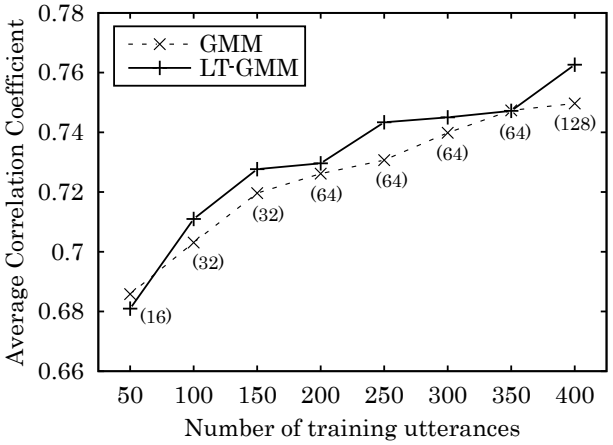

Figure 3: *Average correlation coefficient of estimated articulatory data. Numbers in parentheses are optimum number of mixture components.*

Note that, in the LT-GMM, the inter-frame correlation is well considered in both training and mapping phases by incorporating the soft constraint in Eq. (6). Figure 1 shows the difference of consistency of the training and mapping phases between the conventional GMM and the proposed LT-GMM.

4. Experimental evaluation

4.1. Experimental conditions

We used a set of simultaneously recorded speech and articulatory data provided in MOCHA [22], spoken by a single British male speaker. The total number of utterances was 460. Speech data was sampled at 16 kHz. EMA data was used as the articulatory data. The frame shift was set to 5 ms.

As the acoustic parameters, we used the 1^{st} through 24^{th} mel-cepstral coefficients converted from the spectral envelope, which was extracted at each frame by using STRAIGHT analysis method [23]. As the articulatory parameters, we used the 14-dimensional EMA data, which were converted to z-scores. The EMA data represented the movements of seven articulators, i. e. upper lip, lower lip, lower incisor, tongue tip, tongue body, tongue dorsum, and velum, defined within x- and y- coordinates on the midsagittal plane.

The constant positive-definite matrix $\mathbf{\Sigma}$ was set to the diagonal matrix of global variances. A trained model from the conventional GMM was used as an initial model for the LT-GMM training. In the training phase of LT-GMM, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$ was initialized using the initial model and held fixed. In the mapping phase of LT-GMM, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$ was determined using also the initial model.

We conducted an objective evaluation by calculating the root-mean-square errors (RMSEs) and the correlation coefficients between the estimated articulatory parameters and the measured ones. The number of training utterances was varied to 50, 100, 150, 150, 200, 250, 300, 350 and 400. The number of mixture components was varied to 16, 32, 64 and 128. Optimum number of mixture components for each number of training utterances was determined by using the conventional GMM, as given in both Fig. 2 and Fig. 3. The number of testing utterances was 20.

In order to efficiently perform the training process of LT-GMM, we exploited the structure of banded matrix, which is implied at the huge linear equations in Eq. (20) and Eq. (21). Specifically, we used the multifrontal massively parallel sparse direct solver (MUMPS) [24, 25]. The computational time for training the LT-GMM is about 15 times slower compared to the conventional GMM.

4.2. Experimental results

Figures 2 and 3 show the average values of the RMSE and that of the correlation coefficient over all 14 dimensions of estimated articulatory parameters through all 20 testing utterances. The proposed LT-GMM method gives lower values than the conventional GMM in terms of average RMSE. The LT-GMM method gives overall higher values for also the average correlation coefficient compared to the conventional GMM. These results indicate that the proposed LT-GMM method improves the accuracy of the acoustic-to-articulatory inversion mapping. Such good performance can be achieved because of the consideration of the inter-frame correlation while optimizing the LT-GMM parameters, whereas in the conventional GMM, this consideration is taken into account only in the mapping phase.

5. Conclusions

This paper presents an acoustic-to-articulatory inversion mapping system based on latent trajectory Gaussian mixture model (LT-GMM). In the proposed LT-GMM method, the consistency between training and mapping phases is preserved by imposing a soft constraint between static and dynamic features, where the inter-frame correlation is taken into account. In the training phase, the parameters can be conveniently optimized by using EM algorithm. The experimental results demonstrate that the proposed LT-GMM method improves the accuracy of the inversion mapping compared to the conventional GMM. For the future work, we plan to optimize also the mixture component sequence and incorporate acoustic segment feature consisting of multiple frames of input features.

6. Acknowledgements

This research was supported in part by JSPS KAKENHI Grant Number 26280060.

7. References

[1] S. Parthasarathy, J. Schroeter, C. Coker, and M. M. Sondhi, "Articulatory analysis and synthesis of speech," In *Fourth IEEE Region 10 International Conference*, pp. 760–764, Bombay, India, Nov. 1989.

[2] Z. -H. Ling, K. Richmond, J. Yamagishi, and R. -H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*. Vol. 17, No. 6, pp. 1171–1185, 2009,

[3] P. L. Tobing, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Articulatory controllable speech modification based on Gaussian mixture models with direct waveform modification using spectrum differential," In *Proc. INTERSPEECH*, pp. 3350–3354, Dresden, Germany, Sep. 2015.

[4] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," In *Advances in Speech Signal Processing*, ed. by S. Furui and M. M. Sondhi, Marcel Dekker, New York, pp. 231–267, 1992.

[5] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," In *Proc. ICSLP*, Vol. 4, pp. 145–148, Beijing, China, Oct. 2000.

[6] J. Frankel, K. Richmond, S. King, and P. Taylor, "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces," In *Proc. ICSLP*, Vol. 4, pp. 254–257, Beijing, China, Oct. 2000.

[7] H. Ackermann and W. Ziegler, "Articulatory deficits in parkinsonian dysarthria: an acoustic analysis," *Journal of Neurology, Neurosurgery & Psychiatry*, Vol. 54, No. 12, pp. 1093–1098, 1991.

[8] H. Ackermann, I. Hertrich, I. Daum, G. Scharf, and S. Spieker, "Kinematic analysis of articulatory movements in central motor disorders," *Movement Disorders*, Vol. 12, No. 6, pp. 1019–1027, 1997.

[9] M. Cohen, D. W. Massaro, "Modeling coarticulation in synthetic visual speech," In *Models and Techniques in Computer Animation*, Springer, 1993, pp. 1019–1027.

[10] I. Steiner, K. Richmond, and S. Ouni, "Using multimodal speech production data to evaluate articulatory animation for audiovisual speech synthesis," In *Proc. The 3rd Symposium on Facial Analysis and Animation (FAA2012)*, No. 2, p. 1, Vienna, Austria, Sept. 2012.

[11] 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," *Journal of the Acoustical Society of America*, Vol. 100, No. 3, pp. 1819–1834, 1996.

[12] S. Suzuki, S. Okadome, T. Honda, "Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints," In *Proc. ICSLP*, pp. 2251–2254, Sydney, Australia, 1998.

[13] K. Richmond, K. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, Vol. 17, No. 2, pp. 153–172, 2003.

[14] S. Hiroya, and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Transactions on Speech and Audio Processing*, Vol. 12, No. 2, pp. 175–185, 2004.

[15] 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, No. 3, pp. 215–227, 2008.

[16] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationship between static and dynamic feature vector sequences," *Computer Speech and Language*, Vol. 21, No. 1, pp. 760–764, 2007.

[17] T. Toda, and S. Young, "Trajectory training considering global variance for HMM-based speech synthesis," In *Proc. ICCASP*, pp. 4025–4028, Taipei, Taiwan, Aug, 2009.

[18] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion," In *Proc. ICCASP*, pp. 4859–4863, Brisbane, Australia, Apr., 2015.

[19] H. Kameoka, "Modeling speech parameter sequences with latent trajectory hidden Markov model," In *Proc. The 25th IEEE International Workshop on Machine Learning for Signal Processing (MLSP2015)*, pp. 1–6, Boston, USA, Sept. 2015.

[20] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72–83, 1995.

[21] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[22] A. Wrench, "The MOCHA-TIMIT articulatory database," http://www.cstr.ed.ac.uk/artic/mocha.html, Queen Margaret University College, 1999.

[23] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representation using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, Vol. 7, No. 3–4, pp. 187–207, 1999.

[24] P. R. Amestoy, I. S. Duff, J. Koster, and J. -Y. L'Excellent, "A fully asynchronous multifrontal solver using distributed dynamic scheduling," *SIAM Journal on Matrix Analysis and Application*, Vol. 23, No. 1, pp. 15–41, 2001.

[25] P. R. Amestoy, A. Guermoche, J. -Y. L'Excellent, and S. Pralet, "Hybrid scheduling for the parallel solution of linear systems," *Parallel Computing*, Vol. 32, No. 2, pp. 136–156, 2006.