Journal of Signal Processing, Vol. 11, No. 4, pp.333-336, July 2007

SELECTED PAPER

Spectral Modification for Voice Gender Conversion Using Temporal Decomposition

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 most state-of-the-art voice gender conversion systems, the converted speech still sounds unnatural, which is mainly attributed to the insufficient smoothness of the converted spectra between frames and ineffective spectral modification. In this paper, we present a new method for voice gender conversion using a speech analysis technique called temporal decomposition (TD). TD is used to model spectral evolution effectively. Instead of modifying speech spectra frame by frame, we only need to modify event targets and event functions, and the smoothness of the converted speech is ensured by the shape of the event functions. To overcome the ineffective spectral modification, we explore Gaussian mixture model (GMM) parameter sets for an input of TD to flexibly model the spectral envelope, and develop a new method of modifying GMM parameters in accordance with formant scaling factors. For transforming fundamental frequencies, our system is based on STRAIGHT, which is a very high-quality vocoder. Experimental results show that the quality of the speech converted by the proposed method is significantly improved.

1. Introduction

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). Voice gender conversion has applications in voice output systems such as Text-to-Speech synthesis, multimedia voice applications, or in voice gender normalization for improved speech compression or recognition. The challenge of voice gender conversion 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 [1] showed that grouped formant information gave a higher automatic gender distinction success rate than pitch information. Therefore, both the glottal and vocal-tract-related features of the source speech signal need to be modified in voice gender conversion systems.

A variety of approaches to voice gender conversion have been discussed. Most voice gender conversion methods are based on a parametric source-filter model of speech production [2, 3, 4]. In [2] and [3], formant modification is performed by linear frequency-scale mapping applied to the speech spectrum. This does not accurately reflect the frequency difference between male and female speech, and the quality of the converted speech is not very natural. To overcome the disadvantages of linear frequency-scale mapping, Jung et al. [4] refined the method proposed in [3] by splitting the speech signal into two complementary frequency bands to separate F4 from the other formants, and modifying each subband with different formant scaling factors. However, this method still uses LP (linear prediction) coefficients to represent and modify the spectral envelope. Because of the limitation of standard LP-based techniques in independently modifying important formant characteristics such as amplitude and bandwidth, the quality of the speech is not enhanced.

In addition, all the methods mentioned above modify the speech spectra and fundamental frequency frame by frame, and rarely apply any constraints between frames. When there are unexpected modifications in some frames, the modified speech may be not smooth. As a result, there are some clicks in the converted speech, which lead to a degradation of speech quality.

In this paper, we propose a new voice gender conversion system using temporal decomposition [5]. To model the spectral evolution, we employ the modified restricted temporal decomposition (MRTD) algorithm [6]. For spectral modification, we use GMM parameters [7, 8] to model the speech spectrum, and develop a new method of modifying 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. 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 [9] is utilized in this study.

2. Spectral Modification using Temporal Decomposition

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 temporal decomposition of speech into a sequence of overlapping target functions and corresponding target vectors [5], 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).

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

where 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, $\phi_k(n)$. $\hat{\mathrm{y}}(n)$ is the approximation of the n^{th} spectral parameter vector 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 speech spectra, we only need to modify the speech spectra of event vectors and the corresponding event functions instead of modifying the speech spectrum 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 modified spectra between frames, and thereby enhances the converted 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 to overcome these drawbacks. In this study, we employ the MRTD algorithm [6]. 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 [6]; (ii) event targets can convey the speaker's identity [10].

2.2 Speech spectrum modeling using Gaussian mixture model (GMM)

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

Zolfaghari and coworkers proposed a technique applying the expectation maximization (EM) algorithm to fit a set of Gaussian mixtures to the smoothed magnitude spectra of a speech signal [7, 8]. 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, and 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 work, we model the STRAIGHT smoothed spectrum using a mixture of Gaussians. STRAIGHT is fundamentally a source-filter-type vocoder designed for the high-quality analysis/modification/synthesis of speech. It 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 distribution of six Gaussians, and an STRAIGHT smoothed spectrum that is obtained by the analysis of one frame of speech.

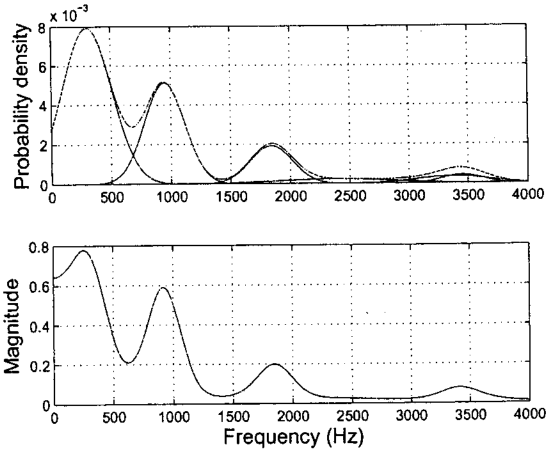

Figure 1: Mixture of Gaussians (6 components) fit to an STRAIGHT smoothed spectrum (top), and an STRAIGHT smoothed spectrum of one frame (bottom)

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, and 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 study. Although it is undesirable to have a mixed set of input parameters at several stages of the TD procedure, other parameters can be decomposed by TD by investigating the relations among parameters or using other event functions, and this will be explored in our future work.

3. New Spectral Modification Algorithm

Formant frequency is one of the most important parameters in characterizing speech, and using formant frequency as a parameter can control parameters that are directly connected to the speech production process. GMM parameters are related to formant information, but they are not true formants 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 modification algorithm is described as follows.

We first extract GMM parameters from the smooth spectral envelope. The precise estimate of the spectral envelope depends on many factors, such as the number of Gaussian components estimated, the sampling frequency, and the number of iterations used in the EM algorithm. 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 divide the frequency range into 4 subbands corresponding to the first four formant frequency ranges, and the scaling factor of each peak is determined to be the scaling factor of the formant to which the peak belongs. Then, on the basic the geometric characteristic of normal distribution, i.e. the empirical rule, we find which GMM 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 component of this Gaussian component by the scaling factor of this peak. Note that every mean component is shifted only once. After shifting the Gaussian components, we reconstruct the modified spectral envelope. An example of spectral modification algorithm applied to a spectrum is shown in Fig. 2.

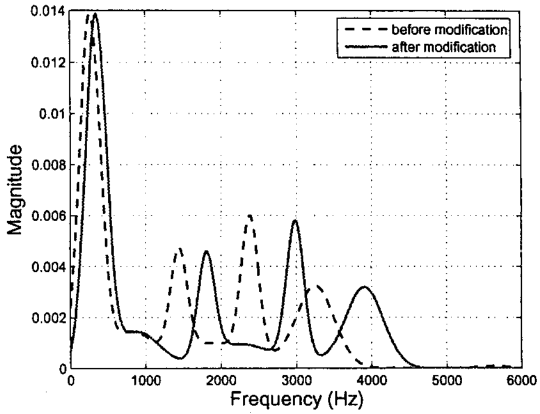

Figure 2: Example of spectral envelope modification algorithm applied to a spectrum: $\Delta F1 = 30\%$, $\Delta F2 = 25\%$, $\Delta F3 = 20\%$, and $\Delta F4 = 15\%$

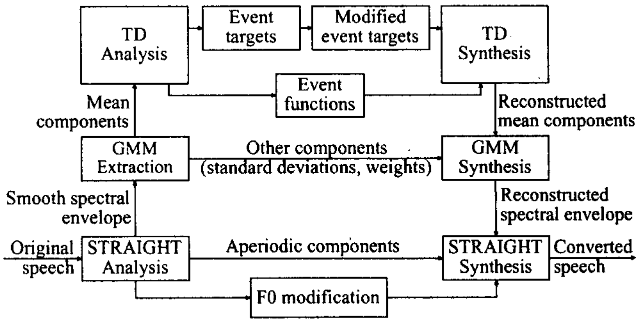

Figure 3: Block diagram of the proposed voice gender conversion system

4. Proposed Voice Gender Conversion System

As mentioned above, the two most important features that show major differences between genders, formant frequencies and fundamental frequencies, are modified in our system. A block diagram of the proposed voice gender conversion system is shown in Fig. 3.

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 re-synthesized 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 multiplying the F0 by a scaling factor. Finally, STRAIGHT synthesis is employed to output the modified speech.

Table 1: Analysis conditions for experiments

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 [12]	LSF order	14

Table 2: Subjective listening results (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

5. Experiments and Results

To evaluate the performance of the proposed method, 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 quality of speeches converted by the proposed method with those converted by two other systems. All three systems use STRAIGHT to modify the fundamental frequency. In the first system, a newly proposed algorithm for formant modification in the LSF domain [12] is employed (STRAIGHT+LSF). In the second system, speech is converted frame by frame by using only GMM parameters to modify the spectral envelope (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 subjective quality of the proposed method is superior to that of the first system and slightly better than that of the second system.

6. Conclusions

In this paper, we have presented a new method for voice gender conversion. The method ensures the smoothness of converted speech by TD. The method also overcomes the problem of ineffective spectral modification. The effectiveness of the proposed method was confirmed by subjective test results. Because of time limitation, we only focused on the two most important features related to gender difference, formant frequencies and fundamental frequencies. We believe that further improvement can be made by analyzing the air transition difference between male and female speakers. To utilize this difference in our proposed voice conversion system, we only need to modify the event functions.

Acknowledgments

A part of this study was supported by SCOPE (071705001) of MIC, Japan.

References

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

[2] B. S. Atal and S. L. Hanauer: Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Am., Vol. 50, No. 2 (Part II), pp. 637-655, 1971.

[3] R. Lawlor and A. D. Fagan: A novel efficient algorithm for voice gender conversion, XIVth Int. Congress of Phonetic Sciences, Berkeley, USA, 1999.

[4] E. Jung, A. Th. Schwarzbacher, K. Humphreys and R. Lawlor: Application of real-time AMDF pitch-detection in a voice gender normalisation system, Proc. ICSLP, pp. 2521-2524, 2002.

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

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

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

[8] 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.

[9] 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., Vol. 27, No. 3-4, pp. 187-207, 1999.

[10] 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.

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

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