Acoustic-to-Articulatory Inversion Mapping with Variational Latent Trajectory Gaussian Mixture Model *

◎ Patrick Lumban Tobing (Nagoya Univ.), Hirokazu Kameoka (NTT), Tomoki Toda (Nagoya Univ.)

1 Introduction

Gaussian mixture model (GMM)-based statistical feature mapping method, using trajectory-based conversion process [1], has been reported to perform effectively in an acoustic-to-articulatory inversion mapping task [2]. However, in the training, the model parameters are optimized with different likelihood criterion compared to in the conversion.

To address this inconsistency issue, the trajectory training method has been proposed [3]. Nonetheless, it still has several limitations in terms of parameter optimization due to the difficulties in finding analytical solution. Based on the latent trajectory model [4], in our previous work [5], we have proposed a latent trajectory Gaussian mixture model (LT-GMM) that is capable of addresing the inconsistency issue while allowing convenient parameter optimization with EM algorithm. However, we still use an approximation of a fixed suboptimum mixture component sequence to train the model.

In this report, we propose a variational LT-GMM to allow the inclusion of all possible mixture component sequences in training the model. To investigate the effectiveness of the approximation in our previous work, we compare the accuracy on the inversion mapping task between the proposed variational LT-GMM and the LT-GMMs with a suboptimum mixture component sequence. The experimental result demonstrates that such approximation resembles the performance of the variational method.

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:

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

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.

In this report, in the conversion process, given an acoustic feature sequence $\boldsymbol{X}$, the estimated articulatory feature sequence $\hat{\boldsymbol{y}}$ is determined by

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

where $\boldsymbol{W}_y$ is a linear transform to append dynamic features to a static feature sequence and $\hat{\boldsymbol{m}} = \{\hat{m}_1, \ldots, \hat{m}_T\}$ is a suboptimum mixture component sequence determined as follows:

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{\operatorname{argmax}} \prod_{t=1}^{T} P(m_t|\boldsymbol{X}_t,\boldsymbol{\lambda}^{(Z)}). \quad (3)$$

3 Variational LT-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 LT-GMM:

$$\boldsymbol{Z} \simeq \boldsymbol{W}_z\boldsymbol{z} = [\boldsymbol{W}_x, \boldsymbol{W}_y][\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top. \quad (4)$$

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

$$P(\boldsymbol{z}|\boldsymbol{\lambda}) = \int \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{z}|\boldsymbol{Z},\boldsymbol{\Sigma})P(\boldsymbol{Z},\boldsymbol{m}|\boldsymbol{\lambda}^{(Z)})\mathrm{d}\boldsymbol{Z}, \quad (5)$$

where $\boldsymbol{m} = \{m_1, \ldots, m_T\}$ is a mixture component sequence, and

$$P(\boldsymbol{z}|\boldsymbol{Z},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{H}\boldsymbol{Z},\boldsymbol{\Lambda}^{-1}) \quad (6)$$

$$\boldsymbol{H} = \boldsymbol{\Lambda}^{-1}\boldsymbol{W}_z^\top\boldsymbol{\Sigma}^{-1} \quad (7)$$

$$\boldsymbol{\Lambda} = \boldsymbol{W}_z^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_z. \quad (8)$$

A functional lower bound $\mathcal{F}(q,\boldsymbol{\lambda})$, corresponding to the log-likelihood of Eq. (5), by following the Jensen's inequality, can be written as follows:

$$\mathcal{F}(q,\boldsymbol{\lambda}) = \int \sum_{\text{all } \boldsymbol{m}} q(\boldsymbol{Z},\boldsymbol{m}) \log\frac{P(\boldsymbol{z},\boldsymbol{Z},\boldsymbol{m}|\boldsymbol{\lambda})}{q(\boldsymbol{Z},\boldsymbol{m})}\mathrm{d}\boldsymbol{Z}, \quad (9)$$

where $q(\boldsymbol{Z},\boldsymbol{m})$ is a variational posterior distribution, which is further approximated as follows:

$$P(\boldsymbol{Z},\boldsymbol{m}) \simeq P(\boldsymbol{Z})P(\boldsymbol{m}). \quad (10)$$

---

* 変分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 by Patrick Lumban Tobing (名古屋大学), 亀岡弘和 (NTT), 戸田 智基 (名古屋大学)

In the expectation step (E-step), the variational posterior distribution $P(\boldsymbol{Z})$ is calculated as follows:

$$\mathbb{E}[\boldsymbol{Z}] = \overline{\boldsymbol{\mu}^{(Z)}} + \overline{\boldsymbol{\Sigma}^{(Zz)}\boldsymbol{\Sigma}^{(zz)1}}(\boldsymbol{z} - \boldsymbol{H}\overline{\boldsymbol{\mu}^{(Z)}}) \quad (11)$$

$$\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top] = \overline{\boldsymbol{\Sigma}^{(Z)}} - \overline{\boldsymbol{\Sigma}^{(Zz)}\boldsymbol{\Sigma}^{(zz)^{-1}}\boldsymbol{\Sigma}^{(zZ)}}, \quad (12)$$

and the variational posterior distribution $P(\boldsymbol{m})$ as:

$$\gamma_{m,t} = \frac{\alpha_m \mathcal{N}(\mathbb{E}[\boldsymbol{Z}]_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma}_m^{(Z)}) C_{m,t}}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\mathbb{E}[\boldsymbol{Z}]_t; \boldsymbol{\mu}_n^{(Z)}, \boldsymbol{\Sigma}_n^{(Z)}) C_{n,t}} \quad (13)$$

$$C_{m,t} = \exp\big\{ -\frac{1}{2} \mathrm{Tr}(\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top]_t \boldsymbol{\Sigma}_m^{(Z)^{-1}}) \big\}. \quad (14)$$

Note that, an overline in Eqs. (11) and (12) denote a sequence of mean vectors or covariance matrices, where at each time-frame t, they are marginalized over all mixture components using the occupancy $\gamma_{m,t}$. Hence, to optimize these distributions, i.e., $P(\boldsymbol{Z})$ and $P(\boldsymbol{m})$, the values in one must be kept while refining the values in the other one.

Then, in the maximization step (M-step), the model parameters $\boldsymbol{\lambda}$ are optimized by using the functional lower-bound in Eq. (9) and the approximated variational posterior distribution $q(\boldsymbol{Z}, \boldsymbol{m})$.

In the conversion, given an acoustic feature sequence $\boldsymbol{x}$, the estimated articulatory parameter sequence $\hat{\boldsymbol{y}}$ is determined as in [5] by

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}), \quad (15)$$

where the suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is given by Eq. (3).

4 Experimental evaluation

4.1 Experimental conditions

A set of speech and articulatory data of a British male speaker in MOCHA [6] was used. As the acoustic parameters, we used the first through 24$^{\text{th}}$ mel-cepstral coefficients extracted using STRAIGHT [7]. As the articulatory parameters, we used the 14-dimensional EMA data.

The error-covariance matrix $\boldsymbol{\Sigma}$ was fixed to the initial values of weighted covariance over all mixture components. We trained three models of LT-GMM: with a fixed suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ (LT-GMM_fixseq), as in [5]; by updating the suboptimum mixture component sequence in the M-step of each iteration (LT-GMM_updseq); and with all possible mixture component sequences, i.e., using the variational method, (LT-GMM_allseqs). The trained conventional GMM was used as the initial model for training the LT-GMMs.

4.2 Experimental results

The result of the experiment is shown in Table 1. It can be observed that the three LT-GMMs gives higher accuracy for the inversion mapping compared to the conventional GMM by yielding lower values of RMSE and higher values of correlation coefficient. Furthermore, the LT-GMM with variational method, i.e., to consider all possible mixture component sequences (LT-GMM_allseqs), gives slightly better accuracy than its two counterparts, i.e., by using a fixed suboptimum mixture component sequence (LT-GMM_fixseq) and by updating the suboptimum sequence (LT-GMM_updseq). Specifically, the observed difference of values between the LT-GMM_allseqs and the other two are 0.006 for the average RMSE and 0.002 for the average correlation coefficient. This result suggests that, even-though the variational technique is capable of giving the best accuracy through exhaustive consideration of all possible mixture component sequences, the approximation of a suboptimum mixture component sequence can still rival that performance with a marginal difference in the accuracy.

Table 1 Average of root-mean-square error (RMSE) and correlation coefficient for inversion mapping using GMM and three models of LT-GMM

	Avg. RMSE	Avg. Corr.
GMM	1.609	0.756
LT-GMM_fixseq	1.554	0.756
LT-GMM_updseq	1.554	0.756
LT-GMM_allseqs	1.548	0.758

5 Conclusion

In this report, we have presented a variational method for the latent trajectory Gaussian mixture model (LT-GMM). The variational technique allows us to optimize the coupled latent variables, i.e., the sequence of joint static and dynamic features and the mixture component sequence. The experimental result demonstrated that by considering all possible mixture component sequences, i.e., using the variational method, a marginal improvement was yielded compared to using only a suboptimum sequence. This implies that, through the low-cost approximation of a suboptimum mixture component sequence, similar performance can be achieved to avoid the high-cost variational LT-GMM procedure.

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

References

[1] T. Toda, et al., *IEEE/ACM Trans. Audio Speech Lang. Process*, vol. 15, no. 8, pp. 2222–2235, 2007.
[2] P. L. Tobing, et al., in *Proc. INTERSPEECH*, pp. 3350–3354, 2015.
[3] H. Zen, et al., *Comput. Speech Lang.*, vol. 21, pp. 153–173, 2007.
[4] H. Kameoka, in *Proc. MLSP*, pp. 1–6, 2015.
[5] P. L. Tobing, et al., *Proc. INTERSPEECH*, pp. 953–957, 2016.
[6] A. Wrench, http://www.cstr.ed.ac.uk/artic/mocha.html, 1999.
[7] H. Kawahara, et al., *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.