Acoust. Sci. & Tech. **30**, 3 (2009)

PAPER

A flexible spectral modification method based on 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, 923–1292 Japan

(*Received 5 February 2008, Accepted for publication 22 September 2008*)

Abstract: Manipulating spectral structure often leads to degradation of speech quality, which is mainly due to insufficient smoothness of the modified spectra between frames, and ineffective spectral modification. This paper presents a new spectral modification method to improve the quality of modified speech. If frames are processed independently, discontinuous features may be generated. Therefore, a speech analysis technique called temporal decomposition (TD), which decomposes speech into event targets and event functions, is used to model the spectral evolution effectively. Instead of modifying the speech spectra frame by frame, we only need to modify event targets and event functions. This feature leads to easy modification of the speech spectra, and the smoothness of modified speech is ensured by the shape of event functions. To improve spectral modification, we explore Gaussian mixture model parameters (spectral-GMM parameters) to model the spectral envelope of each event target, and develop a new algorithm for modifying spectral-GMM parameters in accordance with formant scaling factors. We first evaluate the effectiveness of our proposed method in spectra modeling, and then apply it to two areas which require different amounts of spectral modification, emotional speech synthesis and voice gender conversion. Experimental results show that the effectiveness of our proposed method is verified for spectra modeling and spectral modification.

Keywords: Spectral modification, Temporal decomposition, Gaussian mixture model, STRAIGHT

PACS number: 43.72.Ar, 43.72.Ja, 43.60.Ek [doi:10.1250/ast.30.170]

1. INTRODUCTION

Spectral modification techniques are used to perform a variety of modifications to speech spectra, such as manipulations of the formant structures, amplitude manipulations, and so on. Since spectral processing is closely linked to human perception, it is an effective way to perform sound processing. It can be applied in many areas. Spectral modification methods are a powerful technology for customizing Text-to-Speech (TTS) systems, such as by converting source features to target features [1,2], changing a male voice into a female voice and vice versa [3], and applying to emotional speech synthesis [4]. Spectral modification techniques are often applicable to automatic speech recognition tasks [5], and speech enhancement [6].

The basic idea of spectral processing is to convert a time-domain digital signal into its frequency-domain representation. Most of the approaches start by developing an analysis/synthesis technique from which the speech signal is reconstructed with minimum loss of sound quality. Then, the main issues have to be resolved: what kind of representation and which parameters are chosen for the application of the desired speech processing. The challenge of spectral modification is to modify the spectral/acoustical features without degrading the speech quality.

A variety of spectral modification methods have been discussed in the literature. They can be classified into two popular approaches: linear prediction (LP)-based methods [7,8] and frequency warping methods [9]. LP-based methods are often affected by the pole interaction problem suffered by pole modification techniques. An iterative algorithm for overcoming pole interaction during formant modification was developed by Mizuno *et al.* [7] This method produces spectral envelopes with desired formant amplitudes at the formant frequencies. However, the amplitude and bandwidth of each formant cannot be independently modified, since each formant's bandwidth is dependent on the magnitude of the corresponding pole. Recently, a method for directly modifying formant locations and bandwidths in the line spectral frequency (LSF) domain has been developed [8]. We refer to the method in

*e-mail: npbinh@jaist.ac.jp
†e-mail: akagi@jaist.ac.jp

[8] as the LSF-based method. By taking advantage of the nearly linear relationship between the LSF coefficients and formants, modifications are performed based on desired shifts in formant frequencies and bandwidths. However, the main drawback, i.e. the lack of control of the spectral shape, has not been solved. Frequency warping methods, such as by Turajlic *et al.* [9], give high quality of modified speech. However, frequency warping methods meet difficulties in modifying spectral peaks, such as preserving shapes of peaks, and emphasizing spectral peaks around 3 kHz in transformation of speaking voice into singing voice, since they do not estimate spectral peaks. Moreover, frequency warping methods do not allow formants to merge or split, which is often desired in formant modification processes [10].

In addition, some methods mentioned above [7,8] only mention how to modify the speech spectrum in a frame, and they [7–9] rarely deal with constraints between frames after modification. This limitation may cause a discontinuity problem between adjacent frames. As a result, there are some clicks in the modified speech when unexpected modifications happen in some frames. Moreover, Knagenhjelm and Kleijn [11] point out that spectral discontinuities between adjacent frames are one of the major sources of quality degradation in speech coding systems. Therefore, this problem should be solved to enhance the quality of modified speech.

This paper proposes a new spectral modification method to address two issues, insufficient smoothness of modified spectra between frames and ineffective spectral modification. First, we propose a new modeling of speech spectra for spectral modification based on temporal decomposition (TD) [12,13] and Gaussian mixture model (GMM) [14–16]. To model the spectral evolution, we employ the modified restricted temporal decomposition (MRTD) algorithm [13]. To model the speech spectrum, we use GMM parameters [14–16]. In this paper, GMM parameters [14–16] are called spectral-GMM parameters. Second, we develop a new algorithm for modifying spectral-GMM parameters in accordance with formant scaling factors. Note that the spectral-GMM parameters used here are to approximate a spectral envelope, which are different from those often used to model the distribution of acoustic features in state-of-the-art methods for voice conversion. We first evaluate the effectiveness of our proposed method in spectra modeling. We then evaluate the effectiveness of our proposed method in two areas, emotional speech synthesis, which requires modification of both formant frequency and power, and voice gender conversion, which requires a large amount of spectral modification. A part of this paper was presented at Interspeech'07 [17], and was published in the Journal of Signal Processing [18] as a short paper. This paper introduces more details of the concept and the algorithms in [17,18], and conducts more evaluations of spectra modeling and the two applications, emotional speech synthesis and voice gender conversion.

2. MODELING OF SPEECH SPECTRA

2.1. Temporal Decomposition

A shortcoming of conventional spectral modification methods is that they do not take into account the correlation between frames after modification. There are some clicks in the modified speech because of discontinuous spectral contours. Therefore, we employ TD to deal with the problem.

In articulatory phonetics, speech can be described as a sequence of distinct articulatory gestures. Each gesture produces an acoustic event that should approximate a phonetic target. Adjacent gestures overlap in time, which results in overlap of these phonetic targets.

Atal proposed a method based on the temporal decomposition of speech into a sequence of overlapping target functions and corresponding event targets [12], as given in Eq. (1).

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

where $\boldsymbol{a}_k$ is the spectral parameter vector corresponding to the kth event target. The temporal evolution of this target is described by the kth event function, $\phi_k(n)$. $\hat{\boldsymbol{y}}(n)$ is the approximation of the nth spectral parameter vector $\boldsymbol{y}(n)$, 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 ($N \gg K$).

To modify the speech spectra, we only need to modify the event targets $\boldsymbol{a}_k$ and the corresponding event functions $\phi_k(n)$, instead of modifying the speech spectra frame by frame. The smoothness of modified speech will be ensured by the shape of the event functions $\phi_k(n)$. This feature 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.

The original method of TD is known to have two major drawbacks, high computational cost 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 [13]. 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 [13]; (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 spectral

sensitivity (an adverse alteration of one coefficient results in a spectral change only around that frequency [20]) and their stability and interpolation advantages (LSFs result in low spectral distortion when being interpolated and/or quantized [21]). In this paper, LSF parameters are extracted from spectral envelope information of STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [22]. The STRAIGHT spectra are suitable for TD, because they are smooth in the time-frequency domain. More details of the extraction of LSF parameters from STRAIGHT can be found in [13]. To determine the initial event locations, the MRTD algorithm uses a spectral stability criterion based on LSF parameters. 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.

2.2. Speech Spectrum Modeling Using Gaussian Mixture Model (GMM)

One of the most important requirements of spectral modification is that it be flexible enough to perform a variety of modifications within the spectral envelope. Formant frequency is one of the most important parameters in characterizing speech, and it also plays an important role in specifying speaker characteristics. Therefore, using formant frequency as a parameter can control other features that are directly connected to the speech production process. Conventional spectral modification methods, such as [7–9], often control formants to modify the speech spectrum. However, these methods are limited by their inability to independently control important formant characteristics such as amplitude and bandwidth, or to control the spectral shape.

Zolfaghari *et al.* proposed a technique to fit a Gaussian mixture model to a smoothed magnitude spectrum of a speech signal [14–16]. This technique is briefly described as follows.

In a single frame, the normalized spectrum $X(e^{j\omega_n})$ is viewed as a probability distribution $P(X)$, where $X = \{x_1, \ldots, x_L\}$, x_l $(1 \le l \le L)$ is the frequency bin number, and $2L$ is the FFT size. $P(x_l)$ is simply a spectral density. The overall density of a Gaussian mixture model is written as follows.

$$u(x) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(x; \mu_m, \sigma_m^2) \tag{2}$$

where M is the number of mixture components, $\mathcal{N}(x; \mu_m, \sigma_m^2) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{(x-\mu_m)^2}{2\sigma_m^2}}$ is the mth local Gaussian component, μ_m, σ_m are called mean and standard deviation of Gaussian component m respectively, and $\{\alpha_m\}_{m=1}^{M}$ are mixture weights satisfying $0 \le \alpha_m \le 1$ and $\sum_{m=1}^{M} \alpha_m = 1$.

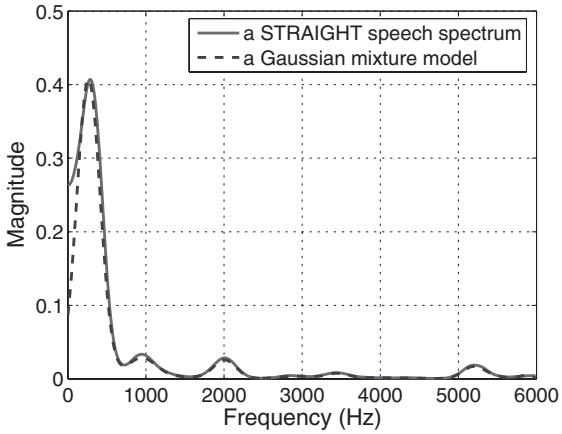

Fig. 1 A Gaussian mixture model ($M = 8$) fits to a STRAIGHT speech spectrum.

Zolfaghari *et al.* assumed that formants could be represented by Gaussian distributions, and a speech spectrum could be represented by a Gaussian mixture model. The EM algorithm [23] is often used to optimize the log likelihood of the histogram of the speech spectrum at time t with respect to the model parameters $u(x)$ in Eq. (2). The estimated means, standard deviations, and mixture weights of the Gaussian components can be related to the locations, bandwidths, and amplitudes of the formants, respectively [14]. The ability to independently control the parameters of each Gaussian component enables precise estimation of the spectral envelope, enables a wide variety of modifications, and enables independent control of the formants. Figure 1 shows that a Gaussian mixture model of eight Gaussian components can fit to a STRAIGHT speech spectrum (at 12 kHz sampling frequency) well.

2.3. Proposed Modeling of Speech Spectra for Spectral Modification

One of the advantages of TD is that it ensures the smoothness of modified speech signals. However, if event targets are represented by linear predictive coding (LPC) parameters, such as LSF parameters, we meet difficulties in spectral modification. To overcome these drawbacks, we use spectral-GMM parameters [14–16] to model each event target. Using TD and GMM, we propose a new modeling of speech spectra for spectral modification which can deal with these two drawbacks of conventional spectral modification methods, the insufficient smoothness of the modified spectra between frames, and the ineffective spectral modification. In addition, since glottal and vocal tract information are not independent, modifying them separately will often degrade the quality of modified speech signals. Therefore, a high-quality analysis-synthesis

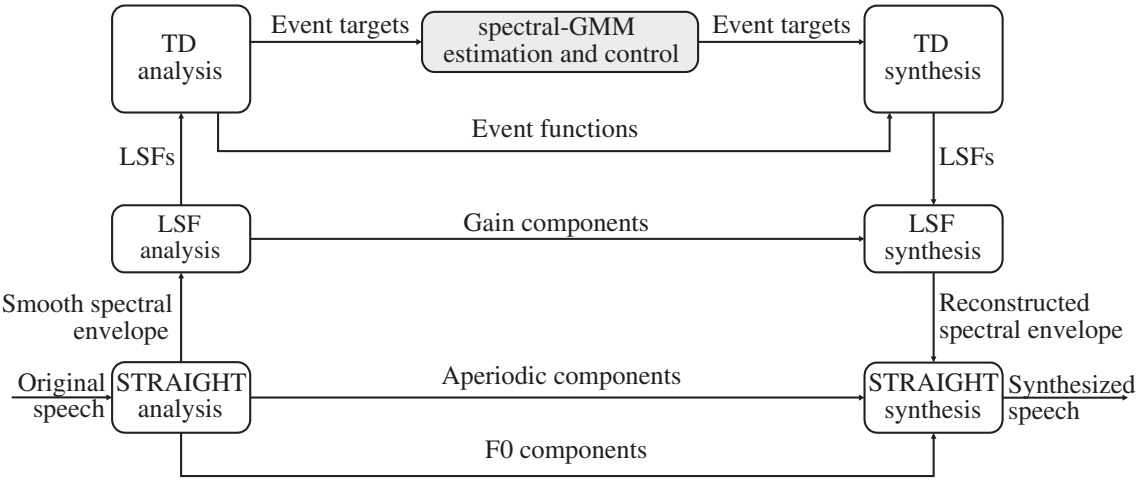

Fig. 2 Diagram of proposed modeling of speech spectra based on temporal decomposition and Gaussian mixture model.

framework, STRAIGHT [22] is utilized in this study. The processing flow of our proposed modeling of speech spectra is as follows, and is shown in Fig. 2.

First, STRAIGHT decomposes input speech signals into spectral envelope information, F0 (fundamental frequency) information, and aperiodic components. Since the spectral envelope information can be further analyzed into LSF parameters, MRTD is employed in the next step to decompose the LSF parameters into event targets and event functions. Since the event targets are valid LSF parameters [13], the spectral envelopes of event targets can be restored, and then the spectral envelopes are converted to spectral-GMM parameters. By using spectral-GMM parameters to model the event targets, we can flexibly perform some modifications of the event targets. The modified event targets are then re-synthesized as modified LSF by TD synthesis. In the following step, the modified LSF parameters are synthesized as spectral envelopes by LSF synthesis. Finally, STRAIGHT synthesis is employed to output the synthesized speech. Note that this paper proposes a spectral modification method, and applying the MRTD algorithm to conversion of other components (i.e. aperiodic, gain, and F0 components) will be our future work.

3. NEW SPECTRAL MODIFICATION ALGORITHM

Control of formants is an effective way to perform modification of a speech spectrum. Spectral-GMM parameters extracted from the spectral envelope are spectral peaks, which may be related to formant information. To modify the spectral-GMM parameters in accordance with formant scaling factors, it is necessary to find relations between formants and the spectral-GMM parameters. When estimating spectral-GMM parameters from a spectral envelope, we just try to minimize the distance between the histogram of the spectral envelope and the Gaussian mixture model. As a result, there may be some components which contribute to one peak of the spectral envelope restored from the spectral-GMM parameters, which make it difficult to modify the spectral-GMM parameters.

In this subsection, we propose a new algorithm for modifying spectral-GMM parameters in accordance with formant frequencies. The spectral modification algorithm is described as follows, corresponding to Fig. 3.

We first extract spectral-GMM parameters from the smooth spectral envelope. In the next step, we find the peaks of the spectral envelope reconstructed from the spectral-GMM parameters. Since not all these peaks are formants, we have to decide how much these peaks will be shifted. For spectral modification, the first formants are most important, and often considered for modification. In this study, we also focus on modifying factors related to the first four formants. We isolate spectral regions of the input signal by dividing it into four non-overlapping bands (0–800 Hz, 800–2,500 Hz, 2,500–3,500 Hz, 3,500–sampling frequency/2 Hz) which cover the first four formant frequency ranges [24]. The scaling factor of each peak will be the scaling factor of the formant to which the peak belongs. Based on the geometric characteristics of the normal distribution, i.e. the empirical rule, we find which Gaussian components contribute to this peak. If this peak is located between $[\mu_m - 3\sigma_m; \mu_m + 3\sigma_m]$, where μ_m is the mean and σ_m is the standard deviation of Gaussian component m, we regard Gaussian component m as contributing to this peak. We shift the mean parameter of this Gaussian component by the scaling factor of this peak. In this paper, we only modify the mean parameters of Gaussian components, and we do not modify the other parameters of Gaussian components (i.e. standard deviations and mixture weights). Note that mean parameters are sorted in ascending order, and every mean parameter is

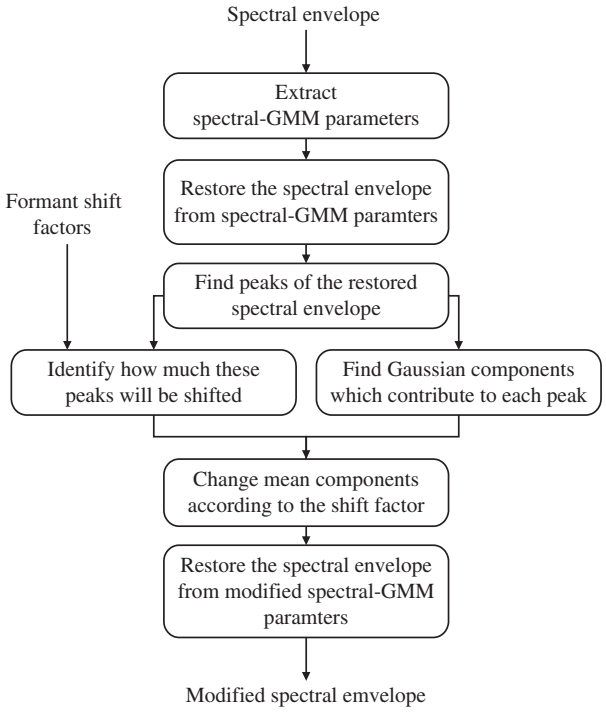

Fig. 3 Block diagram of our spectral modification algorithm.

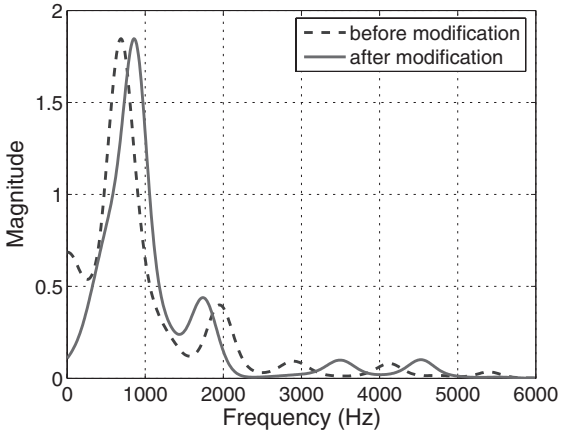

Fig. 4 Example of our spectral modification algorithm applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = -10\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$.

shifted only once. After shifting the mean parameters of the Gaussian components, we reconstruct the modified spectral envelope. Consequently, we can independently modify each spectral peak. An example of our proposed algorithm applied to a spectrum is shown in Fig. 4. For comparison with our method, an example of formant modification of the LSF-based method [8] is shown in Fig. 5. In the LSF-based method, since attributes of a formant depend on properties of a conjugate pole pair, when we change formant frequencies, amplitudes of a speech spectrum are also changed. On the contrary, we can control the spectral shape using our method.

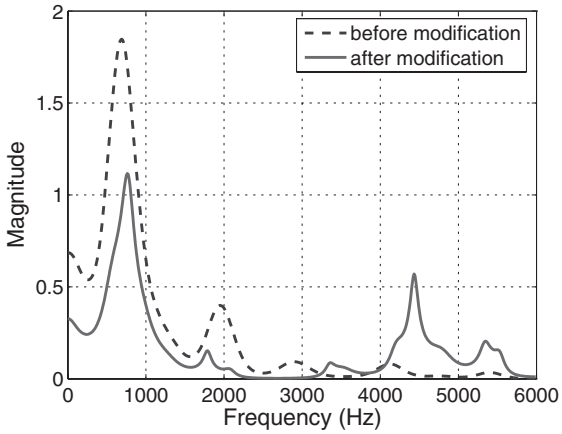

Fig. 5 Example of formant modification algorithm of the LSF-based method [8] applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = -10\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$.

4. EXPERIMENTS AND RESULTS

The two main themes of this paper evaluated in the experiments are (i) the effectiveness of our proposed modeling of speech spectra, and (ii) the effectiveness of our proposed spectral modification method. To evaluate the effectiveness of our proposed modeling of speech spectra, we use three objective measures in Subsection 4.1. To evaluate the effectiveness of our proposed spectral modification method, we investigate it in two areas, emotional speech synthesis in Subsection 4.2, and voice gender conversion in Subsection 4.3.

4.1. Evaluation of Our Proposed Modeling of Speech Spectra

In our proposed modeling of speech spectra, since we use spectral-GMM parameters to model each event target, the order of LSFs has to be high enough to precisely restore the spectral envelope. Via a small experiment, by calculating the average log spectral distortion (LSD) between STRAIGHT spectra and the spectral envelopes restored from LSFs with different orders in a set of 250 sentence utterances of the ATR Japanese speech database [25] at sampling frequency of 16 kHz, we chose the LSF order of 40 in this paper. With this order, the average LSD is smaller than 1 dB, and the average event rate is 24 events/second.

To evaluate the performance of our proposed modeling of speech spectra, we compared our proposed modeling of speech spectra (TD-GMM) with the framewise-GMM method. In the framewise-GMM method, the spectral-GMM parameters are used to model each spectral envelope, frame by frame. In this part, both methods used 10 Gaussian components to model the speech spectrum and each event target. The quality of the synthesized speech

Table 1 Analysis conditions for experiments of modeling of speech spectra.

Sampling frequency	8 kHz
Window length	40 ms
Window shift	1 ms
FFT points	1,024

Table 2 Average Euclidean MFCC distances for testing methods.

Original sounds	0.4103
Framewise-GMM method	0.4268
Proposed method	0.4067

Table 3 Average Euclidean distance between the original delta-MFCCs and the delta-MFCCs extracted from the two testing methods.

Original sounds and the corresponding sounds of the framewise-GMM method	0.3395
Original sounds and the corresponding sounds of the proposed method	0.2561

was evaluated by three objective measures. The first objective measure is used to evaluate the smoothness of synthesized speech utterances. The second objective measure is used to evaluate the modeling of spectral evolution. The last objective measure has high correlation with subjective listening tests.

A set of 150 sentence utterances of the ATR Japanese speech database [25] was selected as the speech data. This speech dataset is spoken by 6 speakers (3 male & 3 female) re-sampled at 8 kHz sampling frequency. The analysis conditions for these experiments are shown in Table 1.

For the first objective test, we used the Euclidean distance between mel-frequency cepstral coefficients (MFCC) (D_{MFCC}) as the objective measure, since this measure was found to be successful at predicting audible discontinuities in synthesized speech utterances in many studies [26]. For the second objective test, we used the Euclidean distance of delta mel-frequency cepstral coefficients (delta-MFCC) between natural and a synthesized spectral sequences ($D_{\text{delta-MFCC}}$). These criteria are defined as follows.

$$D_{\mathrm{MFCC}}(\boldsymbol{c1}, \boldsymbol{c2}) = \sqrt{\sum_{i=1}^{V_{\mathrm{MFCC}}} (\boldsymbol{c1}_i - \boldsymbol{c2}_i)^2} \tag{3}$$

$$D_{\text{delta-MFCC}}(\Delta\boldsymbol{c1}, \Delta\boldsymbol{c2}) = \sqrt{\sum_{i=1}^{V_{\mathrm{MFCC}}} (\Delta\boldsymbol{c1}_i - \Delta\boldsymbol{c2}_i)^2} \tag{4}$$

where $\boldsymbol{c1}$ and $\boldsymbol{c2}$ are MFCC coefficients of two consecutive frames. $\Delta\boldsymbol{c1}$ and $\Delta\boldsymbol{c2}$ are delta-MFCCs of a natural and corresponding synthesized utterances. V_{MFCC} is the MFCC order. The zeroth MFCC coefficient and the corresponding delta-MFCC coefficient are not included in the analysis, since they relate to the overall energy. Throughout this paper, the MFCC order of 24 has been used. In general, a smaller value for D_{MFCC} suggests a better system in terms of the smoothness of speech. A smaller value for $D_{\text{delta-MFCC}}$ suggests a better system in terms of modeling of spectral evolution. The average Euclidean MFCC distances and the average Euclidean delta-MFCC distances are shown in Table 2 and Table 3, respectively. According to a two-tailed t-test, experimental results are statically significant at a 95% confidence level (p-value $= 3.1 \cdot 10^{-4}$ for the first objective test, and p-value $= 1.8 \cdot 10^{-9}$ for the second objective test). Experimental results indicate that the performance of our proposed method is better than that of the framewise-GMM method in terms of both the smoothness of synthesized speech and the modeling of spectral evolution. The greater value of the average D_{MFCC} distance in the framewise-GMM method indicates that some non-smoothed areas occur using the framewise-GMM method.

For the third objective test, we used the perceptual evaluation of speech quality (PESQ) (ITU-T P.862). The PESQ uses a sensory model to compare the original, unprocessed signal (reference signal) with the degraded signal from a network or an analysis/synthesis system. The PESQ scores are calibrated using a large database of subjective tests. Having high correlation ($\rho > 0.92$) with subjective listening tests, the PESQ can be used reliably to predict the subjective speech quality of codec in a very wide range of conditions, including those with background noise, analogue filtering, and/or variable delay [27]. The score of PESQ ranges from -0.5 to 4.5. The higher the score, the better the perceptual quality. In this paper, since both the framewise-GMM method and our proposed method (TD-GMM method) estimate spectral-GMM parameters to model STRAIGHT spectral envelopes, we used the synthesized utterances restored from STRAIGHT (STRAIGHT sounds) as the reference signals, and the synthesized utterances restored using the framewise-GMM method and the TD-GMM method as the degraded signals. We calculated the PESQ between STRAIGHT sounds and sounds restored using the two methods (the framewise-GMM method and the TD-GMM method). The average PESQ results are shown in Table 4. According to a two-tailed t-test, experimental results are statically significant at a 95% confidence level (p-value $= 5.7 \cdot 10^{-27}$). These results indicate that our proposed method is better than the framewise-GMM method in terms of subjective speech quality.

Table 4 Average PESQ for testing methods.

Framewise-GMM method	3.5042
Proposed method	3.7416

4.2. Application to Emotional Speech Synthesis

In this subsection, we investigate our spectral modification method for emotional speech synthesis, where formant frequencies are shifted by small scaling factors (below 8 percent), and power envelopes need to be modified.

Huang and Akagi [28] proposed a novel model for the perception of emotional speech. Unlike most other studies that deal with the direct relationship between emotional speech and acoustic features, this model consists of three layers, emotional speech, semantic primitives, and acoustic features. This model is a rule-based conversion system, and it therefore requires controlling each parameter independently.

In [28], it was necessary to modify both power envelopes and formants. In the standard spectral modification techniques, such as [8], when formant frequencies are shifted, the magnitude of the speech spectrum is also changed accordingly. It is difficult to independently modify both power and formant frequencies with the defined scaling factors. To overcome this drawback, we employ our spectral modification method. Since our method uses spectral-GMM parameters to directly model and modify the spectral envelope, the magnitude of the speech spectrum is almost the same when formant frequencies are shifted, and each parameter's value can be modified independently. In addition, the smoothness of synthesized speech is ensured by using TD.

To verify the effectiveness of our spectral modification method, we conducted a listening experiment to compare it with the LSF-based method [8], which enabled a high level of control of formant characteristics. Both the LSF-based method and our spectral modification method had been applied in the work of Huang and Akagi [28], while other processes and morphing rules were kept the same. A neutral utterance was used to morph emotional utterances, e.g. cold anger, joy, sadness, and hot anger. The analysis conditions are listed in Table 5.

The subjective test was carried out using Scheffe's method of paired comparison [29]. In this subsection, five grades from -2 to 2 were used, as shown in Fig. 6. Eight Japanese graduate students known to have normal hearing ability were recruited for the listening experiment. Paired stimuli A and B were presented to each listener, and listeners were asked to grade stimuli according to his/her perception of speech quality. Experimental results are shown in Fig. 7. According to a two-tailed t-test, these results are statically significant at a 95% confidence level (p-value $= 9.2 \cdot 10^{-3}$). These experimental results also indicate that the speech quality of our proposed method is better than that of the LSF-based method [8]. In this application, since scaling factor is small (less than 8 percent), the difference of the results between the LSF-based and TD-GMM methods is small.

Table 5 Analysis conditions for experiments of emotional speech synthesis.

STRAIGHT	Sampling frequency	22.05 kHz
	Window length	40 ms
	Window shift	1 ms
	FFT points	1,024
LSF-based method	LSF order	24
Proposed method	Gaussian components	24

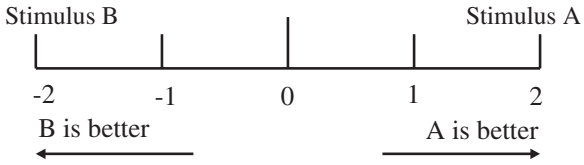

Fig. 6 Evaluation measure of Scheffe's paired comparison (five grades: -2, -1, 0, 1 and 2).

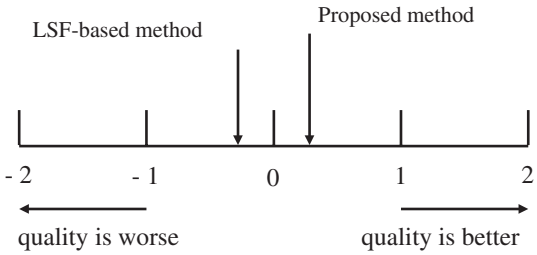

Fig. 7 Subjective listening results for emotional speech synthesis.

4.3. Application to Voice Gender Conversion

In Subsection 4.2, our proposed spectral modification method was effectively applied to shift the small formant frequencies, below 8 percent. In this subsection, we explore the effectiveness of our spectral modification method in a voice gender conversion (VGC) system which requires much spectral modification, about 20 percent.

The aim of VGC is to modify a female (male) speech so that it will sound as if it were spoken by a male (female). The VGC challenge is to convert the gender-related parameters of the speech signal without affecting smoothness and naturalness. For a long time it was believed that pitch was the dominant cue in voice gender perception. However, Childers and Wu [30] showed that grouped formant information was a slightly better determinant of gender than fundamental frequency information. Therefore, the two most important features that show major dif-

ferences between genders, formant frequencies and fundamental frequencies, are modified in our system. The formant frequencies are modified by our proposed method, and the fundamental frequency contour is modified by simply shifting the F0 mean by a scaling factor.

Our perception of spoken-voice gender relies heavily on the phonation or voicing process, which is associated mainly with vowel sounds. We first extracted the fundamental frequencies, and the first four formant frequencies from the five Japanese vowels spoken by two speakers (one male & one female) in the ATR Japanese speech database [25]. We then used these values to formulate the scaling factors for our VGC system. In this subsection, we used labeled data of each utterance to identify the distance between an event location and vowels. The scaling factors for an event target is the scaling factors of the vowel which was nearest to this event target.

To evaluate the performance of our proposed system, we conducted a listening experiment. We compared the performance of our system with the performance of two other systems. All three systems used STRAIGHT to modify fundamental frequencies. In the first system, the LSF-based method [8] was employed to modify formant frequencies (STRAIGHT+LSF). In the second system, speech spectra were modified frame by frame using only the framewise-GMM method, without using TD (STRAIGHT + framewise-GMM method).

A set of 50 sentence utterances of the ATR Japanese speech database was selected as the speech data. This dataset spoken by 2 speakers (one male & one female) was re-sampled at 12 kHz sampling frequency. The analysis conditions are listed in Table 6.

We randomly presented the synthesized sounds of each of six utterances which had been spoken by two speakers (one male & one female), to eight Japanese graduate students with normal hearing ability, and asked them to identify the gender of the person who was speaking, and to rate the perceptual quality of the speech on a five-point scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). The average scores are shown in Table 7. When comparing our proposed method (STRAIGHT + TD-GMM) with the first method (STRAIGHT + LSF), a two-tailed t-test at a 95% confidence level shows that the speech quality of our proposed method is superior to that of the first method for both kinds of conversions (p-value $= 1.0 \cdot 10^{-4}$ for male to female conversion, and p-value $= 1.7 \cdot 10^{-5}$ for female to male conversion). In this application, since scaling factor is large (more than 20 percent), the difference of the results between the LSF-based and TD-GMM methods is large. The LSF-based method cannot give acceptable voice quality, since the quality of most converted speech signals is diminished by a discernible buzzy sound. Our proposed method produces better voice quality than the other methods. The speech quality of the second method (STRAIGHT + framewise-GMM) and our proposed method are almost equivalent (p-value of a two-tailed t-test at a 95% confidence level for male to female and female to male conversions are 0.7529 and 0.6802, respectively). According to the experimental results in Subsection 4.1, our proposed method is better than the framewise-GMM method in terms of spectral modeling. In this application, there are no reference speakers, and the modified speech of the framewise-GMM and our proposed methods are hardly perceptually distinguishable. The reason is that we used the same scaling factor for every frame in a vowel in this application. Therefore, the smoothness of spectra is preserved in voiced frames when modifying, and the effectiveness of TD is not shown clearly. It should be noted that both the second method and our proposed method used the algorithm in Section 3 to perform spectral modification.

Table 6 Analysis conditions for experiments of VGC system.

STRAIGHT	Sampling frequency	12 kHz
	Window length	40 ms
	Window shift	1 ms
	FFT points	1,024
LSF-based method	LSF order	14
Framewise-GMM method	Gaussian components	14
Proposed method	Gaussian components	14

Table 7 Subjective listening results for VGC system (1) STRAIGHT + LSF (2) STRAIGHT + framewise-GMM (3) our proposed system (STRAIGHT + TD-GMM).

Type of conversion	Correct gender identification (%)			Quality evaluation score		
	(1)	(2)	(3)	(1)	(2)	(3)
M to F	83.3	93.8	93.8	2.73	3.15	3.19
F to M	100	100	100	3.10	3.58	3.63

5. CONCLUSIONS

In this paper, we have presented a new modeling of speech spectra based on TD and spectral-GMM, and then developed a new algorithm for modifying spectral-GMM parameters in accordance with formant scaling factors. We utilize TD to model the spectral evolution, and spectral-GMM parameters to model the event targets. Our proposed method not only effectively describes the temporal trajectories between frames, but also flexibly models the event targets. Moreover, processing rules are more effectively applied, since we only need to process the event targets,

instead of processing frame by frame, and the event targets may be associated with ideal articulatory positions. Our proposed method is especially useful when we change the speech spectra by large factors, while conventional methods can not make great changes. The experimental results, in terms of objective and subjective measures, prove the effectiveness of our proposed method.

There are however issues which still remain to be solved. In this paper, we only model and modify the event targets. Since the event functions describe the spectral evolutions of the event targets, we are convinced that these event functions may also contain useful information. Modeling the event functions therefore should be implemented for easy and effective processing, and this consideration will be explored in our future work. In addition, in this paper, we only change mean parameters of Gaussian components to perform spectral modification. It is well-known that amplitudes and bandwidths of spectral peaks are also important. The next stage of this research is how to change other Gaussian components (i.e. standard deviations and mixture weights) to modify amplitudes and bandwidths of spectral peaks.

ACKNOWLEDGMENTS

This study was supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan. We would like to thank anonymous reviewers for their suggestions and comments that improved the quality of our paper.

REFERENCES

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

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

[3] R. Lawlor and A. D. Fagan, "A novel efficient algorithm for voice gender conversion," *XIVth International Congress of Phonetic Sciences*, University of California, Berkeley, USA, pp. 77–80 (1999).

[4] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," *Proc. Interspeech*, pp. 2401–2404 (2003).

[5] D. Giuliani, M. Gerosa and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *J. Comput. Speech Lang.*, **20**, 107–123 (2006).

[6] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by additive noise," *Proc. ICASSP*, pp. 208–211 (1979).

[7] H. Mizuno, M. Abe and T. Hirokawa, "Waveform-based speech synthesis approach with a formant frequency modification," *Proc. ICASSP*, pp. 195–198 (1993).

[8] R. W. Morris and M. A. Clements, "Modification of formants in the line spectrum domain," *IEEE Signal Process. Lett.*, **9**, 19–21 (2002).

[9] E. Turajlic, D. Rentzos, S. Vaseghi and C.-H. Ho, "Evaluation of methods for parameteric formant transformation in voice conversion," *Proc. ICASSP*, pp. 724–727 (2003).

[10] M. E. Lee, "Acoustic models for the analysis and synthesis of the singing voice," *Ph.D. Dissertation, Georgia Institute of Technology* (2005).

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

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

[13] P. C. Nguyen, T. Ochi and M. Akagi, "Modified restricted temporal decomposition and its application to low bit rate speech coding," *IEICE Trans. Inf. Syst.*, **E86-D**, 397–405 (2003).

[14] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of Gaussians," *Proc. ICSLP*, pp. 1229–1232 (1996).

[15] P. Zolfaghari, S. Watanabe, A. Nakamura and S. Katagiri, "Bayesian modelling of the speech spectrum using mixture of Gaussians," *Proc. ICASSP*, pp. 553–556 (2004).

[16] P. Zolfaghari, H. Kato, Y. Minami, A. Nakamura, S. Katagiri and R. Patterson, "Dynamic assignment of Gaussian components in modelling speech spectra," *J. VLSI Signal Process. Syst.*, **45**, 7–19 (2006).

[17] B. P. Nguyen and M. Akagi, "A flexible spectral modification method based on temporal decomposition and Gaussian mixture model," *Proc. Interspeech*, pp. 538–541 (2007).

[18] B. P. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," *J. Signal Process.*, **11**, 333–336 (2007).

[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] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Process.*, 3–14 (1993).

[21] K. K. Paliwal, "Interpolation properties of linear prediction parametric representations," *Proc. Eurospeech*, pp. 1029–1032 (1995).

[22] 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," *J. Speech Commun.*, **27**, 187–207 (1999).

[23] A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B*, **39**, 1–38 (1977).

[24] A. Kain, Q. Miao and J. van Santen, "Spectral control in concatenative speech synthesis," *Proc. 6th ISCA Workshop on Speech Synthesis* (2007).

[25] M. Abe, Y. Sagisaka, T. Umeda and H. Kuwabara, "Speech database user's manual," *ATR Tech. Rep.*, TR-I-0166 (1990).

[26] B. Kirkpatrick, D. O'Brien and R. Scaife, "A comparison of spectral continuity measures as a join cost in concatenative speech synthesis," *Proc. Ir. Signals Syst.*, 515–520 (2006).

[27] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Proc. ICASSP*, pp. 749–752 (2001).

[28] C. F. Huang and M. Akagi, "A rule-based speech morphing for verifying an expressive speech perception model," *Proc. Interspeech*, pp. 2661–2664 (2007).

[29] H. Scheffe, "An analysis of variance for paired comparisons," *J. Am. Stat. Assoc.*, **37**, 381–400 (1952).

[30] D. G. Childers and K. Wu, "Gender recognition from speech. Part II: Fine analysis," *J. Acoust. Soc. Am.*, **90**, 1841–1856 (1991).

Binh Phu Nguyen received the B.E. degree in Electronic & Telecommunication Engineering, and the M.E. degree in Information Technology from Hanoi University of Technology, Vietnam, in 1997 and 2004, respectively. He is currently working towards the Ph.D. degree in Information Science at Japan Advanced Institute of Science and Technology (JAIST), Japan. His research interests include spectral modification and speech analysis/synthesis. He is a member of the Acoustical Society of Japan (ASJ), and International Speech Communication Association (ISCA). Mr. Nguyen received the Student Paper Award at NCSP' 07.

Masato Akagi received the B.E. degree from Nagoya Institute of Technology in 1979, and the M.E. and PhD. Eng. degrees from Tokyo Institute of Technology in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), and is now a professor. His research interests include speech perception, the modeling of speech perception mechanisms of human beings, and signal processing of speech. Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, and the Sato Prize for Outstanding Paper from the ASJ in 1998 and 2005.