HIGH ACCURATE MODEL-INTEGRATION-BASED VOICE CONVERSION USING DYNAMIC FEATURES AND MODEL STRUCTURE OPTIMIZATION

Daisuke Saito[1]*, Shinji Watanabe*[2]*, Atsushi Nakamura*[2]*, and Nobuaki Minematsu*[1]

[1]The University of Tokyo, Tokyo, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
{dsk_saito,mine}@gavo.t.u-tokyo.ac.jp, {watanabe,ats}@cslab.kecl.ntt.co.jp

ABSTRACT

This paper combines a parameter generation algorithm and a model optimization approach with the model-integration-based voice conversion (MIVC). We have proposed probabilistic integration of a joint density model and a speaker model to mitigate a requirement of the parallel corpus in voice conversion (VC) based on Gaussian Mixture Model (GMM). As well as the other VC methods, MIVC also suffers from the problems; the degradation of the perceptual quality caused by the discontinuity through the parameter trajectory, and the difficulty to optimize the model structure. To solve the problems, this paper proposes a parameter generation algorithm constrained by dynamic features for the first problem and an information criterion including mutual influences between the joint density model and the speaker model for the second problem. Experimental results show that the first approach improved the performance of VC and the second approach appropriately predicted the optimal number of mixtures of the speaker model for our MIVC.

Index Terms— Voice conversion, probabilistic integration, dynamic features, information criterion

1. INTRODUCTION

Voice conversion (VC) is a technique to transform an inputted utterance of a speaker to another utterance that sounds like another speaker's voice without changing the linguistic content. VC can be regarded as a technique to modify inputted features to features of a desired target. Then VC techniques have potentials of applying to many research areas of speech processing besides speech synthesis or speech generation [1, 2].

To derive appropriate features of a target speaker from a source speaker's features by VC techniques, two important functions should be considered; to model the proper correspondence of the source features to the target features, and to represent a feature space of the target precisely. Although there have been several proposed techniques for voice conversion based on statistical approaches [1, 3, 4], they strongly focus on the first function. To realize this function, they require the parallel corpus for training, which contains plenty of utterances with the same linguistic content both the source and the target. On the other hand, we have proposed the model-integration-based voice conversion (MIVC) which focuses not only on the first function, but also on the second function, i.e., to model the precise feature space of the target speaker [5]. Our method uses non-parallel speech data of the target speaker to construct the speaker model of the target. Then it effectively mitigates the data sparse problem caused by the requirement of the parallel corpus. There are other approaches focusing on the efficient use of non-parallel data [6, 7]. They have applied parameter adaptation techniques to parameters of *the joint density model*, which is constructed to model the relation between the source and the target speakers. On the other hand, our proposed approach independently constructs *the speaker model* of the target, and integrates it with the joint density model by a probabilistic manner. Therefore it works well even if the amount of training data for the joint density model is small.

In this paper, we try other two problems in voice conversion studies; the degradation of the perceptual quality of the converted speech caused by the discontinuity through the parameter trajectory, and the difficulty to optimize the model structure of conversion models. The first problem is mainly caused by the frame-by-frame mapping where the correlation of the target feature vectors between frames is not considered. Our MIVC also suffers from this problem. In addition, since parameters in the target speaker model in MIVC are independent of a feature sequence of the source speaker, inappropriate spectral movement can occur more often than the conventional VC methods even if each frame in the converted features is modeled more precisely.

The second problem, the determination of an optimal model structure, is one of the most difficult problems in statistical acoustic modeling. For example, in the conventional GMM-based voice conversion, if the number of Gaussian components is increased unnecessarily, it causes the degradation of the performance of the conversion for test sentences. It is well-known as the over-training effect. In our case of MIVC, optimization of model structure is more difficult because mutual influences between the joint density model and the speaker model should be considered.

For the above problems, there have been several proposed approaches in various areas; filter-based approach [8], maximum likelihood estimation of the parameter trajectory [9, 10] for the first problem, and acoustic modeling based on the MDL criterion [11] or variational Bayesian treatment [12] for the second one. Considering them, in this paper, we employ two approaches in our method; a parameter generation algorithm using dynamic features for the first problem and a model optimization based on an information criterion including mutual influences between both the models for the second one. Experimental results show that the first approach improved the performance of VC and the second one appropriately predicted the optimal number of components of the speaker model for our MIVC.

2. MODEL-INTEGRATION-BASED VOICE CONVERSION

This section briefly describes the joint density GMM method [1] and our model-integration-based voice conversion (MIVC) [5]. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}]$ be a vector sequence characterizing an utterance from the source speaker, and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}]$ be that

of the target speaker. When $\boldsymbol{x}_t$ is given, the optimal parameter generation of $\boldsymbol{y}_t$ is based on the conditional probability $P(\boldsymbol{y}_t|\boldsymbol{x}_t)$. The important points of GMM-based VC are how to derive this probability and how to optimize it for the parameter generation.

In voice conversion based on the joint density GMM, $P(\boldsymbol{y}_t|\boldsymbol{x}_t)$ is derived from the probability density of $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$, i.e., the joint vector of the source and the target feature vectors. The notation $^\top$ denotes transposition of the vector. The joint probability density of the source and the target vectors is modeled by a GMM for the joint vector $\boldsymbol{z}_t$ as follows:

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \tag{1}$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}, \tag{2}$$

where $\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ denotes the Gaussian distribution with mean vector $\boldsymbol{\mu}_m^{(z)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$, m is the mixture component index, and the total number of mixture components is M. The weight of the m-th component is w_m. Deriving $P(\boldsymbol{y_t}|\boldsymbol{x}_t)$ with the above parameters and minimizing the mean square error, a mapping function $\mathcal{F}(\cdot)$ to convert the source vector $\boldsymbol{x}_t$ to the target vector $\boldsymbol{y}_t$ is derived as

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{m=1}^{M} \frac{w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})} \boldsymbol{E}_{m,t}^{(y)}, \tag{3}$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}). \tag{4}$$

On the other hand, our MIVC focuses on the following optimization problem of $P(\boldsymbol{y}_t|\boldsymbol{x}_t)$ derived by the Bayes rule:

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\operatorname{argmax}} \underbrace{P(\boldsymbol{x}_t|\boldsymbol{y}_t)}_{\text{from joint density model}} \underbrace{P(\boldsymbol{y}_t)}_{\text{from speaker model}}. \tag{5}$$

Voice conversion should have two important functions; to ensure the proper correspondence of the source features to the target ones that keeps the linguistic content, and to model the speaker individuality of the target. In Equation 5, the first term is derived from the joint density model as Equation 1 and realizes the former function. The second term $P(\boldsymbol{y}_t)$ is derived from the speaker GMM trained by a non-parallel corpus of the utterances of the target speaker, and it corresponds to the latter function.

From the following likelihood function based on Equation 5, a parameter generation algorithm is derived [1]. Let $\boldsymbol{\lambda}^{(s)}$ be the parameters of the speaker model.

$$\mathcal{L}(\boldsymbol{y}_t; \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(s)}) \triangleq P(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{\lambda}^{(s)}). \tag{6}$$

For the optimum solution $\hat{\boldsymbol{y}}_t$ to maximize the function $\mathcal{L}$, the auxiliary function with respect to $\hat{\boldsymbol{y}}_t$ is derived [5]. Finally the following updating equations are derived:

$$\hat{\boldsymbol{y}}_t = \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(y)-1} + \sum_{n=1}^{N} \gamma_{n,t} \boldsymbol{\Sigma}_n^{-1} \right)^{-1} \times \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(y)-1} \boldsymbol{E}_{m,t}^{'(y)} + \sum_{n=1}^{N} \gamma_{n,t} \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_n \right), \tag{7}$$

$$\gamma_{m,t} = P(m|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}), \gamma_{n,t} = P(n|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(s)}), \tag{8}$$

[1] As well as a language model weight in ASR, a weight factor to control the balance between the models also can be derived.

where $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are the mean vector and the covariance matrix of the n-th component in the speaker GMM, and

$$\boldsymbol{E}_{m,t}^{'(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yy)} \boldsymbol{\Sigma}_m^{(xy)+} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}), \tag{9}$$

$$\boldsymbol{D}_m^{'(y)-1} = \left[\boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)} \right]^{-1} - \boldsymbol{\Sigma}_m^{(yy)-1}. \tag{10}$$

The notation $^+$ denotes the pseudo-inverse of the matrix. For the initial values of $\gamma_{m,t}$ and $\gamma_{n,t}$, $P(m|\mathcal{F}(\boldsymbol{x}_t), \boldsymbol{\lambda}^{(z)})$ and $P(n|\mathcal{F}(\boldsymbol{x}_t), \boldsymbol{\lambda}^{(s)})$ are used respectively. Equation 7 becomes the weighted summation of the effects from the joint density model and those from the speaker model. Thus, MIVC can overcome the sparse parallel data problem by reducing the over-estimation effects of the joint density parameters by the speaker model.

3. PARAMETER GENERATION USING DYNAMIC FEATURES AND MODEL OPTIMIZATION

3.1. Constraint from dynamice features

In the frame-by-frame mapping where the correlation of the feature vectors is ignored, the discontinuity of the parameter trajectory becomes a problem. In MIVC, this is more serious than the conventional VC, because even slight skips worse affect the perceptual quality of the whole sentence since each frame in the sequence is converted more precisely. Besides, skips can occur more often because the speaker model of the target is independent of the source features. To compensate for the discontinuity, several approaches that smooth the output parameter sequence have been proposed. Chen *et al*. applied a median filter and a low pass filter for the parameter generation in VC to smooth the parameter trajectory [8]. Toda *et al*. proposed the maximum likelihood estimation of the spectral parameter trajectory considering dynamic features [9].

In this paper, we also employ the parameter generation considering dynamic features to our MIVC by the similar manner as that of the approaches in [9]. From here, let a time sequence of the source features and that of the target ones be $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$, respectively. $\boldsymbol{X}_t = [\boldsymbol{x}_1^\top, \boldsymbol{\Delta x}_1^\top]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_1^\top, \boldsymbol{\Delta y}_1^\top]^\top$ consist of static and dynamic features. $\boldsymbol{\lambda}^{(Z)}$ and $\boldsymbol{\lambda}^{(S)}$ are trained by these features as well as the conventional MIVC. A time sequence of the converted feature vectors $\hat{\boldsymbol{y}}$ in MIVC is derived as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Z)}) P(\boldsymbol{Y}|\boldsymbol{\lambda}^{(S)}) \tag{11}$$

$$\text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \tag{12}$$

where $\boldsymbol{W}$ denotes the matrix to extend the static feature sequence to the static and dynamic feature sequence. By the similar manner as that in [9] and [5], we derive the following updating equations:

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)-1}} \boldsymbol{W} \right) \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)-1} \boldsymbol{E}^{(Y)}}, \tag{13}$$

$$\overline{\boldsymbol{D}^{(Y)-1}} = \operatorname{diag} \left[\overline{\boldsymbol{D}_1^{(Y)-1}}, \ldots, \overline{\boldsymbol{D}_T^{(Y)-1}} \right], \tag{14}$$

$$\overline{\boldsymbol{D}^{(Y)-1} \boldsymbol{E}^{(Y)}} = \left[\overline{\boldsymbol{D}_1^{(Y)-1} \boldsymbol{E}_1^{(Y)}}^\top, \ldots, \overline{\boldsymbol{D}_T^{(Y)-1} \boldsymbol{E}_T^{(Y)}}^\top \right]^\top, \tag{15}$$

$$\overline{\boldsymbol{D}_t^{(Y)-1}} = \left(\textstyle\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(Y)-1} + \sum_{n=1}^{N} \gamma_{n,t} \boldsymbol{\Sigma}_n^{(S)-1} \right), \tag{16}$$

$$\overline{\boldsymbol{D}_t^{(Y)-1} \boldsymbol{E}_t^{(Y)}} = \left(\textstyle\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(Y)-1} \boldsymbol{E}_{m,t}^{'(Y)} + \sum_{n=1}^{N} \gamma_{n,t} \boldsymbol{\Sigma}_n^{(S)-1} \boldsymbol{\mu}_n^{(S)} \right), \tag{17}$$

$$\gamma_{m,t} = P(m|\boldsymbol{Y}_t, \boldsymbol{\lambda}^{(Z)}), \gamma_{n,t} = P(n|\boldsymbol{Y}_t, \boldsymbol{\lambda}^{(S)}). \tag{18}$$

Compared Equation 13 with MLE-based method in [9], the proposed generation has the similar form, but it includes the effects of the independent speaker GMM. Compared it with Equation 7, the proposed generation is regarded as MIVC constrained from dynamic features. Thus the proposed generation is advanced from both the MLE-based method and the conventional MIVC.

3.2. Model optimization using an information criterion

For controlling the complexity of statistical models, information criteria that contain the number of free parameters as the penalty factor are often used [11]. Bayesian information criterion (BIC) defined by the following equation is adopted in some studies:

$$BIC = -2\log(\mathcal{L}) + k\log(n), \tag{19}$$

where $\mathcal{L}$ is a likelihood function of a model, k is the number of free parameters in the model, and n is the number of training data for the model. The second term works as the penalty of the complicated model structure. BIC is equivalent to MDL when Gaussian distributions are focused on. In our case, however, BIC does not perfectly work for the model optimization. For example, in the case of a male speaker GMM trained by 50 sentences, $N = 128$ or 256 should be optimal according to BIC. However, the optimal N derived from the distorion is actually 16 in the previous study [5]. To decide the number of mixtures in the speaker GMM for MIVC, we should also consider mutual influences between the joint density model and the speaker model. Then we modify BIC to the following information criterion considering both the models:

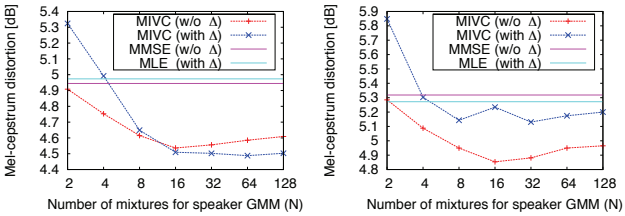
$$BIC' = -2\log(\mathcal{L}_{zs}) + \frac{n_z}{n_s}\left(k\log(n_s)\right), \tag{20}$$

where $\mathcal{L}_{zs}$ means the likelihood function of the speaker model for the training data of *the joint density model*, k_s is the number of free parameters in the speaker model, n_z and n_s are the number of training data for the joint density model and the speaker model, respectively. The former term in Equation 20 focuses on the mutual influences between both the models, and the latter means the penalty factor for the complicated speaker GMM. Although Equation 20 is derived heuristically, we preliminarily use it for our MIVC. For further improvements, we are planning to deal with MIVC on a Bayesian framework including the optimization of model structure [12]. To optimize the model structure, the number of mixtures that minimize Equation 20 is selected.

4. EXPERIMENT

4.1. Experimental conditions

To evaluate the performance of parameter generation using dynamic features in MIVC, voice conversion experiments using Japanese sentences were performed. This experiment used speech samples from 5 speakers (MSH as the source, MMY, MTK, FKS, and FTK as the targets) in the ATR Japanese speech database B-set [13]. The first letters of the speaker names correspond to gender. This database consists of 503 phonetically balanced sentences. They are divided into 9 subsets (subset A to I) consisting of 50 sentences and subset J consisting of 53 sentences. We selected the subset J for test data. For training of the joint density models, one sentence pair was used. The total number of mixture components (M) was fixed to 8 for the methods without dynamic features, and 16 for the methods using dynamic features. On the other hand, for the speaker GMMs, 50 sentences in the subset I were selected and the GMM for each



(a): Male to male conversion (b): Male to female conversion

Fig. 1. Results of averaged distorion as a function of N. M is 8 for "w/o Δ" and 16 for "with Δ".

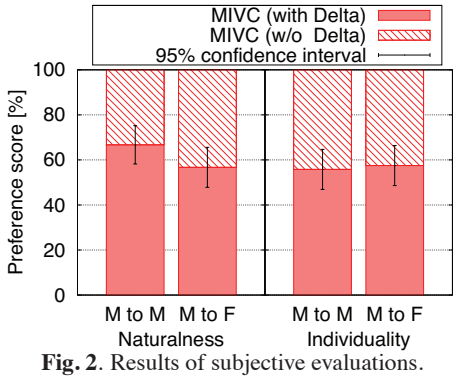


Fig. 2. Results of subjective evaluations.

speaker was trained. The number of mixture components for the speaker GMM (N) was varied from 2 to 128. On the other hand, the model optimization based on Equation 20 was also carried out. The number of iterations for Equation 7 and 13 was fixed to 5. We used 24-dimensional mel-cepstrum vectors for spectrum representation. These are derived by STRAIGHT analysis [14]. Aperiodic components are fixed to -30 dB at all frequencies. The power coefficient and the fundamental frequency were converted in a simple manner that only considers the mean and the standard deviation of the parameters. We compared the MMSE-based method (Eq.3, w/o Δ), the MLE-based method (Eq.13, with Δ), and MIVC with and without Δ parameters.

4.2. Effects of dynamic features

We evaluated the effects of dynamic features by both the objective and the subjective evaluations. For the objective evaluation, we evaluated the conversion performance using mel-cepstral distortion between the converted vectors and the vectors of the targets. For the subjective evaluation, a listening test was carried out to evaluate the naturalness of converted speech and conversion accuracy for speaker individuality. The test was conducted with 10 subjects to compare the utterances converted by the MIVC methods with and without dynamic features. For the MIVC method without dynamic features, the number of mixtures of the joint density model (M) was fixed to 8, and the number of mixtures of the speaker GMM (N) was 16. For the MIVC method with dynamic features, the number of mixtures of the joint density model (M) was fixed to 16. We selected the optimal number for each speaker as the number of mixtures of the speaker GMM (N). To evaluate naturalness, a paired comparison was carried out. In this test, pairs of two different types of the converted speech samples were presented to subjects, and then each subject judged which sample sounded better. To evaluate conversion accuracy, an RAB test was performed. In this test, pairs of two different types of the converted samples were presented after presenting the reference sample of the target speech. The number of sample pairs evaluated by each subject was 24 in each test.

Table 1. The optimal and selected numbers of mixtures for the speaker GMM from BIC and BIC'. The values in parentheses mean the mel-cepstral distortion on each condition.

speaker	Optimal	BIC [Eq.19]	BIC' [Eq. 20]
MMY (w/o Δ)	16 (4.54)	128 (4.63)	32 (**4.58**)
MTK (w/o Δ)	16 (4.53)	256 (4.63)	64 (**4.56**)
FKS (w/o Δ)	16 (4.80)	256 (4.85)	16 (**4.80**)
FTK (w/o Δ)	16 (4.90)	256 (5.11)	64 (**5.04**)
MMY (with Δ)	128 (4.46)	128 (**4.46**)	64 (4.48)
MTK (with Δ)	64 (4.49)	256 (4.56)	64 (**4.49**)
FKS (with Δ)	8 (5.16)	128 (**5.23**)	32 (5.26)
FTK (with Δ)	32 (5.00)	128 (5.16)	64 (**5.08**)

Figure 1 shows the result of average mel-cepstral distortion for the test data as a function of the number of mixture components of *the speaker GMM* and Figure 2 shows preference scores as the results of the subjective evaluation. From Figure 1, compared with MMSE and MLE, MIVC results were better. These results show that MIVC appropriately compensated for the sparse data problem of the joint density model. In comparison of MIVC methods by mel-cepstral distortion, dynamic features reduced the mel-cepstral distortion in the case of male to male conversion, while they did not in the case of male to female conversion. On the other hand, according to the subjective scores, MIVC with dynamic features outperformed MIVC without dynamic features in both the cases; intra-gender and cross-gender. In Figure 1, the similar results also can be found between the MMSE method and the MLE method. That is to say, it is difficult to evaluate the effects of dynamic features precisely by mel-cepstral distortion. However, the subjective scores show the effects that the discontinuity of the spectral trajectory was mitigated by considering the constraint of dynamic features, and the perceptual qualities of the converted speech were improved.

4.3. Effects of model optimization

Table 1 is the comparison of the model optimization for each target speaker. The left column in Table 1 is the optimal number of mixtures selected by the mel-cepstral distortion of MIVC methods. Both BIC and the modified BIC (BIC') constantly worked well for the model optimization. This result confirmed the effects of information criteria for the model optimization. In the case that dynamic features are not used, the model optimization based on BIC' selected better models than those selected by BIC. That is to say, BIC' appropriately captures the influence from the joint density model. Although the proposed information criterion is derived heuristically, the influence from the joint density model trained by a little amount of corpus could be properly included in this criterion. On the other hand, in the case that dynamic features are used, selected models by BIC and BIC' were comparable to both of them. In order to derive more robust optimization of model structure, it is required to deal with MIVC on a full Bayesian framework [12].

5. CONCLUSIONS

We have proposed two approaches for improving model-integration-based voice conversion (MIVC); the parameter generation using the constraint of the dynamic features and the model optimization approach for the speaker GMM based on the information criterion where mutual influences between the joint density model and the speaker model are considered. These approaches mitigate the difficulties with MIVC; the discontinuity through the parameter trajectory and the optimization of model structure for the speaker model. For further improvements of the conversion performance, we are planning to deal with MIVC on a Bayesian framework, which can use the prior knowledge to both the models and apply appropriate selection of model structures for both the models [12].

6. REFERENCES

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP, vol. 1, pp. 285–288, 1998.

[2] A. Kunikoshi, Y. Qiao, N. Minematsu, and K. Hirose, "Speech generation from hand gestures based on space mapping," in Proc. INTERSPEECH, pp. 308–311, 2009.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[4] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in Proc. ICASSP, pp. 3893–3896, 2009.

[5] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Probabilistic integration of joint density model and speaker model for voice conversion," in Proc. INTERSPEECH, pp. 1728–1731, 2010.

[6] A. Mouchtaris, J. V. der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 3, pp. 952–963, 2006.

[7] C. H. Lee and C. H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in Proc. INTERSPEECH, pp. 2254–2257, 2006.

[8] Y. Chen, M. Chu, E. Chang, J. Jiu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," in Proc. EUROSPEECH, pp. 2413–2416, 2003.

[9] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in Proc. ICASSP, pp. 660–663, 1995.

[11] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in Proc. EUROSPEECH, vol. 1, pp. 99–102, 1997.

[12] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," IEEE Trans. on Speech and Audio Processing, vol. 12, pp. 365–381, 2004.

[13] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K.Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol.9, pp. 357–363, 1990.

[14] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, "Restructuring speech representations 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.27, pp. 187–207, 1999.