Probabilistic Integration of Joint Density Model and Speaker Model for Voice Conversion

Daisuke Saito[1,2]*, Shinji Watanabe*[2]*, Atsushi Nakamura*[2]*, Nobuaki Minematsu*[3]

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

Abstract

This paper describes a novel approach to voice conversion using both a joint density model and a speaker model. In voice conversion studies, approaches based on Gaussian Mixture Model (GMM) with probabilistic densities of joint vectors of a source and a target speakers are widely used to estimate a transformation. However, for sufficient quality, they require a parallel corpus which contains plenty of utterances with the same linguistic content spoken by both the speakers. In addition, the joint density GMM methods often suffer from over-training effects when the amount of training data is small. To compensate for these problems, we propose a novel approach to integrate the speaker GMM of the target with the joint density model using probabilistic formulation. The proposed method trains the joint density model with a few parallel utterances, and the speaker model with non-parallel data of the target, independently. It eases the burden on the source speaker. Experiments demonstrate the effectiveness of the proposed method, especially when the amount of the parallel corpus is small.

Index Terms: voice conversion, joint density model, speaker model, probabilistic unification

1. Introduction

Voice conversion (VC) aims at transforming a speaker's voice to make it sound like another speaker's without changing the linguistic content. Applications of VC include the modification of speaker identity in Text-to-Speech (TTS) systems [1], noisy speech to clean speech for speech enhancement [2], hand motion to speech conversion [3], and so on. Since spectral features have a very important role in representing speaker individuality of voices, most current conversion systems mainly focus only on the transformation of spectral features.

There are many ways to implement the conversion from source features to target ones. Statistical approaches have often been used for estimating the transformation, such as codebook mapping method [4], artificial neural networks [5], or Gaussian mixture models (GMM) [1, 6]. Among these, GMM-based approaches are widely used in particular because of their flexibility. GMM-based techniques for statistical mapping use a mixture of Gaussians to model the probabilistic densities of source feature vectors [6] or those of joint vectors of the source and the target speakers [1]. Both approaches derive the transformation function as a weighted summation of linear transformations, each corresponding to each Gaussian component, while the weights are calculated as posterior probabilities of source vectors. Since the latter approach estimates a joint density of the source and the target vectors by allocating mixtures in a single feature space, it can model the relationship between the source and the target feature spaces more precisely [1].

However, the joint density GMM methods absolutely require the training corpus, which contains plenty of utterances with the same linguistic content from both the speakers to achieve sufficient quality. In addition, they suffer from over-training effects when the number of utterance pairs for training is small, since the dimensionality of the vector space is estimated to double [1]. To solve these problems, there have been several proposed approaches which do not require a parallel corpus [7, 8]. They have applied parameter adaptation techniques to parameters of the joint density model, using non-parallel speech data. In this paper, we propose another method to compensate for these problems. In our approach, the function of VC is divided into two functions; to ensure the consistency of the linguistic content between both the speakers, and to model the speaker individuality of the target. The proposed method realizes these functions by different and independent models, i.e. the joint density model constructed by a small parallel corpus and the speaker model trained by a non-parallel but large speech corpus of the target speaker. Finally it integrates the two functions into one function of VC using probabilistic formulation. Since the proposed method can train the joint density model and the speaker model separately, it has the potential to apply precise modeling techniques proposed independently in each research area; such as eigenvoice conversion in VC studies [9] or approaches based on the universal background model in speaker recognition studies [10]. The proposed method also eases the burden on the source speaker, since it is expected to work well when the number of training data for the joint density model is small.

The remainder of this paper is organized as follows. Section 2 describes the conventional GMM-based VC approach using the joint density model [1]. Then, in Section 3, our proposed approach using both the joint density model and the speaker model is described. In Section 4, experimental evaluations are described. Finally Section 5 concludes the paper.

2. GMM-based voice conversion

In this section, the joint density GMM method [1] is briefly described. 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. Note that the two utterances contain the same linguistic content. The dynamic time warping algorithm (DTW) is applied to align the source vectors to their corresponding vectors in the target sequence. Then, a new sequence of joint vectors

$\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]$ where $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ is created. 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:

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

In Equation 1, $\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ denotes the normal 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 and $\sum_{m=1}^{M} w_m = 1$. $\boldsymbol{\lambda}^{(z)}$ denotes a parameter set of the GMM, which consists of weights, mean vectors, and covariance matrices for individual mixture components. Since the feature space of the joint vector $\boldsymbol{z}$ includes the feature spaces for the source and the target speakers as its subspaces, $\boldsymbol{\mu}_m^{(z)}$ and $\boldsymbol{\Sigma}_m^{(z)}$ are written as

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

where $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ are the mean vector of the m-th component for the source and that for the target, respectively. Similarly, the matrices $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ are the covariance matrix of the m-th component for the source and that for the target, respectively. The matrices $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ are the cross-covariance matrices of the m-th component for the source and the target. These parameters in the GMM are estimated by the EM algorithm using the sequence of the joint vectors ($\boldsymbol{Z}$).

A mapping function $\mathcal{F}(\cdot)$ to convert the source vector $\boldsymbol{x}_t$ to the target vector $\boldsymbol{y}_t$ is derived based on the conditional probability density of $\boldsymbol{y}_t$, given $\boldsymbol{x}_t$. This probability density can be represented by the parameters of the joint density model as follows:

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}), \quad (3)$$

where

$$P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \frac{w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}, \quad (4)$$

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}), \quad (5)$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}), \quad (6)$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)}. \quad (7)$$

By minimizing the mean squre error, the mapping function $\mathcal{F}$ is derived as

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) \boldsymbol{E}_{m,t}^{(y)}. \quad (8)$$

When maximum likelihood estimation is adopted for parameter generation [11], the covariance matrix of the conditional probability density in Equation 7 is also considered and the target parameters are generated by the following updating equations:

$$\hat{\boldsymbol{y}}_t = \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)-1} \right)^{-1} \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)-1} \boldsymbol{E}_{m,t}^{(y)} \right),$$

$$\gamma_{m,t} = P(m|\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}). \quad (9)$$

3. Probabilistic integration of joint density model and speaker model

3.1. Outline

Fundamentally, VC is a technique that allows us to convert voice characteristics of the source speaker into those of the target speaker *without changing the linguistic content*. When VC systems are considered as a speech generator of the target speaker, the source speaker's utterance can be regarded as seeds of linguistic content for generation of speech. From this point of view, the joint density model in the conventional VC systems has two functions; to ensure the consistency of the linguistic information and to model the speaker individuality of the target. To realize the first function, a parallel corpus is absolutely necessary for training a proper model for the function. However, about the latter one, a non-parallel corpus is sufficient for only constructing the target speaker model. In this paper, we realize these two functions by different models; a joint density model for the first and a speaker model for the second. For voice conversion, these models are integrated by probabilistic formulation based on the conditional maximum likelihood criterion.

3.2. Formulation

First, as with the conventional VC approach, we focus on the conditional probability density of the target vector $\boldsymbol{y}_t$, given the source vector $\boldsymbol{x}_t$. From the conditional maximum likelihood criterion, the optimum output for the target vector is derived as follow:

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\operatorname{argmax}} P(\boldsymbol{y}_t|\boldsymbol{x}_t). \quad (10)$$

By using the Bayes rule, which is the same manner as that of automatic speech recognition or statistical machine translation, Equation 10 is written as

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\operatorname{argmax}} \underbrace{P(\boldsymbol{x}_t|\boldsymbol{y}_t)}_{\text{from joint density model}} \underbrace{P(\boldsymbol{y}_t)}_{\text{from speaker model}}. \quad (11)$$

In Equation 11, the first term $P(\boldsymbol{x}_t|\boldsymbol{y}_t)$ corresponds to the function that provides the consistency of the linguistic content between the source and the target speakers, because this "feedback" model is trained by a parallel corpus. The second term $P(\boldsymbol{y}_t)$ corresponds to the function that models the speaker individuality of the target. For the first term, we use the parameters of the joint density model trained by the parallel corpus. On the other hand, we can use the speaker GMM for the second term, which is widely used in speaker recognition studies. The speaker GMM is trained by a non-parallel corpus of the utterances of the target speaker.

Here, we derive an algorithm of voice conversion based on Equation 11. Let $\boldsymbol{\lambda}^{(z)}$ and $\boldsymbol{\lambda}^{(s)}$ be the parameters of the joint density model and those of the speaker model, respectively. We define a new likelihood function $\mathcal{L}$ based on Equation 11 as follows:

$$\mathcal{L}(\boldsymbol{y}_t; \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(s)}) \triangleq P(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{\lambda}^{(s)})^\alpha, \quad (12)$$

where the constant α denotes the weight for controlling the balance between the two models, as it is similar to a language model weight in speech recognition. To obtain the optimum solution $\hat{\boldsymbol{y}}_t$ to maximize the function $\mathcal{L}$, we derive the auxiliary function with respect to $\hat{\boldsymbol{y}}_t$. For the following derivation, similar to the conventional joint density GMM, n and N are

the mixture component index and the total number of mixture components in the speaker GMM, respectively.

$$\begin{aligned}
&\log \mathcal{L}(\hat{\boldsymbol{y}}_t) \\
&= \log \sum_{m=1}^{M} P(\boldsymbol{x}_t, m|\hat{\boldsymbol{y}}_t, \boldsymbol{\lambda}^{(z)}) + \alpha \log \sum_{n=1}^{N} P(\hat{\boldsymbol{y}}_t, n|\boldsymbol{\lambda}^{(s)}) \quad (13) \\
&\geq Q_{z_1}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) + Q_{z_2}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) + \alpha Q_s(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t). \quad (14)
\end{aligned}$$

Q_{z_1}, Q_{z_2}, and Q_s are auxiliary functions as follows:

$$Q_{z_1}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \sum_{m=1}^{M} \gamma_{m,t} \log P(m|\hat{\boldsymbol{y}}_t, \boldsymbol{\lambda}^{(z)}), \quad (15)$$

$$Q_{z_2}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \sum_{m=1}^{M} \gamma_{m,t} \log P(\boldsymbol{x}_t|m, \hat{\boldsymbol{y}}_t, \boldsymbol{\lambda}^{(z)}), \quad (16)$$

$$Q_s(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \sum_{n=1}^{N} \gamma_{n,t} \log P(\hat{\boldsymbol{y}}_t, n|\boldsymbol{\lambda}^{(s)}), \quad (17)$$

$$\gamma_{m,t} = P(m|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}), \gamma_{n,t} = P(n|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(s)}). \quad (18)$$

To derive Equation 14, we use Jensen's inequality. Since linguistic contents do not change, we assume that Equation 15 does not change drastically when $\hat{\boldsymbol{y}}_t$ changes, i.e. the derivative of Q_{z_1} with respect to $\hat{\boldsymbol{y}}_t$ can be ignored. Finally, we iteratively maximize the following function to optimize $\hat{\boldsymbol{y}}_t$:

$$Q'(\boldsymbol{y}, \hat{\boldsymbol{y}}_t) = Q_{z_2}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) + \alpha Q_s(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t). \quad (19)$$

By setting the derivative of Equation 19 with respect to $\hat{\boldsymbol{y}}_t$ to zero, the following updating equation is derived:

$$\begin{aligned}
\hat{\boldsymbol{y}}_t = &\left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(y)-1} + \alpha \sum_{n=1}^{N} \gamma_{n,t} \boldsymbol{\Sigma}_n^{-1} \right)^{-1} \times \\
&\left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(y)-1} \boldsymbol{E}_{m,t}^{'(y)} + \alpha \sum_{n=1}^{N} \gamma_{n,t} \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_n \right), \quad (20)
\end{aligned}$$

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

$$\boldsymbol{E}_{m,t}^{'(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yy)} \boldsymbol{\Sigma}_m^{(xy)+} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}), \quad (21)$$

$$\boldsymbol{D}_m^{'(y)-1} = \boldsymbol{D}_m^{(y)-1} - \boldsymbol{\Sigma}_m^{(yy)-1}. \quad (22)$$

The notation $(\cdot)^+$ denotes the pseudo-inverse of the matrix. For the initial values of the iteration, $\gamma_{m,t}$ and $\gamma_{n,t}$ are set as $P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)})$ and $P(n|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(s)})$, respectively. Equation 20 has a similar form to Equation 9, but it becomes the weighted summation of the effects from the joint density model and that from the speaker model. Thus, our proposed method can overcome the sparse parallel data problem by reducing the overestimation effects of the joint density parameters by the speaker model.

3.3. Relationship with other non-parallel approaches

Requirements of the parallel corpus for VC systems become a barrier to flexible applications of VC techniques. Then, there have been several proposed approaches which do not require a parallel corpus. Mouchtaris *et al.* proposed an unsupervised training method based on maximum likelihood constrained adaptation of the GMM trained with an existing parallel data set of a different speaker-pair [7]. Lee *et al.* proposed another approach based on maximum a posteriori (MAP) adaptation [8]. Compared with our proposed method, these approaches mainly focus on flexible control of the speaker individuality of the target. Then, the source speaker provides plenty of utterances for the joint density model. On the other hand, our method is expected to work well even if the number of training data for the joint density model is very small. Then, it eases the burden of the source speaker, i.e. the user of the VC application. In addition, since the proposed method can train the joint density model and the speaker model separately, it has the potential to apply precise modeling techniques proposed independently in VC and speaker recognition studies.

4. Experiment

4.1. Experimental conditions

To evaluate the performance of our proposed approach, voice conversion experiments using Japanese sentences were performed. In this experiment, we used speech samples from 5 speakers (MSH as the source speaker, MMY, MTK, FKS, and FTK as the target speakers) in the ATR Japanese speech database B-set [12]. This database consists of 503 phonetically balanced sentences. The first letters of the speaker names correspond to gender. We selected the last 53 sentences for test data. For training of the joint density models, one sentence pair was used. The total number of mixture components (M) was 4 or 16. On the other hand, for the speaker GMMs, 50 sentences were selected and the GMM for each speaker was trained. Note that the sentence used for training of the joint density model was not included in the 50 sentences for the speaker GMMs. The number of mixture components for the speaker GMM (N) was varied from 4 to 128. The weight for controlling the balance between both the models (α) was selected from 0.5, 1, and 5. The number of iterations for Equation 20 was fixed to 5.

We used 24-dimensional mel-cepstrum vectors for spectrum representation. These are derived by STRAIGHT analysis [13]. Aperiodic components, which are features to construct STRAIGHT mixed excitation, are not converted in this study, and they are fixed to -30 dB at all frequencies. Prosodic features, 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 proposed approach with the conventional VC technique based on only the joint density model.

4.2. Objective evaluations

We evaluated the conversion performance using mel-cepstral distortion between the converted vectors and the vectors of the targets. Figure 1 shows the result of average mel-cepstral distortion for the test data as a function of the number of mixture components of *speaker GMM*. The solid and dashed lines of "Conventional" are results of the joint density models where the number of utterance pairs are 1 and 32, respectively. The respective optimal numbers of mixture components for each of them are selected. Compared with "the conventional method (1 pair)", the proposed method was better. This is because the sparse data problem, i.e. "the conventional (1 pair)" could not train the model parameters sufficiently and sometimes caused the over-training effect. On the other hand, the proposed method improved the performance of "the conventional (1 pair)" even when the both methods used the only 1 parallel sentence. Thus, we show the effectiveness of the proposed method by mitigating the sparse data problem by using the speaker model, as discussed in Section 3.2. In this exper-

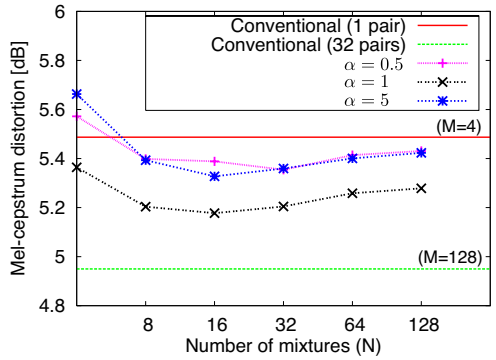

Figure 1: Results of objective evaluations as a function of N. M for the proposed method is 4.

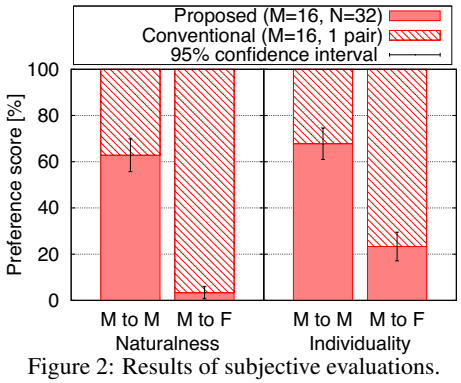

Figure 2: Results of subjective evaluations.

iment, the optimal weight for controlling the balance between the models became $\alpha = 1$. It might depend on the balance of training data between the two models.

4.3. Subjective evaluations

A listening test was carried out to evaluate the naturalness of converted speech and conversion accuracy for speaker individuality. The test was conducted with 15 subjects to compare the utterances converted by the proposed method and those by the conventional method. 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.

Figure 2 shows preference scores. In the case of male to male conversion, the proposed method outperformed the conventional one. In male to female conversion, the proposed method did not outperform the conventional method. This may be because the initialization of $\gamma_{m,t}$ and $\gamma_{n,t}$ in Equation 20 does not work very well in cases of cross-gender conversion. For a solution of this, derivation of initial parameters only from the convetional joint density model and iteration of Equation 20 might be effective. Speaker model based on the universal background model also might be another solution [10]. We observe that the preference scores for speaker individuality were better than those for naturalness in both the cases. This result reflects the properties of our method which uses the joint density model trained by a small parallel corpus and the well-trained speaker model.

Although some improvements for cross-gender conversion are required, experimental results demonstrate the effectiveness of our method in the case that a parallel corpus is small.

5. Conclusions

We have proposed a new method for voice conversion which integrates the speaker model and the joint density model into one function of VC using probabilistic formulation. This approach uses non-parallel data from the target speaker effectively and works well when the amount of parallel data is limited. Since the proposed method can train the joint density model and the speaker model separately, it has the potential to apply more precise modeling to both the joint density model and the speaker model. For further improvements of the conversion performance, we are planning to apply the prior knowledge to both the models and to integrate speech recognition models into our approach for more intelligible voice conversion.

6. Acknowledgment

The authors would like to thank Dr. T. Toda of NAIST, Japan, for fruitful discussion on VC approaches.

7. 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] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," Proc. ICASSP, pp. 301–304, 2001.

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

[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," Proc. ICASSP, pp. 655–658, 1988.

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

[6] 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.

[7] 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.

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

[9] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," Proc. INTERSPEECH, pp. 2446–2449, 2006.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, no. 1–3, pp. 19–41, 2000.

[11] 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.

[12] 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.

[13] 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.