# A Bayesian Approach to Voice Conversion Based on GMMs Using Multiple Model Structures

*Lei Li, Yoshihiko Nankaku, Keiichi Tokuda*

Department of Computer Science and Engineering
Nagoya Institute of Technology, Nagoya, Japan
{li-lei,nankaku}@sp.nitech.ac.jp, tokuda@nitech.ac.jp

## Abstract

A spectral conversion method using multiple Gaussian Mixture Models (GMMs) based on the Bayesian framework is proposed. A typical spectral conversion framework is based on a GMM. However, in this conventional method, a GMM-appropriate number of mixtures is dependent on the amount of training data, and thus the number of mixtures should be determined beforehand. In the proposed method, the variational Bayesian approach is applied to GMM-based voice conversion, and multiple GMMs are integrated as a single statistical model. Appropriate model structures are stochastically selected for each frame based on the Bayesian frame work.

**Index Terms**: voice conversion, speech synthesis, GMM, model structure

## 1. Introduction

Voice conversion is a technique for converting a certain speaker's voice into another speaker's voice. It can modify speech characteristics using conversion rules statistically extracted from a small amount of data. One typical spectral conversion framework is based on a Gaussian Mixture Model (GMM)[1]. This method achieves continuous mapping based on soft clustering. A more accurate formulation of spectral conversion based on a Maximum Likelihood (ML) criterion has been presented [2]. In the ML-based conversion, both training and conversion processes are consistently derived based on the single ML objective function. In this method, spectral conversion is performed using a single GMM structure with a static number of mixtures, assuming that the optimal structure is common among the utterances. However, the assumption is not sufficient, since voice features may vary even within a sentence and thus the optimal model structure may change at every frame.

This paper describes a method to solve this problem: the Variational Bayesian (VB) approach is applied to GMM-based voice conversion, and multiple GMMs are integrated as a single statistical model[4, 5]. The VB approach, which uses the variational approximation technique[6, 7], has been proposed and recently been applied to many classifications using latent variable models. Appropriate model structures are stochastically selected for each frame based on the Bayesian framework. One of the methods proposed for voice conversion, the Dynamic Models Selection method (DMS)[3], is similar to this proposed method with ML criterion.

This paper is organized as follows. Section 2 explains the conventional voice conversion technique based on GMMs, and section 3 explains voice conversion based on the Variational Bayesian Method. The proposed method of spectral conversion using multiple structures is described in Section 4, and experimental results are reported in Section 5. Finally, a summary of this work is given in Section 6.

## 2. Spectral Conversion Based on GMM

To convert spectral feature sequences of a source speaker to that of a target speaker, the joint probability of two features is modeled by a GMM [2]. Let a vector $\boldsymbol{x}_t = \left[\boldsymbol{x}_t^{(1)\top}, \boldsymbol{x}_t^{(2)\top}\right]^\top$ be a joint feature vector of the source $\boldsymbol{x}_t^{(1)}$ and the target $\boldsymbol{x}_t^{(2)}$ at time $t$, where $\cdot^\top$ denotes transposition of the vector. Alignment between two feature sequences is obtained by the Dynamic Programming (DP) matching. In the GMM-based voice conversion, the vector sequence $\boldsymbol{x} = \left[\boldsymbol{x}_1{}^\top, \boldsymbol{x}_2{}^\top, \cdots, \boldsymbol{x}_T{}^\top\right]^\top$ is modeled by a GMM to learn a relation between source and target features. The output probability of $\boldsymbol{x}$, given GMM $\boldsymbol{\lambda}$, can be written as

$$P(\boldsymbol{x}|\boldsymbol{\lambda}) = \prod_{t=1}^{T}\left[\sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right] \tag{1}$$

where

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{(1)} \\ \boldsymbol{\mu}_i^{(2)} \end{bmatrix}, \boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(1,1)} & \boldsymbol{\Sigma}_i^{(1,2)} \\ \boldsymbol{\Sigma}_i^{(2,1)} & \boldsymbol{\Sigma}_i^{(2,2)} \end{bmatrix}, \tag{2}$$

and $M$ means the number of mixtures, $w_i = P(i|\boldsymbol{\lambda})$ is the mixture weight of the $i$-th component, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix, respectively. These model parameters are estimated using the Expectation Maximization (EM) algorithm.

### 2.1. Maximum Likelihood Spectral Conversion

In the maximum likelihood spectral conversion, the optimal converted feature sequence $\boldsymbol{x}^{(2)} = \left[\boldsymbol{x}_1^{(2)\top}, \boldsymbol{x}_2^{(2)\top}, \cdots, \boldsymbol{x}_T^{(2)\top}\right]^\top$, given a source feature sequence $\boldsymbol{x}^{(1)} = \left[\boldsymbol{x}_1^{(1)\top}, \boldsymbol{x}_2^{(1)\top}, \cdots, \boldsymbol{x}_T^{(1)\top}\right]^\top$, is obtained by maximizing the following conditional distribution:

$$\begin{aligned} &P(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(1)}, \boldsymbol{\lambda}) \\ &= \sum_{\boldsymbol{m}}\left[P(\boldsymbol{m}|\boldsymbol{x}^{(1)}, \boldsymbol{\lambda})\prod_{t=1}^{T} P(\boldsymbol{x}_t^{(2)}|\boldsymbol{x}_t^{(1)}, m_t, \boldsymbol{\lambda})\right] \end{aligned} \tag{3}$$

where $\boldsymbol{m} = [m_1, m_2, \cdots, m_T]$ is a mixture index sequence. The conditional distribution can also be written as a GMM, and its output probability distribution is presented as

$$P(\boldsymbol{x}_t^{(2)}|\boldsymbol{x}_t^{(1)}, m_t = i, \boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{x}_t^{(2)}; \boldsymbol{E}_i(t), \boldsymbol{D}_i\right) \tag{4}$$

where

$$\boldsymbol{E}_i(t) = \boldsymbol{\mu}_i^{(2)} + \boldsymbol{\Sigma}_i^{(2,1)}\boldsymbol{\Sigma}_i^{(1,1)^{-1}}\left(\boldsymbol{x}_t^{(1)} - \boldsymbol{\mu}_i^{(1)}\right) \quad (5)$$

$$\boldsymbol{D}_i = \boldsymbol{\Sigma}_i^{(2,2)} - \boldsymbol{\Sigma}_i^{(2,1)}\boldsymbol{\Sigma}_i^{(1,1)^{-1}}\boldsymbol{\Sigma}_i^{(1,2)}. \quad (6)$$

Since Eq. (3) includes latent variables, the optimal sequence of $\boldsymbol{x}^{(2)}$ is estimated using the EM algorithm. The EM algorithm is an iterative method for approximating the maximum likelihood estimation. It maximizes the expectation of the complete data log-likelihood, i.e., the $Q$-function (auxiliary function):

$$Q(\boldsymbol{x}^{(2)}, \hat{\boldsymbol{O}}^{(2)}) = \sum_{\boldsymbol{m}} \left[P(\boldsymbol{x}^{(2)}, \boldsymbol{m}|\boldsymbol{x}^{(1)}, \boldsymbol{\lambda}) \ln P(\hat{\boldsymbol{O}}^{(2)}, \boldsymbol{m}|\boldsymbol{x}^{(1)}, \boldsymbol{\lambda})\right]. \quad (7)$$

By taking the derivative of the $Q$-function, the spectral sequence $\hat{\boldsymbol{O}}^{(2)}$ that maximizes the $Q$-function is given by

$$\hat{\boldsymbol{O}}^{(2)} = \left(\overline{\boldsymbol{D}^{-1}}\right)^{-1}\overline{\boldsymbol{D}^{-1}\boldsymbol{E}} \quad (8)$$

where

$$\overline{\boldsymbol{D}^{-1}} = \mathrm{diag}\left[\overline{\boldsymbol{D}_1^{-1}}, \overline{\boldsymbol{D}_2^{-1}}, \cdots, \overline{\boldsymbol{D}_T^{-1}}\right] \quad (9)$$

$$\overline{\boldsymbol{D}_t^{-1}} = \sum_{i=1}^{M}\gamma_t(i)\boldsymbol{D}_i^{-1} \quad (10)$$

$$\overline{\boldsymbol{D}^{-1}\boldsymbol{E}} = \left[\overline{\boldsymbol{D}^{-1}\boldsymbol{E}_1}^{\top}, \overline{\boldsymbol{D}^{-1}\boldsymbol{E}_2}^{\top}, \cdots, \overline{\boldsymbol{D}^{-1}\boldsymbol{E}_T}^{\top}\right]^{\top} \quad (11)$$

$$\overline{\boldsymbol{D}^{-1}\boldsymbol{E}_t} = \sum_{i=1}^{M}\gamma_t(i)\boldsymbol{D}_i^{-1}\boldsymbol{E}_t(i) \quad (12)$$

$$\gamma_t(i) = P(m_t = i|\boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(2)}, \boldsymbol{\lambda}). \quad (13)$$

# 3. Voice Conversion Based on Variational Bayesian Method

## 3.1. Voice Conversion Based on Bayesian Criterion

In the ML criterion, the optimal target feature sequence $\overline{\boldsymbol{x}}^{(2)}$ is generated by maximizing the posterior probability $P(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(1)}, \boldsymbol{\lambda})$ using the maximum likelihood model parameters. In the Bayesian criterion, the model parameters are treated as random variables and the predicted distribution of $\overline{\boldsymbol{x}}^{(2)}$ is estimated by marginalizing model parameters. The target feature sequence of Bayesian criterion is calculated by maximizing the predictive distribution as follows

$$\overline{\boldsymbol{x}}^{(2)} = \underset{\boldsymbol{x}^{(2)}}{\mathrm{argmax}}\, P\left(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(1)}, \boldsymbol{X}\right) = \underset{\boldsymbol{x}^{(2)}}{\mathrm{argmax}} \int P\left(\boldsymbol{x}^{(2)}, \boldsymbol{\lambda}|\boldsymbol{x}^{(1)}, \boldsymbol{X}\right) d\boldsymbol{\lambda} \quad (14)$$

where $\boldsymbol{X}$ is a joint feature sequence of training data consisting of source feature sequence $\boldsymbol{X}^{(1)}$ and target feature sequence $\boldsymbol{X}^{(2)}$. If assuming model parameter $\boldsymbol{\lambda}$ is independent of $\boldsymbol{x}^{(1)}$, Eq. (14) can be written as

$$\overline{\boldsymbol{x}}^{(2)} = \underset{\boldsymbol{x}^{(2)}}{\mathrm{argmax}} \int P\left(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(1)}, \boldsymbol{\lambda}\right) P(\boldsymbol{\lambda}|\boldsymbol{X}) d\boldsymbol{\lambda}. \quad (15)$$

Comparing with the ML criterion, the Bayesian criterion uses all possible parameters represented by posterior probabilities $P(\boldsymbol{\lambda}|\boldsymbol{X})$ given training data $\boldsymbol{X}$, even though the ML criterion estimates a constant model parameter set which maximizes the likelihood $P(\boldsymbol{\lambda}|\boldsymbol{X})$ in the training process. This marginalization avoids the over-training and potentially high generalization performance is obtained. However, solving the integral and expectation calculations is difficult because the calculation becomes more complicated. To over-come this problem, the VB method has been proposed as a tractable approximation method for the Bayesian approach.

## 3.2. Approximation of Posterior Probability using Variational Bayesian Method

Eq. (14) can be rewritten as

$$\overline{\boldsymbol{x}}^{(2)} = \underset{\boldsymbol{x}^{(2)}}{\mathrm{argmax}}\, P\left(\boldsymbol{x}^{(2)}|\boldsymbol{x}^{(1)}, \boldsymbol{X}\right) = \underset{\boldsymbol{x}^{(2)}}{\mathrm{argmax}}\, P\left(\boldsymbol{x}, \boldsymbol{X}\right) \quad (16)$$

where $\boldsymbol{x}$ is a joint feature sequences of testing (conversion) data consisting of $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, and $\boldsymbol{X}$ is a joint feature sequences of training data consisting of $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. Note that the independence assumption between model $\boldsymbol{\lambda}$ and $\boldsymbol{x}^{(1)}$ in 3.1 is not used for accurate approximation. Taking account of latent variables including model parameters, marginal likelihood $P(\boldsymbol{x}, \boldsymbol{X})$ can be defined as

$$P\left(\boldsymbol{x}, \boldsymbol{X}\right) = \sum_{\boldsymbol{m}}\sum_{\boldsymbol{M}}\int P\left(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda}\right) d\boldsymbol{\lambda} = \sum_{\boldsymbol{m}}\sum_{\boldsymbol{M}}\int P\left(\boldsymbol{X}, \boldsymbol{M}|\boldsymbol{\lambda}\right) P\left(\boldsymbol{x}, \boldsymbol{m}|\boldsymbol{\lambda}\right) P\left(\boldsymbol{\lambda}\right) d\boldsymbol{\lambda} \quad (17)$$

where $\boldsymbol{m}, \boldsymbol{M}$ are mixture index sequences of conversion and training data, respectively. Here, $P(\boldsymbol{x}, \boldsymbol{m}|\boldsymbol{\lambda})$ and $P(\boldsymbol{X}, \boldsymbol{M}|\boldsymbol{\lambda})$ are the complete likelihood function of GMM for conversion and training data respectively, and $P(\boldsymbol{\lambda})$ is a prior distribution of model parameters. In the proposed method, the marginal likelihood $P(\boldsymbol{x}, \boldsymbol{X})$ is approximated using the variational method. The lower bound of log marginal likelihood $\mathcal{F}$ is defined by introducing $Q(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda})$ to a marginalized likelihood function that is the log of Eq. (17), and by using Jensen's inequality.

$$\log P\left(\boldsymbol{x}, \boldsymbol{X}\right) = \log\sum_{\boldsymbol{m}}\sum_{\boldsymbol{M}}\int Q\left(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda}\right)\frac{P\left(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda}\right)}{Q\left(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda}\right)}d\boldsymbol{\lambda} \geq \left\langle \log\frac{P\left(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{X}\right)}{Q\left(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda}\right)}\right\rangle_{Q(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda})} = \mathcal{F} \quad (18)$$

where $\langle\cdot\rangle_{Q(x)}$ is the expectation of $Q(x)$. The lower bound of log marginal likelihood $\mathcal{F}$ is the objective function with variational function $Q(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda})$, and any distribution $Q(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda})$ is given some constraint conditions to enable an integral calculation to be solved. Thus, the optimal approximate distribution can be solved that maximize the lower bound of log marginal likelihood $\mathcal{F}$. Here, the following constraint conditions are given.

$$Q\left(\boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda}\right) = Q\left(\boldsymbol{m}\right)Q\left(\boldsymbol{M}\right)Q\left(\boldsymbol{\lambda}\right) \quad (19)$$

The optimal distributions of $Q(\boldsymbol{\lambda})$, $Q(\boldsymbol{M})$, and $Q(\boldsymbol{m})$ are obtained by maximizing $\mathcal{F}$.

$$Q(\boldsymbol{\lambda}) \propto \exp\langle \log P(\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{M}, \boldsymbol{\lambda})\rangle_{Q(\boldsymbol{m})Q(\boldsymbol{M})} \quad (20)$$

$$Q(\boldsymbol{M}) \propto \exp\langle \log P(\boldsymbol{X}, \boldsymbol{M}|\boldsymbol{\lambda})\rangle_{Q(\boldsymbol{\lambda})} \quad (21)$$

$$Q(\boldsymbol{m}) \propto \exp\langle \log P(\boldsymbol{x}, \boldsymbol{m}|\boldsymbol{\lambda})\rangle_{Q(\boldsymbol{\lambda})} \quad (22)$$

where $Q(\boldsymbol{m})$ and $Q(\boldsymbol{M})$ depend on each other. These updates should be iterated as the EM algorithm, which increases the value of objective function $\mathcal{F}$ at each iteration until convergence.

$$\overline{\boldsymbol{x}}^{(2)} = \underset{\boldsymbol{x}^{(2)}}{\operatorname{argmax}}\, \mathcal{F} \quad (23)$$

where $\mathcal{F}$ can be defined by $Q(\boldsymbol{\lambda})$, $Q(\boldsymbol{m})$, and $Q(\boldsymbol{M})$.

# 4. Voice Conversion Using Multiple Structures

In voice conversion based on a GMM using multiple structures, multiple GMMs of different numbers of mixtures form a single statistical model. If $N$ GMMs are integrated, the likelihood of observation vector sequence $\boldsymbol{\Lambda} = \{p^{(n)}, \boldsymbol{\lambda}^{(n)} | n = 1, 2, \ldots, N\}$ for the model parameters of the integrated model is shown by the following expressions.

$$\begin{aligned} & P(\boldsymbol{x}|\boldsymbol{\Lambda}) \\ &= \prod_{t=1}^{T} \sum_{n=1}^{N} p^{(n)} P\left(\boldsymbol{x}_t \mid n, \boldsymbol{\lambda}^{(n)}\right) \\ &= \prod_{t=1}^{T} \sum_{n=1}^{N} p^{(n)} \left[\sum_{i=1}^{M_n} \omega_i^{(n)} \mathcal{N}\left(\boldsymbol{x}_t | \boldsymbol{\mu}_i^{(n)}, \boldsymbol{\Sigma}_i^{(n)}\right)\right] \end{aligned} \quad (24)$$

where $M_n$ is the mixture of the $n$-th GMM, and $p^{(n)}$ is the integration weight for each GMM and defined by $\sum_{n=1}^{N} p^{(n)} = 1$. The model structure of the output probability using multiple GMMs is shown in Fig. 1. Eq. (24) can be expressed as the follows by reversing:

$$P(\boldsymbol{x}|\boldsymbol{\Lambda}) = \prod_{t=1}^{T} \sum_{n=1}^{N} \sum_{i=1}^{M_n} p^{(n)} \omega_i^{(n)} \mathcal{N}\left(\boldsymbol{x}_t | \boldsymbol{\mu}_i^{(n)}, \boldsymbol{\Sigma}_i^{(n)}\right) \quad (25)$$

If $p^{(n)}\omega_i^{(n)}$ is regarded as new weight parameter $\hat{\omega}_i^{(n)}$, Eq. (25) is in the same form as the likelihood function for a GMM that the mixture is $\sum_{i=1}^{N} M_n$. Spectral conversion using the optimal model at each frame can be achieved by adjusting the spectral conversion based on the GMM. Additionally, the integration weight $p^{(n)}$ must be set appropriately in the proposed method. Two weight determination methods are discussed in this paper: uniform distribution and the Bayesian Information Criterion (BIC)[8].

### 4.1. Setting Weight Based on Uniform Distribution

Uniform distribution is the easiest integration weight determination method to use. This means $p^{(n)} = 1/N$. In this case, no prior information is assigned for the appropriate number of mixtures.

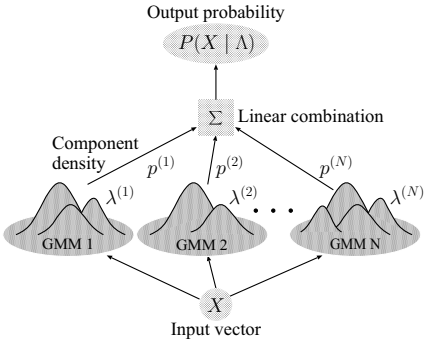


Figure 1: *Output probability for multiple structures.*

### 4.2. Setting Weight Based on Bayesian Information Criterion

Estimation of the integration weights can be regarded as a model structure selection problem. Therefore, this work applies BIC to the weight determination method. When the likelihood function of the training vector sequence $\boldsymbol{x} = \left[\boldsymbol{x}_1{}^\top, \boldsymbol{x}_2{}^\top, \ldots, \boldsymbol{x}_T{}^\top\right]^\top$ is $\ln P(\boldsymbol{x}|\boldsymbol{\lambda})$, BIC is defined as

$$\mathrm{BIC} = \ln P(\boldsymbol{x}|\boldsymbol{\lambda}) - \frac{k}{2} \ln T \quad (26)$$

where $k$ is the number of free parameters of the model and $T$ is the number of samples. The likelihood increases when the number of mixtures grows. In contrast, the penalty term (the second term in Eq. (26)) decreases BIC in proportion to the number of parameters, and BIC selects the model structure that has a good balance when fitting to training data. Thus, appropriate weights corresponding to the amount of training data can be obtained by a BIC-based approach. In the integration weight determination based on BIC, the integration weight $p^{(n)}$ is defined as

$$p^{(n)} = \frac{\exp \mathrm{BIC}^{\frac{\alpha}{T}}(n)}{\sum_{j=1}^{N} \exp\left(\mathrm{BIC}^{\frac{\alpha}{T}}(j)\right)} \quad (27)$$

where BIC($n$) is the BIC measure of the $n$-th GMM, and $\alpha$ is an adjustive parameter for smoothing.

# 5. Experiment

Obtained voice conversion experiments using the ATR Japanese speech database b-set were carried out to evaluate the effects of the proposed method. Two male speakers were selected as a source and a target speaker (source: MTK, target: MHT). The speech data were down-sampled from 20 to 16 kHz, windowed at a 5-ms frame rate using a 25-ms Blackman window, and parameterized into 24 mel-cepstral coefficients excepting the 0-th coefficients. Their first order derivatives were used as the dynamic features. The mel-cepstral distortion was used as the objective evaluation criterion. This distortion was calculated using the distance between the mel-cepstrum of the converted utterance and the target utterance as follows:

$$\text{Mel-CD} = \frac{10}{\ln 10} \sqrt{2 \sum_{s=1}^{D} \left(c_s^{(1)} - c_s^{(2)}\right)^2} \quad (28)$$

where $c_s^{(1)}$ and $c_s^{(2)}$ were the $s$-th mel-cepstrum coefficients extracted from the target utterance and the converted utterance, respectively.

In the experiment, the following methods were compared.

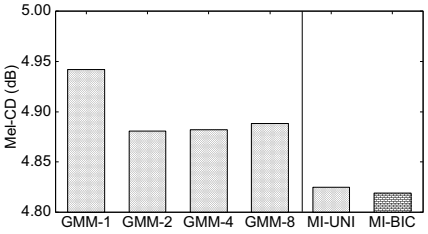


Figure 2: *Results of Objective Evaluation (10 utterances training data).*

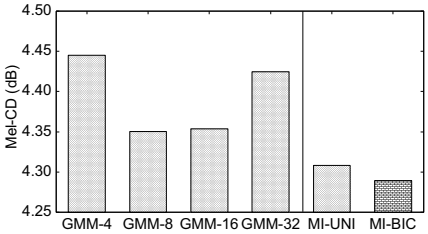


Figure 3: *Results of Objective Evaluation (50 utterances training data).*

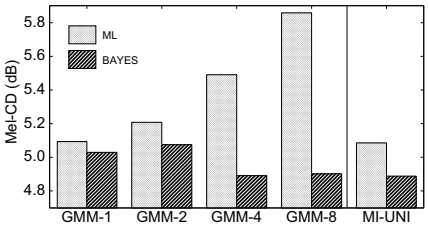


Figure 4: *Results of Objective Evaluation (speaker independent, three utterances training data).*

- GMM: standard GMMs
- MI-UNI: Model Integration of GMMs using uniform weights
- MI-BIC: Model Integration of GMMs using BIC-based weights

First, to confirm the effects of model integration, experiments were carried out using the ML criterion. The results for 10 and 50 utterances are shown in Figs. 2 and 3 respectively. The number of mixtures was varied among 1, 2, 4, and 8. The adjustive parameter of $\alpha$ in the BIC-based method was set to 4.0 for the 10 utterances test and 3.0 for the 50 utterances test.

By comparing GMMs with a varying number of mixtures, it can be seen that the optimal number of mixtures depended on the amount of training data. The distortions of the integrated models (MI-UNI and MI-BIC) were smaller than those of the conventional GMMs for both the 10 and 50 utterances training data. This result indicated that the appropriate number of mixtures was successfully selected stochastically at each frame by the smoothed model structure obtained from the integration of multiple GMMs.

When comparing MI-UNI and MI-BIC, MI-BIC obtained no significant improvement over MI-UNI, even though the adjustable parameter $\alpha$ was determined to minimize the distortion. Therefore, the integration weights based on uniform distribution were used in the next experiment.

In the second experiment, the model integration using the Bayesian criterion was evaluated. The results are shown in Fig. 4. In this experiment, three utterances were used as the training data. The VB learning in the integrated GMMs was examined by comparing the ML criterion (ML) and Bayesian criterion (BAYES). Prior distributions in BAYES were given by using universal background models. From the results of the GMMs, the Bayesian approach achieved smaller distortion than ML. The reason for this result was that the overtraining problem was decreased by using VB method. The Bayesian method still outperformed the ML method when the model structures were integrated. Although there was no significant effect of the model structure integration, the performance was comparable to the best result of a single GMM (GMM-4). It seems that the proposed method can automatically select the optimal models without determining the optimal number of mixtures beforehand.

## 6. Conclusions

A Bayesian approach to voice conversion based on a GMM using multiple models of different numbers of mixtures was proposed and evaluated. Experimental results in an objective evaluation showed that model structure integration was effective for the ML criterion and that appropriate structures were selected automatically for the Bayesian criterion. Future works is to carry out subjective evaluation to show the effectiveness of the proposed method.

## 7. Acknowledgement

The research leading to these results was partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communication, Japan.

## 8. References

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *Proc. of IEEE Trans. Speech Audio*, vol. 6, pp. 131–142, Mar. 1998.

[2] T. Toda, *et al.*, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[3] P. Lanchantin, X. Rodet, " Dynamic Model Selection for Spectral Voice Conversion, " *INTERSPEECH 2010*, pp.1720-1724, Sep. 2010.

[4] A. Kain, M. W. Macon, "Spectral Voice Conversion for Text-to-Speech Synyhesis", *Proc. of ICASSP*, vol.1, pp. 285–288, May 1998.

[5] A. Mouchtaris, J. V. Spiegel, P. Mueller, "Nonparallel Training for Voice Conversion Based on a Parameter Adaptation Approach," *Proc. of IEEE Trans. Speech Audio*, vol. 14, pp. 952–963, no. 3, May 2006.

[6] Z. Ghahramani and M. I. Jordan, "On Structured Variational Approximations," *University of Toronto Technical Repor*t, 1997, revised 2002.

[7] M. I. Jordan, Z. Ghahramani, T, S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, vol. 37, pp. 183–233, 1999.

[8] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, pp. 461–464, no. 2, Jul. 1978.

[9] Y. Nankaku, K. Nakamura, T. Toda, and K. Tokuda, "Spectral Conversion Based on Statistical Models Including Time-Sequence Matching," *Proc. of ISCA Speech Synthesis Workshop*, pp. 333–338, Aug. 2007.