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, Ishikawa, 923-1292, Japan
{npbinh, akagi}@jaist.ac.jp

10.21437/Interspeech.2007-244

Abstract

This paper presents a new spectral modification method to solve two drawbacks of conventional spectral modification methods, insufficient smoothness of the modified spectra between frames and ineffective spectral modification. To overcome the insufficient smoothness, a speech analysis technique called temporal decomposition (TD) is used to model the spectral evolution. Instead of modifying the speech spectra frame by frame, we only need to modify event targets and event functions, and the smoothness of the modified speech is ensured by the shape of the event functions. To overcome the ineffective spectral modification, we explore Gaussian mixture model (GMM) parameters for an input of TD to model the spectral envelope, and develop a new method of modifying GMM parameters in accordance with formant scaling factors. Experimental results show that the effectiveness of the proposed method is verified in terms of the smoothness of the modified speech and the effective spectral modification.

Index Terms: spectral modification, Temporal Decomposition, Gaussian mixture model, STRAIGHT

1. Introduction

Spectral modification techniques are capable of performing a variety of modifications to a speech spectra such as manipulations of the formant structures, amplitude manipulations, and so on. They can be applied to many applications, such as transforming the identity of a speaker, enhancing speech, etc.. The challenge of spectral modification is to modify the spectra 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: LPC-based methods [1, 2] and frequency warping methods [3]. LPC-based methods often meet 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. [1]. While this method produces spectral envelopes with desired formant amplitudes at the formant frequencies, one drawback to this technique is that the bandwidth of each formant cannot be controlled. Recently, a method for modifying formant locations and bandwidths directly in the line spectral frequency (LSF) domain has been developed in [2]. By taking advantage of the nearly linear relationship between the LSFs and formants, modifications are performed based on desired shifts in formant frequencies and bandwidths. However, the main drawback to this type of modification, the lack of control over the spectral shape, has not been solved. Frequency warping methods such as [3] allow a high level of control over formant characteristics, but only when the original and modified formants are spaced far enough apart so as to be nearly independent of one another. When positions of formants are too close to one another, it is difficult to modify their bandwidths to appropriate specifications. This is similar to the pole interaction problem suffered by pole modification techniques.

In addition, some methods mentioned above [1, 2] only mention the way to modify the spectral modification in a frame, and all of them [1, 2, 3] rarely deal with constraints between frames after modification. When there are unexpected modifications in some frames, the modified speech may be not smooth. As a result, there are some clicks in the modified speech, which lead to a degradation of speech quality.

In this paper, we propose a new spectral modification method based on temporal decomposition [4] and Gaussian mixture model (GMM) [5, 6]. To model the spectral evolution, we employ the modified restricted temporal decomposition (MRTD) algorithm [7]. For spectral modification, we use GMM parameters [5, 6] to model the speech spectrum, and develop a new method to modify GMM parameters in accordance with formant scaling factors. Note that the GMM parameters used here are different from those often used to model the distribution of acoustic features in state-of-the-art methods for voice conversion. We evaluate the effectiveness of the proposed method in two areas which require different amounts of spectral modification, voice gender conversion and emotional speech synthesis.

2. Spectral modification based on Temporal Decomposition and Gaussian mixture model

2.1. Temporal decomposition

As mentioned earlier, a shortcoming of conventional spectral modification methods is that they do not take into account the correlation between frames, resulting in clicks in the modified speech because of the discontinuous spectral contour. Therefore, we employ TD to solve this 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. Because of the overlap of the gestures, these phonetic targets are often only partly realized.

Atal proposed a method based on the so-called temporal decomposition of speech into a sequence of overlapping target functions and corresponding target vectors [4], in which the target vectors may be associated with ideal articulatory positions, and the target functions describe the temporal evolution of these targets, as given in Eq. (1).

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

where $\mathbf{a}_k$, the k^{th} event vector, is the speech parameter corresponding to the k^{th} 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.

To modify the speech spectra, we only need to modify the speech spectrum of each event vector and the corresponding event function 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 spectral envelope, as well as ensuring the smoothness of the spectral envelopes between frames, and thereby enhances the modified speech quality.

The original method of TD is known to have two major drawbacks of high computational cost and high parameter sensitivity to the number and locations of events. A number of modifications have been explored in the literature to overcome these drawbacks. In this paper, we employ the MRTD algorithm [7]. The reasons for using MRTD in this work are twofold: (i) the MRTD algorithm enforces a new property on the event functions, named the "well-shapedness" property, to model the temporal structure of speech more effectively [7]; (ii) event targets can convey the speaker identity [8].

2.2. Speech spectrum modeling using Gaussian mixture model (GMM)

One of the most important properties of spectral modification is that it is flexible enough to perform a variety of modifications within the spectral envelope. The standard spectral modification techniques are limited by their inability to independently control important formant characteristics such as amplitude and bandwidth.

Zolfaghari et al. proposed a technique to fit a set of Gaussian mixtures to the smoothed magnitude spectrum of a speech signal [5, 6]. The estimated means, standard deviations, and mixture weights of the Gaussians can be related to the locations, bandwidths, and amplitudes of the formants, respectively. The ability to independently control the parameters of each Gaussian component enables a precise estimate of the spectral envelope, a wide variety of modifications, as well as independent control of the formants.

2.3. Smoothed-spectrum representation by STRAIGHT

The characteristic shape of the speech spectrum can present problems for estimating a set of Gaussian components. The voiced speech spectrum is characterized by a number of pitch peaks separated by the fundamental frequency. If the pitch peaks are separated by a high fundamental frequency, a maximum can be found by estimating a Gaussian component for a single-pitch peak, and ignoring the adjacent harmonics. This results in a very small variance for that Gaussian. Therefore, the high-quefrency effects of the excitation from the spectrum are removed to improve the representation of the spectral envelope by the Gaussian mixture fitting method. In this paper, we model the STRAIGHT spectral envelope using mixture of Gaussians. STRAIGHT [9] uses a pitch-adaptive spectral analysis scheme combined with a surface reconstruction method in the time-frequency plane to remove signal periodicity. This results in a smooth spectral representation free of glottal excitation information. Fig. 1 shows an estimated mixture of six Gaussians, and an STRAIGHT smoothed spectrum that is obtained by the analysis of one frame of speech.

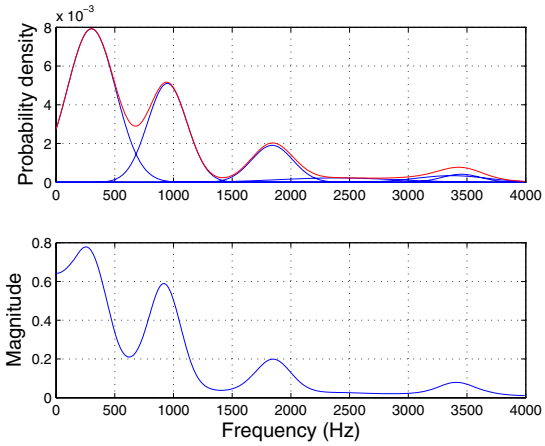

Figure 1: Mixture of Gaussians (6 components) fit to an STRAIGHT smoothed spectrum (top) and an STRAIGHT smoothed spectrum of one frame (bottom).

2.4. GMM parameters as an input of TD

As mentioned earlier, speech spectrum modeling using GMM enables a precise estimate of the spectral envelope, a wide variety of modifications, as well as independent control of the formants in a frame. However, if frames are processed independently, it may generate discontinuous features. To overcome this drawback, we investigate GMM parameters as an input of TD. Using TD and GMM, we can deal with the two drawbacks of conventional spectral modification methods, the insufficient smoothness of the modified spectra between frames and the ineffective spectral modification.

Among GMM parameters, the mean components are the most significant parameters, since they are related to formant locations. To apply TD for analyzing GMM parameters, only the mean components are used as input parameters in this paper.

3. New spectral modification algorithm

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 parameters that are directly connected to the speech production process. GMM parameters extracted from the spectral envelope are related to formant information. However, this scheme is not a formant detector in terms of obtaining the resonances in the speech signal. To modify GMM parameters in accordance with formant scaling factors, it is necessary to find a relation between formants and GMM parameters. We propose a new method of modifying GMM parameters in accordance with formant frequencies. The spectral envelope modification algorithm is described as follows, corresponding to Fig. 2.

We first extract GMM parameters from the smooth spectral envelope. In the next step, we find the peaks of the spectral envelope reconstructed from the GMM parameters. Since not all these peaks are formants, we have to identify by how much these peaks will be shifted. We isolate spectral regions of the input signal by dividing it into N non-overlapping bands which correspond to the first N formant frequency ranges. Scaling factor of each peak will be the scaling factor of the formant to which the peak belongs. N is defined by the requirement of each specific task. Based on the geometric characteristic of normal distribution, i.e. Empirical rule, we find which GMM components contribute to this peak. If this peak is located be-

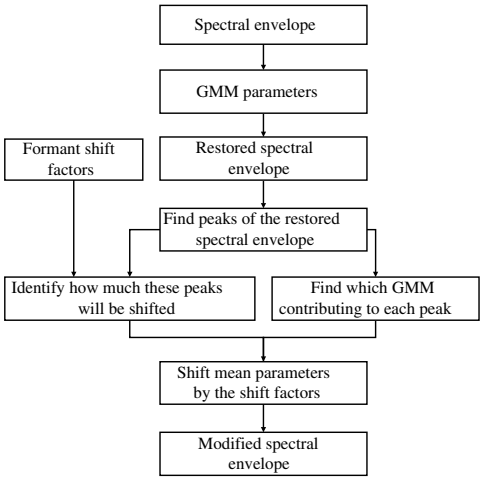

Figure 2: Block diagram of spectral modification algorithm.

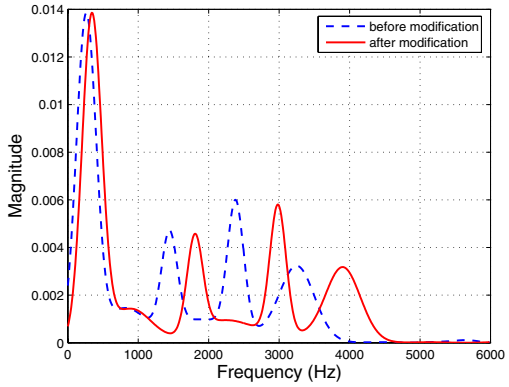

Figure 3: Example of spectral envelope modification algorithm applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = 25\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$.

tween $[\mu_m - 3\sigma_m; \mu_m + 3\sigma_m]$, 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. Note that every mean parameter is shifted only once. After shifting the Gaussian components, we reconstruct the modified spectral envelope. An example of the proposed algorithm applied to a spectrum is shown in Fig. 3.

4. Experiments and results

In order to evaluate the effectiveness of the proposed method, we investigate it in two different areas which require different amounts of spectral modification, voice gender conversion and emotional speech synthesis.

4.1. Application to voice gender conversion

The aim of voice gender conversion is to modify female (male) speech so that it sounds as if it was spoken by a male (female). The voice gender conversion 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 [10] showed that grouped formant information gave a higher automatic gender distinction success rate than pitch information. Therefore, the two most important features which show major differences across gender, formant frequencies and fundamental frequencies, are modified in our system. The processing flow of our voice gender conversion system is as follows.

Table 1: Analysis conditions for experiments of voice gender conversion.

STRAIGHT	Sampling frequency	12 kHz
	Window length	40 ms
	Window shift	1 ms
	FFT points	1024
Proposed method	Iteration of EM algorithm	30 times
	GMM components	14
Method in [2]	LSF order	14

First, STRAIGHT decomposes input speech signals into spectral envelopes, F0 (fundamental frequency) information, and aperiodic components. Since the spectral envelopes can be further analyzed into GMM parameters, MRTD is employed in the next step to decompose the mean components of GMM parameters into event targets and event functions. These targets are modified in accordance with shift factors, and then resynthesized as mean parameters by TD reconstruction. In the next step, the modified GMM parameters are synthesized as spectral envelopes by GMM synthesis. The fundamental frequency contour is modified by simply shifting the F0 mean by a scaling factor. Finally, STRAIGHT synthesis is employed to output the modified speech.

To evaluate the performance of the system, a number of experiments were conducted.

Our perception of spoken-voice gender relies heavily on the phonation or voicing process, which is associated mainly with vowel sounds. We therefore extract the fundamental frequency, and the first four formant frequencies from the five Japanese vowels spoken by two speakers (one male and one female) in the ATR Japanese speech database [11] to formulate the scaling factors for our voice gender conversion system. To modify other syllables, we use the same scaling factors of the vowel that is nearest to the syllable.

We then compare the performance of the system with the performance of two other systems. All three systems use STRAIGHT to modify the fundamental frequencies. In the first system, a newly proposed algorithm for formant modification in the LSF domain [2] is employed to modify formants frame by frame (STRAIGHT+LSF). In the second system, speech is converted frame by frame using only GMM parameters to modify the spectral envelopes, without using TD (STRAIGHT+GMM). Six utterances of the ATR Japanese speech corpus spoken by two speakers (one male and one female) are used for evaluation. The analysis conditions are listed in Table 1.

We presented the synthesized sounds to 8 listeners, 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). Table 2 shows the average scores, which indicate that the speech quality of the proposed method is superior to that of the first system and slightly better than that of the second system.

4.2. Application to emotional speech synthesis

In Subsection 4.1, the proposed spectral modification method was effectively applied to shifting large formant frequencies, approximately 20 percent. In this subsection, we investigate our proposed spectral modification method in emotional speech synthesis, where formant frequencies are shifted by small scal-

Table 2: Subjective listening results for voice gender conversion (1) STRAIGHT + LSF (2) STRAIGHT + GMM (3) the proposed system (STRAIGHT + TD + GMM).

Type of Conversion	Correct Gender Identification (%)			Mean Opinion 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

Table 3: Analysis conditions for experiments of emotional speech synthesis.

STRAIGHT	Sampling frequency	22 kHz
	Window length	40 ms
	Window shift	1 ms
	FFT points	1024
Proposed method	Iteration of EM algorithm	30 times
	GMM components	24
Method in [2]	LSF order	24

ing factors, about 10 percent, and power envelopes need to be modified.

Huang and Akagi propose a novel model for the perception of emotional speech [12]. 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.

In the work of Huang and Akagi [12], it is necessary to modify both power envelopes and formants. In the standard spectral modification techniques, such as [2], when formant frequencies are shifted, the amplitudes of FFT bins are also changed dependently. It is difficult to independently modify both power and formant frequencies with the defined scaling factors. To overcome this drawback, we employ our proposed method. Since our method uses GMM parameters to directly model and modify the spectral envelope, the amplitudes of FFT bins are almost the same when formant frequencies are shifted, and each GMM parameter's values can be modified independently. In addition, the smoothness of synthesized speech is ensured by using TD.

To verify the effectiveness of our proposed method, we have conducted experiments to compare it with the formant modification method in [2] which enables a high level of control over formant characteristics. Both methods were in turn applied to [12], while other processes and morphing rules were kept the same. A neutral speech was used to morph emotional utterances, e.g. cold anger, joy, sad, and hot anger. The analysis conditions are listed in Table 3.

We presented the emotional utterances created from two methods to 8 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 4 shows the average scores, which indicate that the speech quality of the proposed method is slightly better than that of the method in [2]. The quality of synthesized emotional speech can be referred to [12].

5. Conclusions

In this paper, we have presented a new method for spectral modification to solve the two drawbacks of standard spectral modification techniques, the insufficient smoothness of the modified spectra between frames and the ineffective spectral modification. The method ensures the smoothness of modified speech by using TD to model the spectral evolution. The method also overcomes the problem of ineffective spectral modification by using GMM to model and modify the spectral envelope. To evaluate the effectiveness of the proposed method, we applied it to voice gender conversion requiring much spectral modification, and to emotional speech synthesis requiring little spectral modification. The experimental results prove the effectiveness of the proposed method.

There is however an issue which is still open. In this paper, the proposed method only ensures the smoothness of mean components of GMM parameters. We are convinced that other parameters can be decomposed by TD by investigating relations among parameters or using other event functions, and this will be explored in our future work.

Table 4: Subjective listening results for emotional speech synthesis.

Method [2]	Proposed method
3.45	3.52

6. Acknowledgements

A part of this study was supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan.

7. References

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

[2] Morris, R. W. and Clements, M. A., "Modification of formants in the line spectrum domain," IEEE Signal Processing Lett., Vol. 9, No. 1, pp. 19-21, 2002.

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

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

[5] Zolfaghari, P. and Robinson, T., "Formant analysis using mixtures of Gaussians," Proc. ICSLP, pp. 1229-1232, 1996.

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

[7] Nguyen, P. C., Ochi, T. and Akagi, M., "Modified restricted temporal decomposition and its application to low bit rate speech coding," IEICE Trans. Inf. Syst., Vol. E86-D, No. 3, pp. 397-405, March 2003.

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

[9] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A., "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., Vol. 27, No. 3-4, pp. 187-207, 1999.

[10] Childers, D. G. and Wu, K., "Gender recognition from speech. Part II: Fine analysis," J. Acoust. Soc. Amer., Vol. 90, pp. 1841-1856, 1991.

[11] Abe, M., Sagisaka, Y., Umeda, T., and Kuwabara, H., "Speech database user's manual," ATR Technical Report, TR-I-0166, 1990.

[12] Huang, C-F. and Akagi, M., "A rule-based speech morphing for verifying an expressive speech perception model," Proc. Interspeech, 2007, to appear.