Phoneme-based Spectral Voice Conversion Using Temporal Decomposition and Gaussian Mixture Model

Binh Phu Nguyen and Masato Akagi
School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, JAPAN
E-mail: {npbinh, akagi}@jaist.ac.jp

***Abstract*—In state-of-the-art voice conversion systems, GMM-based voice conversion methods are regarded as some of the best systems. However, the quality of converted speech is still far from natural. There are three main reasons for the degradation of the quality of converted speech: (i) modeling the distribution of acoustic features in voice conversion often uses unstable frames, which degrades the precision of GMM parameters (ii) the transformation function may generate discontinuous features if frames are processed independently (iii) over-smooth effect occurs in each converted frame. This paper presents a new spectral voice conversion method to deal with the two first drawbacks of standard spectral modification methods, insufficient precision of GMM parameters and insufficient smoothness of the converted spectra between frames. A speech analysis technique called temporal decomposition (TD), which decomposes speech into event targets and event functions, is used to effectively model the spectral evolution. For improvement of estimation of GMM parameters, we use phoneme-based features of event targets as spectral vectors in training procedure to take into account relations between spectral parameters in each phoneme, and to avoid using spectral parameters in transition parts. For enhancement of the continuity of speech spectra, we only need to convert event targets, instead of converting source features to target features frame by frame, and the smoothness of converted speech is ensured by the shape of the event functions. Experimental results show that our proposed spectral voice conversion method improves both the speech quality and the speaker individuality of converted speech.**

Keywords: spectral voice conversion, temporal decomposition, Gaussian mixture model (GMM)

I. INTRODUCTION

The aim of voice conversion is to convert a speaker voice (source speaker) to sound as if it were the voice of a defined speaker (target speaker). Applications of voice conversion systems can be found in several fields, such as Text-to-Speech customization, automatic translation, education, medical aids and entertainment, etc..

The core process in a voice conversion system is the transformation of the spectral envelope of the source speaker to match that of the target speaker. There are many approaches which have been proposed to implement the transform function for converting source features to target features, such as codebook-based conversion [1], neural network-based conversion [2], hidden Markov model (HMM)-based conversion [3], and Gaussian mixture model (GMM)-based conversion [4] [5] [6] [7] [8] [9] [10] [11] [12] [13]. Among those techniques, the vast majority of the current voice conversion systems focus on data-driven GMM-based transformation on the spectral aspects of conversion. Research results found in the literature have shown that the GMM-based approaches can be used successfully in voice conversion. These approaches are still regarded as robust and capable of producing high speech quality [10]. Although the GMM-based voice conversion methods can give reasonably acceptable speech, the quality of converted speech is still far from natural. Three major problems remain to be solved, i.e. insufficient precision of GMM parameters, insufficient smoothness of the converted spectra between frames, and over-smooth effect in each converted frame. This paper deals with the first two of the three drawbacks in a GMM-based voice conversion system, insufficient precision of GMM parameters, and insufficient smoothness of the converted spectra between frames.

A GMM-based voice conversion method normally includes two parts, a training procedure and a transformation procedure. In the training procedure, the methods are often based on parallel training data, where both the source and target speakers utter the same sentences. In this case, the dynamic time warping (DTW) algorithm is often used to align the two signals, to extract matching source and target training vectors. Both unstable frames, which often come from transition parts between phonemes, and stable frames are used to model the distribution of acoustic features. This leads to addition of noise to the GMM parameters. To overcome this drawback, some solutions have been proposed. In the work of Kumar and Verma [8], acoustic space of a speaker was partitioned explicitly into phones using the phonetic alignments and GMM was used for finer modeling of each phone. This approach could prevent the interference of frames between phones. However, it still used unstable frames in each phone. Liu et al. [11] segmented frames according to each phoneme, and eliminated unstable frames in each phoneme by proposing a method for identifying stable frames based on limitation of maximal variation range for the first three formant frequencies. After getting the stable frames, Liu et al. also used GMM to model the distribution of acoustic features. Nguyen and Akagi [13] used event targets as spectral vectors to estimate

1-4244-2426-9/08/$20.00 ©2008 IEEE

GMM parameters, instead of using spectral parameters of aligned frames. However, all methods in [8], [11], and [13] did take into account the relations between frames when estimating GMM parameters. GMM parameters therefore are more precisely estimated when being considered the relations between frames.

In the transformation procedure, there are two main drawbacks, i.e. insufficient smoothness of the converted spectra between frames, and over-smooth effect in each converted frame. Until now, most voice conversion methods perform voice transformation function frame by frame. This means that to convert one frame, the information about past and future frames is not relevant. This may cause a discontinuity problem between adjacent frames when unexpected modifications happen in some frames. As a result, there are some clicks in the converted speech. Moreover, Knagenhjelm and Kleijn [14] pointed out that spectral discontinuities between adjacent frames were one of the major sources of quality degradation in speech coding systems. Some approaches to deal with this problem were discussed. In the work of Chen et al. [7], to maintain a continuous transformation in consecutive frames, the converted features were smoothed along the time axis by employing a median filter and a low pass filter. However, applying these filters could lead to a loss of temporal resolution, and it was a relatively crude implementation. Duxans et al. [9] included dynamic information in their GMM-based voice conversion system to take into the relations between frames. However, according to Duxans et al. [9], this method did not improve the performance of a GMM-based voice conversion system. Therefore, the discontinuity problem between adjacent frames should be solved to enhance the quality of converted speech. The problem of over-smooth effect happens in each converted frame, because of the statistical averaging operation [6]. Some works attempted to solve it [6] [7] [10], but defining solutions for this problem is beyond the scope of this paper.

This paper addresses two of the three main issues mentioned above, insufficient precision of GMM parameters, and insufficient smoothness of the converted spectra between frames. We propose a new spectral voice conversion method based on temporal decomposition (TD) [15] [16] and GMM [4] [5]. In our proposed method, we employ the modified restricted temporal decomposition (MRTD) algorithm [16] in both training and transformation procedures. We extract a set of phoneme-based features of event targets. We then use them as spectral vectors for training to take into the relations between spectral parameters in each phoneme, and to avoid using spectral parameters in transition parts. In the transformation procedure, we only need to convert event targets, instead of converting spectral parameters frame by frame, and the smoothness of converted speech is ensured by the shape of the event functions. In addition, since the fundamental frequency and vocal tract information are not independent, modifying them separately will often degrade the quality of converted speech. Therefore, a high quality analysis-synthesis framework, STRAIGHT [17] is utilized in this paper.

II. Conventional GMM-based voice conversion

As previously mentioned, the GMM-based voice conversion methods are found to be superior to other methods. In this section, we describe the basic GMM-based voice conversion method which is employed as our baseline system. A GMM-based voice conversion method often includes two parts, the training procedure and the transformation procedure.

A. Training Procedure

The source speech is represented by a time sequence $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$, where $\mathbf{x}_i$ is a D dimensional feature vector for the i^{th} frame, i.e. $\mathbf{x}_i = [x_1, x_2, ..., x_D]^T$. The target speech is represented by a time sequence $Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m]$, where $\mathbf{y}_j = [y_1, y_2, ..., y_D]^T$. The DTW algorithm is then adopted to align source features with their counterparts in target series to obtain feature pair series $Z = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_q]$ where $\mathbf{z}_q = [\mathbf{x}_i^T, \mathbf{y}_j^T]^T$.

The distribution of Z is modeled by Gaussian mixture model, as in Eq. (1).

$$p(z) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(z; \mu_m, \Sigma_m) = p(x, y) \tag{1}$$

where M is the number of Gaussian components. $\mathcal{N}(z; \mu_m, \Sigma_m)$ denotes the 2D dimension normal distribution with the mean μ_m and the covariance matrix Σ_m. α_m is the prior probability of $\mathbf{z}$ having been generated by component m, and it satisfies $0 \leq \alpha_m \leq 1$, $\sum_{m=1}^{M} \alpha_m = 1$. The parameters $(\alpha_m, \mu_m, \Sigma_m)$ for the joint density $p(x, y)$ can be estimated using the expectation maximization (EM) algorithm [18].

B. Transformation Procedure

The transformation function that converts source feature $\mathbf{x}$ to target feature $\mathbf{y}$ is given by Eq. (2).

$$F(x) = E(y|x) = \int y p(y|x) dy$$

$$= \sum_{m=1}^{M} p_m(x) \left(\mu_m^y + \Sigma_m^{yx} (\Sigma_m^{xx})^{-1} (x - \mu_m^x) \right) \tag{2}$$

$$p_m(x) = \frac{\alpha_m \mathcal{N}(x; \mu_m^x, \Sigma_m^{xx})}{\sum_{m=1}^{M} \alpha_m \mathcal{N}(x; \mu_m^x, \Sigma_m^{xx})} \tag{3}$$

where $\mu_m = \begin{bmatrix} \mu_m^x \\ \mu_m^y \end{bmatrix}$, $\Sigma_m = \begin{bmatrix} \Sigma_m^{xx} & \Sigma_m^{xy} \\ \Sigma_m^{yx} & \Sigma_m^{yy} \end{bmatrix}$, and $p_m(x)$ is the probability of $\mathbf{x}$ belonging to the m^{th} Gaussian component.

III. Temporal Decomposition

A shortcoming of the conventional GMM-based voice conversion methods is that they do not take into account the correlation between frames in both training and transformation procedures. As a result, the precision of estimated GMM parameters is degraded, and there are some clicks in the converted speech because of discontinuous spectral contours. Therefore, we employ TD to deal with the problem.

In articulatory phonetics, speech is described as a sequence of distinct articulatory gestures, each of which produces an acoustic event that should approximate a phonetic target. Due to the overlap of the gestures, these phonetic targets are often only partly realized.

Atal [15] proposed a method based on the temporal decomposition of speech into a sequence of overlapping target functions and corresponding event targets, as given in Eq. (4).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \qquad 1 \le n \le N \tag{4}$$

where $\mathbf{a}_k$ is the speech parameter corresponding to the k^{th} event target. The temporal evolution of this target is described by the k^{th} event function, $\phi_k(n)$. $\hat{\mathbf{y}}(n)$ is the approximation of the n^{th} spectral parameter vector $\mathbf{y}(n)$, and is produced by the TD model. *N* and *K* are the number of frames in the speech segment, and the number of event functions, respectively ($N \gg K$).

Many applications of TD have been explored in the literature, such as speech coding [16], and speaker identification [19]. In this paper, we investigate the application of TD in speech modification. To modify the speech spectra, we only need to modify the speech spectra of event targets and the corresponding event functions, instead of modifying the speech spectra frame by frame. The smoothness of modified speech will be ensured by the shape of the event functions. This leads to easy modification of the speech spectra, as well as ensuring the smoothness of the speech spectra between frames, and thereby enhances the quality of modified speech.

A. Modified Restricted Temporal Decomposition (MRTD)

The original method of TD is known to have two major drawbacks, high computational costs, and high parameter sensitivity to the number and locations of events. A number of modifications have been explored to overcome these drawbacks. In this study, we employ the MRTD algorithm [16]. The reasons for using the MRTD algorithm in this work are twofold: (i) the MRTD algorithm enforces a new property on event functions, named the "well-shapedness" property, to model the temporal structure of speech more effectively [16]; (ii) event targets can convey the speaker's identity [19]. In the MRTD algorithm, LSF parameters are chosen for the input of TD because of their sensitivity (an adverse alteration of one coefficient results in a spectral change only around that frequency) and efficiency (LSFs result in low spectral distortion when being interpolated and/or quantized). In this paper, LSF parameters are extracted from spectral envelopes of STRAIGHT [17]. The STRAIGHT spectra are suitable for TD, because they are smooth in the time-frequency domain.

B. Phoneme-based Determination of Event Locations

The MRTD algorithm uses a spectral stability criterion to determine the initial event locations [16]. It is assumed that each acoustic event that exists in speech gives rise to a spectrally stable point in its neighborhood. Therefore, the locations of the spectrally stable points and the corresponding spectral parameter sets can be used as good approximations of event locations and event targets, respectively. This algorithm is automatically performed, and the subsequent computation of refined event targets and event functions is much less demanding than the traditional TD method. This algorithm is useful for applications in speech coding [16], and speaker identification [19]. However, this algorithm does not ensure a one-to-one correspondence between event locations and phonemic units. This makes it difficult to align parallel training data in voice conversion systems.

Shibata and Akagi [20] proposed a new method for determination of event locations based on phonemes. To increase the accuracy of phoneme segmentation, this algorithm is effectively used when labeled data of utterances are available. Each phoneme is divided into four equal segments, and the five points marking these segments are used for identifying the event locations. Using this algorithm, the quality of synthesized speech is very high. Specially, since we can represent each phoneme by five event targets, these five event targets of each phoneme can be regarded as a "voice font". It should be noted that we can easily increase the quality of synthesized speech by increasing the number of event locations in each phoneme.

IV. Phoneme-based Spectral Voice Conversion Using Temporal Decomposition and Gaussian Mixture Model

A. Spectral Parameters

The overall shape of the spectral envelope provides an effective representation of the vocal tract characteristics of the speaker. However, the dimension of the spectral envelope is rather high, and it is not effective for direct use in a voice conversion system. We therefore often use another representation of the spectral envelope. MFCC coefficients are used to represent the spectral envelope in [4] [6] [8], while line spectral frequency (LSF) coefficients are used in [5] [7] [10] [9] [12] for the reason that LSFs have better linear interpolation attributes. In our voice conversion system, we choose LSFs for the representation of the spectral envelope. The reason for selecting LSFs is that these parameters closely relate to formant frequencies, but in contrast to formant frequencies they can be estimated quite reliably. Also, they have good interpolation characteristics, and a badly predicted component adversely affects only a portion of the frequency spectrum. Moreover, they are easily integrated with the MRTD algorithm, which uses LSFs as its input.

B. Proposed Spectral Voice Conversion Method

As previously mentioned, our proposed method focuses on spectral voice conversion, and is based on the GMM method [4] [5]. The processing flow of our spectral voice conversion system, which includes training and transformation procedures, is described as follows, and is shown in Fig. 1.

In the training procedure, STRAIGHT [17] decomposes input speech signals into spectral envelopes, F0 (fundamental

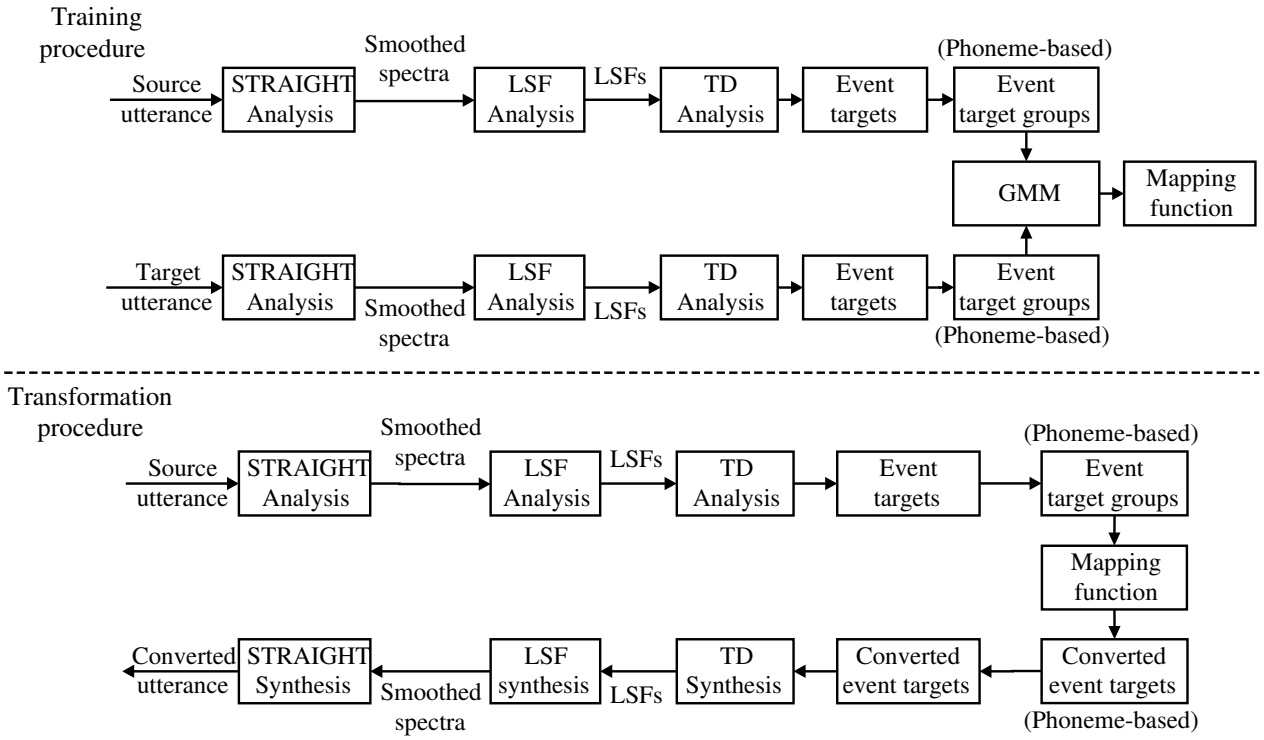

Fig. 1. Diagram of our proposed voice conversion method training procedure (top), and transformation procedure (bottom).

frequency) information, and aperiodic components (AP). Since the spectral envelopes can be further analyzed into LSF parameters, MRTD [16] is employed in the next step to decompose the LSF parameters into event targets and event functions. Note that the method for determination of event locations is from [20]. Each phoneme is represented by five event targets, and a vector of phoneme-based features of event targets $\mathbf{EV} = [\mathbf{a}_1^T, \mathbf{a}_2^T, \mathbf{a}_3^T, \mathbf{a}_4^T, \mathbf{a}_5^T]$, where $\mathbf{a}_k (1 \leq k \leq 5)$ is the k^{th} event target in each speech segment (a phoneme), can be a good vector to present the relations between event targets in a phoneme. Moreover, each event target $\mathbf{a}_k$ in the MRTD algorithm [16] is a valid LSF coefficient. An important property of LSFs $\{LSF_i\}$ is that they are ordered $(0, \pi)$, as follows.

$$0 < LSF_1 < LSF_2 < \ldots < LSF_P < \pi \tag{5}$$

where P is the order of LSF. To prevent a bad initialization in estimation of GMM parameters, we normalize the vectors of phoneme-based features of event targets extracted from each phoneme in utterances of source and target speakers, $\mathbf{EV_x}$ and $\mathbf{EV_y}$, as follows.

$$\mathbf{EV_x} = [\mathbf{a}_{x1}^T, \mathbf{a}_{x2}^T + \pi, \mathbf{a}_{x3}^T + 2\pi, \mathbf{a}_{x4}^T + 3\pi, \mathbf{a}_{x5}^T + 4\pi]^T \tag{6}$$

$$\mathbf{EV_y} = [\mathbf{a}_{y1}^T, \mathbf{a}_{y2}^T + \pi, \mathbf{a}_{y3}^T + 2\pi, \mathbf{a}_{y4}^T + 3\pi, \mathbf{a}_{y5}^T + 4\pi]^T \tag{7}$$

where $\mathbf{a}_{xk}$, $\mathbf{a}_{yk}$ are the k^{th} event targets in each phoneme of the source and target speakers, respectively. As a result, the vectors $\mathbf{EV_x}$ and $\mathbf{EV_y}$ are ordered $(0, 5\pi)$. All the phoneme-based features are then aligned according to each phoneme, and modeled by GMM parameters in Eq. (1).

In the transformation procedure, normalized phoneme-based features are also extracted from each utterance of the source speaker by using STRAIGHT and MRTD. We then convert each of the normalized phoneme-based features by using Eq. (2), and convert back to event targets. The converted event targets are re-synthesized as converted LSF by MRTD synthesis. In the following step, the converted LSF parameters are synthesized as spectral envelopes by LSF synthesis. Finally, STRAIGHT synthesis is employed to output the converted speech. Note that this paper does not deal with prosodic, energy conversion. Therefore, to implement a complete voice conversion system, our proposed method should be integrated with some methods for prosodic, energy conversion, such as in [6] [12].

V. Experiments and Results

A. Experimental Conditions

The corpus used for the experiments is a dataset consisting of 460 sentences spoken once each by two speakers (one male & one female) in the MOCHA-TIMIT English speech database [21]. The speech data was recorded at 16KHz sampling rate. In our experiments, two different voice conversion tasks were investigated: male-to-female, and female-to-male conversion.

TABLE I
ANALYSIS CONDITIONS FOR EXPERIMENTS ON THE VOICE CONVERSION METHODS.

Sampling frequency	16 kHz
Window length	40 ms
Window shift	1 ms
FFT points	1024
LSF order	18
Gaussian components	20

For each kind of conversion, we used 300 pair utterances for training, and 30 other pair utterances for evaluation.

To evaluate the performance of our proposed method, we performed an objective test, and also subjective evaluation experiments regarding speech quality and speaker individuality. We compared our proposed method (the phoneme-based TD+GMM method) with two other methods. The first method used for comparison is the conventional method (the GMM method) [4] [5]. The second method used for comparison also used event targets for training, and the transformation procedure was performed for each event target (the TD+GMM method). The difference between the second method and our proposed method is that the second method does not take into account the relations between event targets in training and transformation procedures. Since we only focus on spectral voice conversion, we automatically copy the prosody information and energy from the utterances of the target speaker to converted utterances. In addition, because the problem of the over-smooth effect in each converted frame is outside the scope of this paper, without loss of generality, all three methods utilize the same transformation mapping function of the conventional method [4] [5] (see Eq. (2)). The analysis conditions for these experiments are shown in Table I.

B. Objective Test

We use LSF performance index PI_{LSF} for the objective test. This measure is defined as follows.

$$PI_{LSF} = 1 - \frac{E_{LSF}(t(n), \widehat{t}(n))}{E_{LSF}(t(n), s(n))} \tag{8}$$

where $t(n)$ represents the utterance of the target speaker, $s(n)$ represents the utterance of the source speaker, and $\widehat{t}(n)$ represents the converted utterance. $E_{LSF}(t(n), \widehat{t}(n))$ is the mean transform LSF error, and $E_{LSF}(t(n), s(n))$ is the mean inter-speaker LSF error, defined as follows.

$$E_{LSF}(A, B) = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{P}\sum_{i=1}^{P}\left(LSF_A^{l,i} - LSF_B^{l,i}\right)^2} \tag{9}$$

where L is the number of frames, P is the order of LSF, and $LSF^{l,i}$ is the LSF component i in the frame l.

PI_{LSF} = 0 indicates that the output of the system is no more similar to the target than the source is, whereas PI_{LSF} = 1 indicates that the output of the system is identical to the target. In general, a higher value for PI_{LSF} suggests a better system.

TABLE II
OBJECTIVE RESULTS FOR THE VOICE CONVERSION METHODS (1) CONVENTIONAL METHOD (GMM METHOD) (2) TD+GMM METHOD (3) OUR PROPOSED METHOD (PHONEME-BASED TD+GMM METHOD).

Type of conversion	LSF performance index		
	(1)	(2)	(3)
Male to Female	0.3692	0.3819	0.4013
Female to Male	0.3517	0.3745	0.3829

TABLE III
MOS RESULTS FOR VOICE CONVERSION METHODS (1) CONVENTIONAL METHOD (GMM METHOD) (2) TD+GMM METHOD (3) OUR PROPOSED METHOD (PHONEME-BASED TD+GMM METHOD).

Type of conversion	Mean opinion score		
	(1)	(2)	(3)
Male to Female	3.17	3.50	3.89
Female to Male	2.67	3.13	3.67

The results of this objective test are shown in Table II. These results indicate that the performance of our proposed method is significantly better than that of the conventional method, and also better than that of the second method (the TD+GMM method).

C. Subjective Tests

Subjective tests concerning speech quality and speaker individuality were carried out. Six graduate students known to have normal hearing ability were recruited for the listening experiments.

In the test of speech quality, we randomly presented each of ten converted utterances from both kinds of conversion (male-to-female and female-to-male) to listeners, and asked them to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). Table III shows the average scores, which indicate that the speech quality of our proposed method (the phoneme-based TD+GMM method) is superior to that of the conventional method (the GMM method), and also better than that of the second method (the TD+GMM method).

In the test of speaker individuality, an ABX test was conducted. A represents the source speaker, B represents the target speaker, and X represents the converted speech, which supplied from each one of the three test systems. The listeners were asked to select if X was closer to A or B, and adjusted the score from 1 to 5 according to his/her perception of speaker individuality when comparing. The score of 1 means that the converted speech is very similar to the source speaker, and the score of 5 means that the converted speech is very similar to the target speaker. Results of the ABX test are shown in Table IV. These results also indicate that the speech

TABLE IV
ABX RESULTS FOR VOICE CONVERSION METHODS (1) CONVENTIONAL METHOD (GMM METHOD) (2) TD+GMM METHOD (3) OUR PROPOSED METHOD (PHONEME-BASED TD+GMM METHOD).

Type of conversion	ABX score		
	(1)	(2)	(3)
Male to Female	4.06	4.17	4.50
Female to Male	3.44	3.56	4.00

individuality of converted utterances of our proposed method is the most similar to the target speaker among the three methods. It should be noted that the score of the test of speaker individuality is rather high because in this paper, we only focus on spectral conversion, and we therefore copied prosodic information and energy from utterances of the target speaker for all three methods.

VI. CONCLUSIONS

In this paper, we proposed a new spectral voice conversion method to deal with two of three main drawbacks of standard voice conversion techniques, insufficient precision of GMM parameters, and insufficient smoothness of the converted spectra between frames. Our proposed method considers the relations between frames when estimating GMM, by using a set of phoneme-based features of event targets as spectral vectors for training. Therefore, our approach can improve the precision of GMM parameters. Our proposed method also ensures the smoothness of converted speech by performing the conversion procedure for event targets, instead of converting the spectral parameters frame by frame. The experimental results prove the effectiveness of our proposed method.

There are however issues which still remain to be solved. Although prosodic conversion, and duration conversion are outside of the scope of this paper, they are important features for the realization of speaker personality, and to improve the natural quality of converted speech. Prosodic conversion, duration conversion, and the problem of over-smooth effect in each converted frame will be considered in our future work.

ACKNOWLEDGMENTS

This study was supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan. Some coding procedures for the conventional GMM-based voice conversion method are from [22]. We would like to thank Prof. Mary Ann Mooradian for checking our English.

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP*, pp. 655–658, 1998.

[2] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on radial basis function networks," *Proc. ICSLP*, pp. 285–288, 2002.

[3] E. K. Kim, S. Lee, and Y. H. Oh, "Hidden Markov model based voice conversion using dynamic characteristics of speaker," *Proc. Eurospeech*, pp. 2519–2522, 1997.

[4] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Proc. IEEE Trans. Speech Audio*, vol. 6, pp. 131–142, 1998.

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

[6] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," *Proc. ICASSP*, pp. 841–844, 2001.

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

[8] A. Kumar and A. Verma, "Using phone and diphone based acoustic models for voice conversion: A step towards creating voice fonts," *Proc. ICASSP*, pp. 720–723, 2003.

[9] H. Duxans, A. Bonafonte, A. Kain, and J. van Santen, "Including dynamic and phonetic information in voice conversion systems," *Proc. ICSLP*, pp. 1193–1196, 2004.

[10] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Trans. on Audio, Speech and lang. Proc.*, pp. 1301–1312, 2006.

[11] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through combining modified GMM and formant mapping for Mandarin," *Proc. ICDT*, p. 10, 2007.

[12] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," *Proc. Interspeech*, pp. 1965–1968, 2007.

[13] B. P. Nguyen and M. Akagi, "Control of spectral dynamics using temporal decomposition in voice conversion and concatenative speech synthesis," *Proc. NCSP*, pp. 279–282, 2008.

[14] H. P. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," *Proc. ICASSP*, pp. 732–735, 1995.

[15] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," *Proc. ICASSP*, pp. 81–84, 1983.

[16] P. C. Nguyen, T. Ochi, and M. Akagi, "Modified restricted temporal decomposition and its application to low bit rate speech coding," *IEICE Transactions on Information and Systems*, vol. E86-D, pp. 397–405, 2003.

[17] 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," *Journal of Speech Communication*, vol. 27, pp. 187–207, 1999.

[18] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, pp. 1–38, 1977.

[19] P. C. Nguyen, M. Akagi, and T. B. Ho, "Temporal decomposition: A promising approach to VQ-based speaker identification," *Proc. ICASSP*, pp. 184–187, 2003.

[20] T. Shibata and M. Akagi, "A study on voice conversion method for synthesizing stimuli to perform gender perception experiments of speech," *Proc. NCSP*, pp. 180–183, 2008.

[21] A. Wrench, "The MOCHA-TIMIT articulatory database," *Queen Margaret University College, http://www.cstr.ed.ac.uk/artic/mocha.html*, 1999.

[22] D. Suendermann, "Voice conversion Matlab toolbox," *Technical Report, Siemens Corporate Technology, Munich, Germany*, 2007.