1-2-8

An Investigation of Acoustic-to-Articulatory Inversion Mapping with Latent Trajectory Gaussian Mixture Model *

◎ Patrick Lumban Tobing (NAIST), Tomoki Toda (Nagoya University/NAIST),
Hirokazu Kameoka (NTT), Satoshi Nakamura (NAIST)

1 Introduction

An acoustic-to-articulatory inversion mapping using a Gaussian mixture model [1] is effective for developing a new speech modification framework [2]. Smoothly varying articulatory parameter trajectory is well estimated from a given acoustic parameters considering interframe correlation by imposing an explicit relationship between static and dynamic features on a maping process. However, this constraint is not taken into accout in a training process of the GMM. The trajectory training method [3] was proposed to address this issue, but it makes the training process too complicated to analytically optimize model parameters.

In this paper, as an alternative method to address this issue, we propose a latent trajectory GMM (LTGMM)-based inversion mapping method inspired by the latent trajectory HMM [4], which makes it possible to use EM algorithm to optimize the model parameters. We conduct an experimental evaluation using a single speaker's articulatory-acoustic data, demonstrating that higher mapping accuracy is achieved using the LTGMM than the traditional GMM.

2 Conventional GMM for acoustic-to-articulatory inversion mapping

Let $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_T^\top]^\top$ be a time sequence of D_x-dimensional static acoustic feature vectors and $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ be that of D_y-dimensional static articulatory feature vectors. At frame t, $2D_x/2D_y$-dimensional acoustic/articulatory feature vectors are denoted as $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$, consisting of D_x/D_y-dimensional joint static and dynamic features. Their joint vector is denoted as $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$. Moreover, their time sequences are written respectively 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 $\boldsymbol{Z} = [\boldsymbol{Z}_1^\top, \cdots, \boldsymbol{Z}_T^\top]^\top$.

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

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

where $\boldsymbol{\lambda}^{(Z)}$ is a set of model parameters consisting of a mixture weight α_m, a mean vector $\boldsymbol{\mu}_m^{(Z)}$ and a covariance matrix $\boldsymbol{\Sigma}_m^{(Z)}$ for the mth mixture component with M total number of mixture components. These parameters are optimized for training data with EM algorithm [1].

In the mapping process, given an acoustic feature sequence $\boldsymbol{X}$, the estimated articulatory feature sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}) \ \text{s.t.}\ \boldsymbol{Y} = \boldsymbol{W}_y \boldsymbol{y}, \qquad (2)$$

where $\boldsymbol{W}_y$ is a linear transform to append dynamic features to a static feature sequence. This ML estimate can be determined with EM algorithm [1].

In this paper, an approximation of the above conditional p.d.f. is employed 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. Then, the estimated articulatory feature sequence $\hat{\boldsymbol{y}}$ is generated from the approximated conditional p.d.f., where its ML estimate can be analytically determined with EM algorithm [5].

Note that the interframe correlation is explicitly considered with the constraint ($\boldsymbol{Y} = \boldsymbol{W}_y \boldsymbol{y}$) in the mapping process. In contrast, it is ignored while optimizing GMM parameters in the training process.

3 Proposed latent trajectory GMM for the inversion mapping

Let the observed variable be a time sequence of joint static feature vectors $\boldsymbol{z} = [\boldsymbol{z}_1^\top, \cdots, \boldsymbol{z}_T^\top]^\top$, where $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$. The following soft constraint is used in the LTGMM:

$$\boldsymbol{Z} \simeq \boldsymbol{W}_z \boldsymbol{z} = [\boldsymbol{W}_x, \boldsymbol{W}_y][\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top. \qquad (3)$$

The joint probability density of the acoustic and articulatory feature vector sequences is modeled with an LTGMM as follows:

$$P(\boldsymbol{z}|\boldsymbol{\lambda}^{(z)}) = \int P(\boldsymbol{z}|\boldsymbol{Z}, \boldsymbol{\Sigma}) P(\boldsymbol{Z}|\boldsymbol{\lambda}^{(Z)}) \mathrm{d}\boldsymbol{Z}, \qquad (4)$$

where

$$\begin{aligned} P(\boldsymbol{z}|\boldsymbol{Z}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{z}; & (\boldsymbol{W}_z^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}_z)^{-1} \boldsymbol{W}_z^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}, \\ & (\boldsymbol{W}_z^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}_z)^{-1}). \end{aligned} \qquad (5)$$

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 [4]. In this paper, an approximation of the above joint p.d.f. is employed using the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$. The approximated joint p.d.f. is given by:

$$P(\boldsymbol{z}|\boldsymbol{\lambda}^{(z)}) = \int P(\boldsymbol{z}|\boldsymbol{Z}, \boldsymbol{\Sigma}) P(\boldsymbol{Z}|\hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(Z)}) \mathrm{d}\boldsymbol{Z}. \qquad (6)$$

In this case, the model parameters can be optimized with EM algorithm.

In the inversion mapping process, given an acous-

*潜在トラジェクトリ混合正規分布モデルによる調音運動逆推定の検討 by Patrick Lumban Tobing (NAIST), 戸田 智基 (名古屋大学/奈良先端大), 亀岡弘和 (NTT), 中村 哲（奈良先端大）

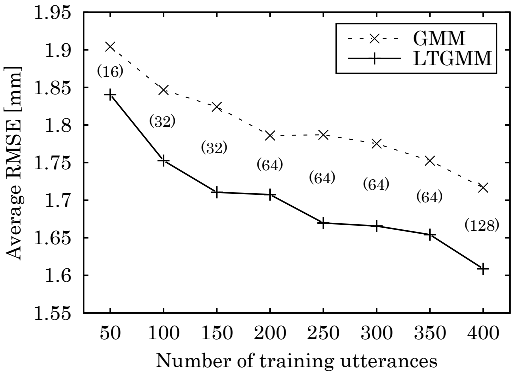

Fig. 1 Average root-mean-square error of testing data and optimum numbers of mixture components.

tic feature sequence $\boldsymbol{x}$, the estimated articulatory feature sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\lambda}^{(z)}), \tag{7}$$

where

$$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\lambda}^{(z)}) = \int P(\boldsymbol{y}|\boldsymbol{Y},\boldsymbol{\Sigma}) \int P(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{\lambda}^{(z)})\mathrm{d}\boldsymbol{X}\mathrm{d}\boldsymbol{Y}. \tag{8}$$

This ML estimate can be determined with variational EM algorithm. In this paper, an approximation of the above conditional p.d.f. is employed by first determining the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$. Then, the estimated articulatory parameter sequence $\hat{\boldsymbol{y}}$ is determined as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{m}},\boldsymbol{\lambda}^{(z)}), \tag{9}$$

where

$$P(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{m}},\boldsymbol{\lambda}^{(z)}) = \int P(\boldsymbol{y}|\boldsymbol{Y},\boldsymbol{\Sigma}) \int P(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{m}},\boldsymbol{\lambda}^{(Z)})P(\boldsymbol{X}|\boldsymbol{x},\hat{\boldsymbol{m}},\boldsymbol{\lambda}^{(z)})\mathrm{d}\boldsymbol{X}\mathrm{d}\boldsymbol{Y}. \tag{10}$$

The ML estimate can be determined analytically.

Note that the interframe correlation is considered, both in the training and conversion process, by imposing explicitly the constraint in Eq. (3).

4 Experimental evaluation

4.1 Experimental conditions

A set of speech and articulatory data of a single British male speaker in MOCHA [6] was used. As the acoustic parameters, we used the 1st-to-24th mel-cepstral coefficients extracted with STRAIGHT analysis [7] from 16 kHz sampled speech data. As the articulatory parameters, we used 14-dimensional EMA data converted to z-score, which represented the movements of 7 articulators, as used in [1]. Frame shift was set to 5 ms.

The constant positive-definite matrix $\boldsymbol{\Sigma}$ in Eq. (5) was set to the diagonal matrix of global variances. In the training process, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$ was initialized beforehand and held fixed. The trained conventional GMM was used as an initial model for the LTGMM training.

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, 200, 250, 300, 350, and 400. The number of mixture components was optimized for each number of training utterances, as given in both Fig. 1 and Fig. 2, using the conventional GMM. The number of testing utterances was 20.

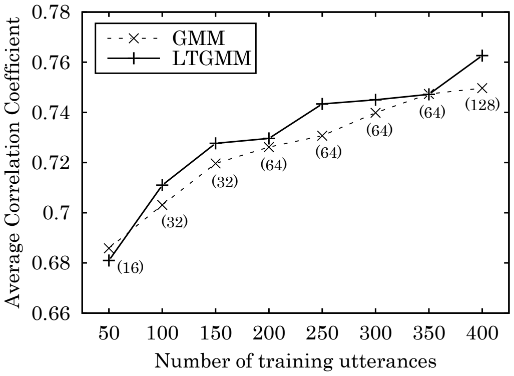

Fig. 2 Average correlation coefficient of testing data and optimum numbers of mixture components.

4.2 Experimental results

Figures 1 and 2 show the RMSE and the correlation coefficient averaged over all 14 dimensions articulatory parameters through all 20 testing utterances. Higher accuracy of inversion mapping is achieved with the LTGMM. This is because the LTGMM can be optimized while considering the interframe correlation, which is considered in only the mapping process if using the conventional GMM.

5 Conclusion

We have proposed an inversion mapping method based on the latent trajectory GMM (LTGMM). The experimental results have demonstrated that higher accuracy of inversion mapping with LTGMM is achieved than the traditional GMM. We will further investigate its performance with the use of an acoustic segment feature consisting of multiple frames of input features and by taking into account all possible mixture component sequences.

Acknowledgements Part of this work was supported by JSPS KAKENHI Grant Number 26280060.

References

[1] T. Toda, *et al.*, Speech Communication, Vol. 50, No. 3, pp. 215–227, 2008.

[2] P. L. Tobing, *et al.*, Proc. INTERSPEECH, pp. 3350–3354, 2015.

[3] H. Zen, *et al.*, Computer Speech and Language, Vol. 21, pp. 153–173, 2007.

[4] H. Kameoka, IEEE 25th International Workshop on MLSP, pp. 1–6, 2015.

[5] T. Toda, *et al.*, IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235, 2007.

[6] A. Wrench, http://www.cstr.ed.ac.uk/artic/mocha.html, 1999.

[7] H. Kawahara, *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.