INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Acoustic-to-articulatory mapping based on mixture of probabilistic canonical correlation analysis

Hidetsugu Uchida, Daisuke Saito, Nobuaki Minematsu

The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
{uchida,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, we propose a novel acoustic-to-articulatory mapping model based on mixture of probabilistic canonical correlation analysis (mPCCA). In PCCA, it is assumed that two different kinds of data are observed as results from different linear transforms of a common latent variable. It is expected that this variable represents a common factor which is inherent in the different domains, such as acoustic and articulatory feature spaces. mPCCA is an expansion of PCCA and it can model a much more complex structure. In mPCCA, covariance matrices of a joint probabilistic distribution of acoustic-articulatory data are structuralized reasonably by using transformation coefficients of the linear transforms. Even if the number of components in mPCCA increases, the structuralized covariance matrices can be expected to avoid over-fitting. Training and mapping processes of the mPCCA-based mapping model are reasonably derived by using the EM algorithm. Experiments using MOCHA-TIMIT show that the proposed mapping method has achieved better mapping performance than the conventional GMM-based mapping.

Index Terms: acoustic-to-articulatory mapping, mixture of probabilistic canonical correlation analysis, Gaussian mixture model

1. Introduction

Acoustic-to-articulatory (a2a) mapping has drawn many researchers' attention as a technique to estimate articulatory movements from acoustic observations. Many a2a mapping techniques have been studied, and those can be grouped into two kinds of statistical models, generative model and discriminative model. Hidden Markov model (HMM)-based mapping technique [1] and Gaussian mixture model (GMM)-based mapping technique [2] are good examples of the former, and deep neural networks (DNN)-based mapping technique [3] is that of the latter. Both of the two kinds are developed in common with acoustic-articulatory data which are observed simultaneously. Based on the data, HMM- or GMM-based mapping techniques represent relationships between acoustic and articulatory parameters by several linear transforms. The linear transforms of HMM are constructed for each state and those of GMM are constructed for each Gaussian distribution. On the other hand, DNN-based mapping techniques represent those by a stack of multiple non-linear transforms. Compared with DNN, GMM and HMM require less training data to develop the mapping models than DNN and their parameters have clear meanings to be interpreted.

In GMM-based mapping, a joint probability of acoustic and articulatory features is modeled as a weighted sum of Gaussian distributions. The more the total number of Gaussian distributions (i.e. components) increases, the more complicated structure the model can characterize. However, increase of the total number of components may result in over-fitting. In voice conversion, to avoid over-fitting, matrix variate Gaussian mixture model (MVGMM) -based mapping technique has been studied [4]. Covariance matrices of MVGMM are structuralized reasonably and the total number of parameters of those is smaller than that of GMM. Therefore, MVGMM-based mapping technique does not suffer from over-fitting even if the total number of components increases. However, MVGMM cannot be applied to a2a mapping because it requires inputs and outputs of mapping to be in the same feature domain, such as acoustic-to-acoustic.

In this paper, we propose a mapping model which has covariance matrices structuralized reasonably like MVGMM but can be applied to a2a mapping. The proposed model is based on mixture of probabilistic canonical correlation analysis (mPCCA) [5]. In PCCA, it is assumed that two different kinds of data are observed as results from different linear transforms of a common latent variable [6]. It is expected that this variable represents a common factor which is inherent in the different domains, such as acoustic and articulatory feature spaces. mPCCA is an extension of PCCA and it can model a much more complex structure. In mPCCA, covariance matrices of a joint probabilistic distribution of acoustic-articulatory data are structuralized reasonably by using transformation coefficients of the linear transforms. Since mPCCA does not require the two kinds of data to be in the same feature domain, it can be applied to a2a mapping.

2. Proposed method

2.1. Mixture of probabilistic canonical correlation analysis

Figure 1 shows a graphical model of mPCCA. Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ are observed data. $\boldsymbol{h}$ is a latent variable and $\boldsymbol{z}$ is a component indicator variable which is a M-dimensional vector $(z_1,z_2,\cdots,z_m,\cdots,z_M)$. z_m becomes 1 when the data of $\boldsymbol{x}$ and $\boldsymbol{y}$ are generated by the m-th component otherwise 0.

A joint probability of those random variables are represented by

$$p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{h},\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{h},\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{h},\boldsymbol{z})p(\boldsymbol{h})p(\boldsymbol{z}). \tag{1}$$

Here,

$$\begin{aligned} p(\boldsymbol{h}) &= \mathcal{N}(0,\boldsymbol{I}), & (2)\\ p(\boldsymbol{x}|\boldsymbol{h},z_m) &= \mathcal{N}(\boldsymbol{U}_m\boldsymbol{h}+\boldsymbol{b}_m,\boldsymbol{\Gamma}_m), & (3)\\ p(\boldsymbol{y}|\boldsymbol{h},z_m) &= \mathcal{N}(\boldsymbol{V}_m\boldsymbol{h}+\boldsymbol{d}_m,\boldsymbol{\Lambda}_m), & (4)\\ p(z_m) &= \phi_m. & (5)\end{aligned}$$

http://dx.doi.org/10.21437/Interspeech.2017-1010

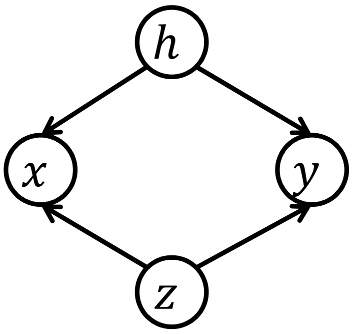

Figure 1: *Graphical model of mPCCA*

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The model assumes linear transforms between the observed data and the latent variable $\boldsymbol{h}$ for each component as follows;

$$\boldsymbol{x} = \boldsymbol{U}_m\boldsymbol{h} + \boldsymbol{b}_m + \boldsymbol{\gamma}_m, \tag{6}$$

$$\boldsymbol{y} = \boldsymbol{V}_m\boldsymbol{h} + \boldsymbol{d}_m + \boldsymbol{\lambda}_m. \tag{7}$$

$\boldsymbol{\gamma}_m$ and $\boldsymbol{\lambda}_m$ are Gaussian noises following $\mathcal{N}(0, \boldsymbol{\Gamma}_m)$ and $\mathcal{N}(0, \boldsymbol{\Lambda}_m)$, respectively. The parameter set of the model is $\Theta = \{\boldsymbol{U}_m, \boldsymbol{V}_m, \boldsymbol{\Gamma}_m, \boldsymbol{\Lambda}_m, \phi_m\}$.

2.2. Acoustic-to-articulatory mapping based on mPCCA

In a2a mapping, the observed data $\boldsymbol{x}$ and $\boldsymbol{y}$ are acoustic and articulatory features, respectively. The latent variable $\boldsymbol{h}$ means a feature which is inherent in both of acoustic and articulatory feature domains. The feature is a continuous value which can be given an interpretation such as activation of speech. On the other hand, another latent variable $\boldsymbol{z}$ is discrete. If the total number of components is set to an adequate number, $\boldsymbol{z}$ can be expected to capture the phonetic features.

A joint probability of $\boldsymbol{x}$ and $\boldsymbol{y}$, $p(\boldsymbol{x}, \boldsymbol{y})$ is represented as that of a joint vector $[\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ by a GMM below;

$$p(\boldsymbol{x}, \boldsymbol{y}) = \sum_m \phi_m \mathcal{N}(\boldsymbol{\mu}_m^{(x,y)}, \boldsymbol{\Sigma}_m^{(x,y)}), \tag{8}$$

$$\boldsymbol{\mu}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{b}_m \\ \boldsymbol{d}_m \end{bmatrix}, \tag{9}$$

$$\boldsymbol{\Sigma}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{U}_m\boldsymbol{U}_m^\top + \boldsymbol{\Gamma}_m & \boldsymbol{U}_m\boldsymbol{V}_m^\top \\ \boldsymbol{V}_m\boldsymbol{U}_m^\top & \boldsymbol{V}_m\boldsymbol{V}_m^\top + \boldsymbol{\Lambda}_m \end{bmatrix}. \tag{10}$$

A covariance matrix in Eq.(10) is structuralized reasonably by using coefficients of the linear transforms in Eqs.(6) and (7). Let D_h, D_x and D_y be dimensions of $\boldsymbol{h}$, $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, the total number of parameters in the covariance matrix is

$$(D_x \times D_h + D_x^2) + (D_y \times D_h + D_y^2). \tag{11}$$

Eq.(11) is less than that of a conventional GMM, when D_h satisfies

$$2\frac{D_x D_y}{D_x + D_y} > D_h. \tag{12}$$

When $\boldsymbol{\Gamma}_m$ and $\boldsymbol{\Lambda}_m$ are assumed diagonal matrices, Eq.(11) can be revised as follows;

$$(D_x \times D_h + D_x) + (D_y \times D_h + D_y). \tag{13}$$

At this assumption, Eq.(12) can be also revised as follows;

$$D_x + D_y - 1 > D_h. \tag{14}$$

2.3. Model training

The model parameter set Θ is trained to maximize the joint probability represented in Eq.(1). When the acoustic-articulatory data are observed as results of Eq.(1), the joint probability is maximized by using EM algorithm. An auxiliary function of the EM algorithm is written by

$$\mathcal{Q}(\Theta, \Theta^{old}) = \int \sum_{m,t} p(\boldsymbol{h}_t, z_{m_t} | \boldsymbol{x}, \boldsymbol{y}; \Theta^{old}) \times \log p(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{h}_t, z_{m_t}; \Theta) d\boldsymbol{h} \tag{15}$$

$$= \sum_{t,m} \Bigg\{ \langle z_{m_t} | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \ln \phi_m \tag{16}$$

$$+ \langle z_{m_t} | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \Big[-\frac{1}{2} \ln |\boldsymbol{\Gamma}_m| - \frac{1}{2} (\boldsymbol{x}_t - \boldsymbol{b}_m)^\top \boldsymbol{\Gamma}_m^{-1} (\boldsymbol{x}_t - \boldsymbol{b}_m) \Big]$$
$$+ (\boldsymbol{x}_t - \boldsymbol{b}_m)^\top \boldsymbol{\Gamma}_m^{-1} \boldsymbol{U}_m \langle z_{m_t} \boldsymbol{h}_t | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle$$
$$- \frac{1}{2} \Big[\mathrm{Tr} \Big\{ \Big(\boldsymbol{U}_m^\top \boldsymbol{\Gamma}_m^{-1} \boldsymbol{U}_m + \boldsymbol{I} \Big) \langle z_{m_t} \boldsymbol{h}_t \boldsymbol{h}_t^\top | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \Big\} \Big]$$
$$+ \langle z_{m_t} | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \Big[-\frac{1}{2} \ln |\boldsymbol{\Lambda}_m| - \frac{1}{2} (\boldsymbol{y}_t - \boldsymbol{d}_m)^\top \boldsymbol{\Lambda}_m^{-1} (\boldsymbol{y}_t - \boldsymbol{d}_m) \Big]$$
$$+ (\boldsymbol{y}_t - \boldsymbol{d}_m)^\top \boldsymbol{\Lambda}_m^{-1} \boldsymbol{V}_m \langle z_{m_t} \boldsymbol{h}_t | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle$$
$$- \frac{1}{2} \Big[\mathrm{Tr} \Big\{ \Big(\boldsymbol{V}_n^\top \boldsymbol{\Lambda}_m^{-1} \boldsymbol{V}_m + \boldsymbol{I} \Big) \langle z_{m_t} \boldsymbol{h}_t \boldsymbol{h}_t^\top | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \Big\} \Big] \Bigg\}. \tag{17}$$

Here $\langle \cdot \rangle$ means an expectation and t is a time index of observation data.

E-step

In the E-step, expectations of hidden variables $\boldsymbol{h}$ and $\boldsymbol{z}$ are obtained as follows;

$$\langle z_{m_t} | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle = \frac{\phi_m p(\boldsymbol{x}_t, \boldsymbol{y}_t | z_m)}{\sum_j^M \phi_j p(\boldsymbol{x}_t, \boldsymbol{y}_t | z_j)} \tag{18}$$

$$\langle z_{m_t} \boldsymbol{h}_t | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \approx \langle z_{m_t} | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \langle \boldsymbol{h}_t | \boldsymbol{x}_t, \boldsymbol{y}_t, z_m \rangle \tag{19}$$

$$\langle z_{m_t} \boldsymbol{h}_t \boldsymbol{h}_t^\top | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \approx \langle z_{m_t} | \boldsymbol{x}_t, \boldsymbol{y}_t \rangle \langle \boldsymbol{h}_t \boldsymbol{h}_t^\top | \boldsymbol{x}_t, \boldsymbol{y}_t, z_m \rangle \tag{20}$$

$$\langle \boldsymbol{h}_t | \boldsymbol{x}_t, \boldsymbol{y}_t, z_{m_t} \rangle = \boldsymbol{L}_m^{-1} \Big(\boldsymbol{U}_m^\top \boldsymbol{\Gamma}_m^{-1} (\boldsymbol{x}_t - \boldsymbol{b}_m) + \boldsymbol{V}_m^\top \boldsymbol{\Lambda}_m^{-1} (\boldsymbol{y}_t - \boldsymbol{d}_m) \Big) \tag{21}$$

$$\langle \boldsymbol{h}_t \boldsymbol{h}_t^\top | \boldsymbol{x}_t, \boldsymbol{y}_t, z_{m_t} \rangle = \boldsymbol{L}_m^{-1} + \langle \boldsymbol{h}_t | \boldsymbol{x}_t, \boldsymbol{y}_t, z_m \rangle \times \langle \boldsymbol{h}_t | \boldsymbol{x}_t, \boldsymbol{y}_t, z_m \rangle^\top \tag{22}$$

$$\boldsymbol{L}_m = \boldsymbol{I} + \boldsymbol{U}_m^\top \boldsymbol{\Gamma}_m^{-1} \boldsymbol{U}_m + \boldsymbol{V}_m^\top \boldsymbol{\Lambda}_m^{-1} \boldsymbol{V}_m \tag{23}$$

To derive Eq.(21) and (23), following equations are used;

$$p(\boldsymbol{h}_t|\boldsymbol{x}_t, \boldsymbol{y}_t, z_{m_t}) \propto p(\boldsymbol{x}_t|\boldsymbol{h}_t, z_{m_t})p(\boldsymbol{y}_t|\boldsymbol{h}_t, z_{m_t})p(\boldsymbol{h}_t) \tag{24}$$

$$\propto \exp\Bigg[-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{U}_m\boldsymbol{h}_t - \boldsymbol{b}_m)^\top \boldsymbol{\Gamma}_m^{-1}(\boldsymbol{x}_t - \boldsymbol{U}_m\boldsymbol{h}_t - \boldsymbol{b}_m)$$
$$-\frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{V}_m\boldsymbol{h}_t - \boldsymbol{d}_m)^\top \boldsymbol{\Lambda}_m^{-1}(\boldsymbol{y}_t - \boldsymbol{V}_m\boldsymbol{h}_t - \boldsymbol{d}_m) - \frac{1}{2}\boldsymbol{h}_t^\top \boldsymbol{h}_t\Bigg] \tag{25}$$

$$\propto \exp\Bigg[\boldsymbol{h}_t^\top \left(\boldsymbol{U}_m^\top \boldsymbol{\Gamma}_m^{-1}(\boldsymbol{x}_t - \boldsymbol{b}_m) + \boldsymbol{V}_m^\top \boldsymbol{\Lambda}_m^{-1}(\boldsymbol{y}_t - \boldsymbol{d}_m)\right)\boldsymbol{h}_t$$
$$-\frac{1}{2}\boldsymbol{h}_t^\top \left(\boldsymbol{I} + \boldsymbol{U}_m^\top \boldsymbol{\Gamma}_m^{-1}\boldsymbol{U}_m + \boldsymbol{V}_m^\top \boldsymbol{\Lambda}_m^{-1}\boldsymbol{V}_m\right)\boldsymbol{h}_m\Bigg]. \tag{26}$$

M-step

In the M-step, the parameter set Θ is renewed as follows;

$$\phi_m = \frac{\sum_t^T \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle}{\sum_t^T \sum_m^M \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle} \tag{27}$$

$$\boldsymbol{U}_m = \left[\sum_{t=1}^T (\boldsymbol{x}_t - \boldsymbol{b}_m)\langle z_{m_t}\boldsymbol{h}_t|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle^\top\right] \left[\sum_{t=1}^T \langle z_{m_t}\boldsymbol{h}_t\boldsymbol{h}_t^\top|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle\right]^{-1} \tag{28}$$

$$\boldsymbol{b}_m = \frac{\sum_t^T \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle\boldsymbol{x}_t}{\sum_t^T \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle} \tag{29}$$

$$\boldsymbol{\Gamma}_m = \frac{1}{\sum_t^T \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle} \times \sum_{t=1}^T \Bigg[(\boldsymbol{x}_t - \boldsymbol{b}_m)(\boldsymbol{x}_t - \boldsymbol{b}_m)^\top \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle - \boldsymbol{U}_m\Big(\langle z_{m_t}\boldsymbol{h}_t|\boldsymbol{x}_t, \boldsymbol{a}_t\rangle\Big)(\boldsymbol{x}_t - \boldsymbol{b}_m)^\top\Bigg] \tag{30}$$

$$\boldsymbol{V}_m = \left[\sum_{t=1}^T (\boldsymbol{y}_t - \boldsymbol{d}_m)\langle z_{m_t}\boldsymbol{h}_t|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle^\top\right] \left[\sum_{t=1}^T \langle z_{m_t}\boldsymbol{h}_t\boldsymbol{h}_t^\top|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle\right]^{-1} \tag{31}$$

$$\boldsymbol{d}_m = \frac{\sum_t^T \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle\boldsymbol{y}_t}{\sum_t^T \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle} \tag{32}$$

$$\boldsymbol{\Lambda}_m = \frac{1}{\sum_t^T \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle} \times \sum_{t=1}^T \Bigg[(\boldsymbol{y}_t - \boldsymbol{d}_m)(\boldsymbol{y}_t - \boldsymbol{d}_m)^\top \langle z_{m_t}|\boldsymbol{x}_t, \boldsymbol{y}_t\rangle - \boldsymbol{V}_m\Big(\langle z_{m_t}\boldsymbol{h}_t|\boldsymbol{x}_t, \boldsymbol{a}_t\rangle\Big)(\boldsymbol{y}_t - \boldsymbol{d}_m)^\top\Bigg] \tag{33}$$

2.4. Mapping process

In the rest of the paper, the time index t is omitted for readability. In the proposed model, when an acoustic feature $\boldsymbol{x}$ is given, an estimated articulatory feature $\hat{\boldsymbol{y}}$ is represented as

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \log p(\boldsymbol{y}|\boldsymbol{x}), \tag{34}$$

$$= \arg\max_{\boldsymbol{y}} \log \int \sum_m p(\boldsymbol{y}, \boldsymbol{h}, z_m|\boldsymbol{x})d\boldsymbol{h}. \tag{35}$$

Since $\boldsymbol{h}$ and $\boldsymbol{z}$ can not be observed, EM algorithm is employed for Eq.(35). An auxiliary function of the EM algorithm is written by

$$\mathcal{Q}(\hat{\boldsymbol{y}}, \hat{\boldsymbol{y}}^{old}) = \int \sum_m p(\boldsymbol{h}, z_m|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}) \log p(\hat{\boldsymbol{y}}, \boldsymbol{h}, z_m|\boldsymbol{x})d\boldsymbol{h}. \tag{36}$$

In the **E-step**, expectation $\langle z_m|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}\rangle$ and $\langle \boldsymbol{h}|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}, z_m\rangle$ are calculated using Eq.(18) and Eq.(21). In the **M-step**, $\hat{\boldsymbol{y}}$ is renewed as follows;

$$\hat{\boldsymbol{y}} = \left(\sum_m \langle z_m|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}\rangle\boldsymbol{\Lambda}_m^{-1}\right)^{-1} \times \sum_m \langle z_m|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}\rangle\boldsymbol{\Lambda}_m^{-1}\Big(\boldsymbol{V}_m\langle \boldsymbol{h}|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}, z_m\rangle + \boldsymbol{d}_m\Big) \tag{37}$$

This mapping process requires an initial value of the articulatory feature $\boldsymbol{y}$. We propose another mapping process which does not require the initial value. For the mapping process, two approximation are introduced in the **E-step** as follows;

$$\langle z_m|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}\rangle \approx \langle z_m|\boldsymbol{x}\rangle \tag{38}$$

$$= \frac{\phi_m p(\boldsymbol{x}_t|z_m)}{\sum_j^M \phi_j p(\boldsymbol{x}_t|z_j)}, \tag{39}$$

$$\langle \boldsymbol{h}|\boldsymbol{x}, \hat{\boldsymbol{y}}^{old}, z_m\rangle \approx \langle \boldsymbol{h}|\boldsymbol{x}, z_m\rangle \tag{40}$$

$$= \boldsymbol{L}_m^{-1}\Big(\boldsymbol{U}_m^\top \boldsymbol{\Gamma}_m^{-1}(\boldsymbol{x}_t - \boldsymbol{b}_m)\Big), \tag{41}$$

$$\boldsymbol{L}_m = \boldsymbol{I} + \boldsymbol{U}_m^\top \boldsymbol{\Gamma}_m^{-1}\boldsymbol{U}_m. \tag{42}$$

In the **M-step**, the estimated articulatory feature is obtained as follows;

$$\hat{\boldsymbol{y}} = \left(\sum_m \langle z_m|\boldsymbol{x}\rangle\boldsymbol{\Lambda}_m^{-1}\right)^{-1} \times \sum_m \langle z_m|\boldsymbol{x}\rangle\boldsymbol{\Lambda}_m^{-1}(\boldsymbol{V}_m\langle \boldsymbol{h}|\boldsymbol{x}, z_m\rangle + \boldsymbol{d}_m). \tag{43}$$

This mapping process is deterministic unlike the original process in Eq.(37).

3. Experimental evaluations

3.1. Conditions

To evaluate the performance of our proposed mapping model, we conducted experimental evaluations using MOCHA database [7]. This database includes acoustic-articulatory parallel data of one male speaker and one female speaker. Articulatory data are measured by an electromagnetic articulography, where its sensors are placed at lower incisor (LI), upper and lower lips (UL and LL), tongue tip (TT), tongue body (TB), tongue dorsum (TD) and velum (V) in the med-sagittal plane. Locations of measuring points are schematically shown as Figure 2. The articulatory data of each point are two-dimensional data of horizontal and vertical directions. The parallel data were divided into 5 parts for 5-fold cross validation. Here, 4 parts (368 utterances) were used for training and the other (92 utterances) was used for testing. Those parts are prepared for each speaker.

For an acoustic feature, 24-dimensional mel-cepstrums were extracted from acoustic data. The first 75 principal components obtained from PCA of 11 frames of the mel-cepstrums

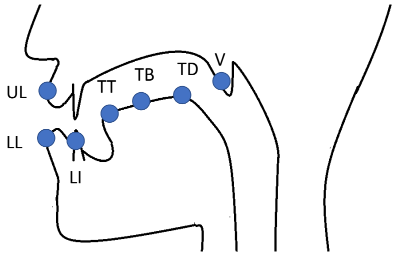

Figure 2: *Measuring points of sensors*

Table 1: *Estimation errors* (mm) *when the total number of components is 64.*

	female	male
GMM-mmse	1.58	1.51
mPPCA-w/o:Init	1.62	1.52
mPPCA-w/:Init	**1.56**	**1.48**

centering at the target frame were used as the acoustic feature. The 14-dimensional articulatory data (7 points × 2 dimensions) were used as the articulatory feature. The frame shift of those features were set to 10 ms.

We conducted a2a mapping using the two models; a conventional GMM-based mapping model and the proposed mapping model. In the mapping process of the GMM-based model, minimum mean square error criteria were adopted (**GMM-mmse**). On the other hand, in the proposed model, the two mapping processes were employed; **mPPCA-w/:Init** in Eq.(37) and **mPPCA-w/o:Init** in Eq.(43). We introduced an approximation to E-step of mPPCA-w/:Init as follows;

$$\langle z_m | \boldsymbol{x}, \hat{\boldsymbol{y}}^{old} \rangle \approx \begin{cases} 1, & \text{if } m = \arg\max_m \langle z_m | \boldsymbol{x}, \hat{\boldsymbol{y}}^{old} \rangle \\ 0, & \text{otherwise} \end{cases} . \quad (44)$$

In the experiments, the dimension of $\boldsymbol{h}$ were set to 14, which is the same as that of articulatory data. $\boldsymbol{\Gamma}_m$ and $\boldsymbol{\Lambda}_m$ were assumed to be diagonal metrices.

3.2. Results

A metric of evaluation is root mean square error (RMSE) between the estimated and measured articulatory data. Figure 3 shows estimation errors as a function of the total number of components in GMM and mPCCA. The estimation errors are averaged values of the RMSE over all 14 dimensions of articulatory data. The estimated results of GMM-mmse were used for the initial articulatory features of mPPCA-w/:Init, where the total number of components in the GMM is the same as that in the mPPCA. We can find that the estimation errors of mPPCA-w/:Init are lower than those of GMM-mmse when the total number of components increases (see Table 1). This trend is common to both the speakers. These results indicate that the proposed mapping model can avoid the over-fitting. However, mPPCA-w/o:Init is inferior to GMM-mmse and mPPCA-w/:Init in all the cases. These results mean that the initial articulatory features in advance are required to obtain the better mapping results.

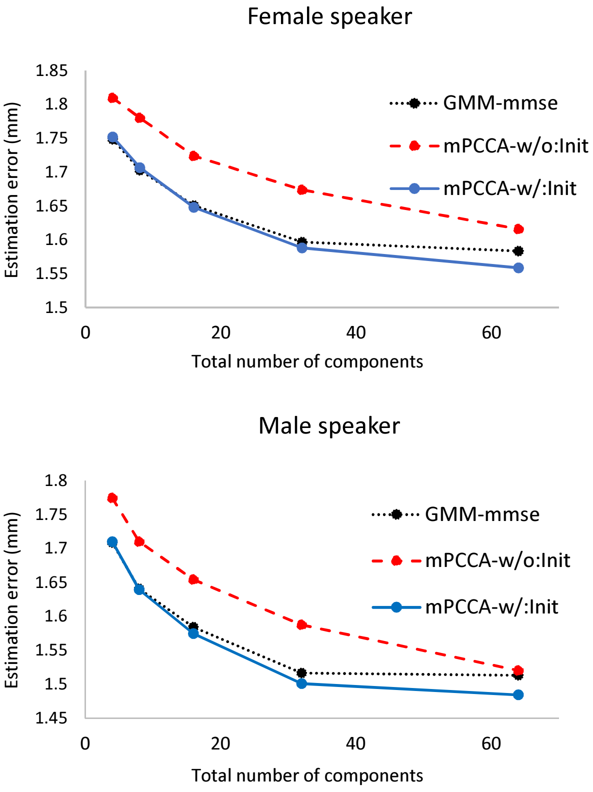

Figure 3: *Estimation errors for each speaker*

4. Conclusions

In this paper, we have proposed a novel a2a mapping model based on mPCCA. The proposed method has two kinds of latent variables, one is the component indicator variable which is the same as that of GMM. The other is the continuous latent variable which is unique to mPCCA. Articulatory and acoustic features are assumed to be generated as results from linear transforms of the continuous latent variable. The covariance matrices of mPCCA are structuralized reasonably, and the structures are expected to avoid over-fitting. The experiments showed that the proposed model has achieved better performance compared with the GMM-based model when the total number of components increases. This result indicates that the structuralized covariance matrices relieve over-fitting as expected.

5. References

[1] S. Hiroya, and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech and Audio Processing*, **12**, 175–185 (2004).

[2] T. Toda, W. A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model", *Speech Commun.*, **50**, 215–227 (2007).

[3] B. Uria, S. Renals, and K. Richmond, "A deep neural network for acoustic-articulatory speech inversion," In *Proc. NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning* (2011).

[4] D. Saito, H. Doi, N. Minematsu, K. Hirose, "Application of Matrix Variate Gaussian Mixture Model to Statistical Voice Conversion," In *Proc. Interspeech2014*, (2014).

[5] B. Zhang, J. Hao, G. Ma, J. Yue, J. Zhang and Z. Shi, "Mixture of Probabilistic Canonical Correlation Analysis," *Journal of Computer Research and Development*, **52(7)**, 1463–1476 (2014). (In Chinese)

[6] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Technical Report 688, Department of Statistics, University of California, Berkeley, (2005)

[7] A. Wrenh, "The MOCHA-TIMIT articulatory database," http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (accessed 2016-11-13).