Studies on Spectral Modification in Voice Transformation

Binh Phu Nguyen
School of Information Science,
Japan Advanced Institute of Science and Technology

March, 2009

Abstract

This dissertation aims to propose spectral modelings and spectral modification algorithms to improve the quality of modified speech in voice transformation.

Voice transformation is a process of changing certain perceptual properties of speech while leaving other properties unchanged. Voice transformation has many applications in our lives. For example, we employ voice transformation techniques to create various wave sounds from a limited pre-recorded database in a Text-to-Speech system. In foreign language learning, it will be much easier to listen when slowing down the speed of sounds. To enhance the hearing abilities of deaf people, we can adjust the frequency of sounds so that it is located in their hearing portion.

One of the core processes in voice transformation is spectral modification. Since spectral processing is closely linked to human perception, it is an effective way to perform sound processing, such as manipulations of the formant structures, amplitude manipulations. The challenge of spectral modification is to modify the spectral/acoustical features without degrading the speech quality. Most high-quality spectral modification methods operate in time-frequency domain. The representations of speech signals in the time-frequency domain can describe the speech signals well. The reason is that when a human produces voices, the continual motion of the articulators forms a time-varying acoustic filter which is responsible for the generation of the speech waveform.

Although many high-quality spectral modification methods have been proposed in the literature, they still have issues. In the time axis, most of them process speech signals frame by frame. They lack of a model to describe the temporal evolution of parameters. Therefore, they do not ensure the smoothness of synthesized speech after modification. In addition, they either do not guarantee to keep the natural evolution of parameters of speech signals. These limitations degrade the quality of modified speech. In the frequency axis, spectral modification can be performed by the rule-based approach or the statistical approach. The rule-based approach often needs small training data. However, this approach only stores basic rules, and the basic rules do not reflect the natural characteristics of the speech signals. While the statistical approach performs spectral modification by using machine learning techniques, and it requires large training data. Therefore, applying which approach depends on applications and specific conditions. Three main issues of spectral modification methods can be summarized as follows.

1. The first issue is lack of efficient spectral modelings for speech modification. The spectral modelings determine which attributes can be processed and how these attributes can be modified. One of requirements of these spectral modelings is that they allow to ensure smoothness of modified speech signals and to perform efficient spectral modification.

2. The second issue is insufficient smoothness of modified spectra between frames. Conventional methods of spectral modification often perform spectral modification frame by frame. When unexpected modifications happen, there are discontinuities of speech spectra between frames. This leads to degradation of the quality of modified speech.

3. The third issue is ineffective spectral modification. The rule-based techniques of spectral modification, such as linear prediction (LP)-based methods and frequency warping methods, are limited by their inability to independently control important formant characteristics such as amplitude, bandwidth or to control the spectral shape. The statistical techniques of spectral modification take advantages of mathematical or statistical models, but they lack of acoustic constraints. It therefore requires a new method for performing efficient spectral modification.

The main goals of this dissertation are to deal with the three major issues of spectral modification mentioned above, i.e. lack of spectral modelings for speech modification, insufficient smoothness of the modified spectra between frames, and ineffective spectral modification.

To perform spectral modification, we first develop an analysis/synthesis framework. When human beings produce speech voices, the continual motion of the articulators forms a time-varying acoustic filter which is responsible for the generation of the speech waveform. To characterize this type of properties, we need a joint time-frequency representation. In the first part of this dissertation, we first introduce improvements of speech spectral envelope modeling, and then we present a new modeling of speech spectral sequence. Conventional representations of a speech spectral envelope, such as LP coefficients or non-parametric representations, meet difficulties in controlling spectral peaks and spectral shape. In our framework, we explore the Gaussian mixture model parameters proposed by Zolfaghari et al. (called spectral-GMM parameters in this dissertation) to model a speech spectral envelope. In this technique, formants are assumed to be represented by Gaussian distributions, and a speech spectral envelope could be represented by a Gaussian mixture model. Although the original method well models and flexibly controls the speech spectral envelope, it still has two main problems. The first problem is difficulty in modifying a speech spectral envelope in both axes, frequency and amplitude. The second problem is that a Gaussian distribution does not always fit to a formant very well, especially for other kinds of voices which are not reading voices, such as singing voices. To solve these drawbacks, we propose two improvements in the speech spectral envelope modeling. We employ constraints to model a spectral peak by using only one Gaussian component. We also use an asymmetric Gaussian mixture model to model a speech spectral envelope, instead of a Gaussian mixture model. We then develop a new modeling of speech spectral sequence based on temporal decomposition (TD), which originally proposed by Atal, and spectral-GMM parameters. In our modeling, the TD algorithm is utilized to model the temporal evolution (in the time domain), and spectral-GMM parameters are used to model the speech spectral envelope (in the frequency domain). Experimental results show that our modeling models speech signals very well. In addition, our analysis/synthesis method is potential to ensure the smoothness of modified speech, and to perform efficient spectral modification.

To solve the second issue, the insufficient smoothness of modified spectra between frames, one of efficient ways is to control spectral dynamics. In this dissertation, we employ the TD technique, which decomposes speech into event targets and event functions, to control spectral dynamics to improve the quality of synthesized speech. Based on the TD technique, we propose a new method to improve the quality of modified speech signals in concatenative speech synthesis. Concatenative speech synthesis systems form utterances by concatenating pre-recorded speech units. In concatenative speech synthesis systems, output speech is limited by the contents of the pre-recorded databases, and inevitable concatenation errors can lead to audible discontinuities. To reduce the discontinuities between speech units, many methods have been presented in the literature. However, some steps in these methods need to be manually performed, due to preparation of "fusion" units, or extraction of formants. Therefore, it is necessary to have a new method which can automatically smooth the mismatch between speech segments. In our method, we automatically decompose speech units by using the TD technique. The same event functions evaluated for the spectral parameters are also used to describe the temporal patterns of the excitation parameters. The modifications of spectral parameters, F0 and gain information at the joint parts of the speech units are performed by altering the last and the first event and excitation targets of the first and the last speech units, respectively. As a result, the mismatch of spectral, F0, and gain information are reduced at the concatenation points.

To solve the third issue, the ineffective spectral modification, we develop a new efficient algorithm of spectral modification which is applied for rule-based methods, and propose an improvement of GMM-based voice conversion methods.

In our proposed algorithm of spectral modification, we utilize spectral-GMM parameters to model a speech spectral envelope. Spectral-GMM parameters extracted from the spectral envelope are spectral peaks, which may be related to formant information. It is well-know that formant frequencies are some of the most important parameters in characterizing speech, and control of formants can effectively modify the spectral envelope. 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. Therefore, we develop a new algorithm to modify spectral-GMM parameters in accordance with formant frequencies. We then apply our algorithm to two areas, emotional speech synthesis which requires modifications of both formant frequency and power, and voice gender conversion which requires a large amount of spectral modification. Experimental results show the effectiveness of our spectral modification algorithm.

In state-of-the-art voice conversion systems, GMM-based spectral voice conversion methods are regarded as some of the best systems. However, the quality of converted speech is still far from natural. This dissertation presents a new spectral voice conversion method to deal with two drawbacks of the conventional GMM-based spectral voice conversion methods, insufficient precision of GMM parameters and insufficient smoothness of the converted spectra between frames. For improvement of the estimation of GMM parameters, we utilize phoneme-based features of event targets as spectral vectors 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 the converted speech is ensured by the shape of the event functions. Experimental results confirm the high-quality of our converted speech.

In summary, the main contributions of this dissertation are as follows.

1. Propose a new approach of spectral modeling for speech modification. Our method is potential to ensure smoothness of modified speech, and to perform efficient spectral modification.
2. Propose an efficient framework to model and control temporal evolution, which can ensure the naturalness and the smoothness of modified speech. This feature improves the quality of modified speech.
3. Propose a new efficient algorithm of spectral modification. Our algorithm performs spectral modification directly on the speech spectral envelope, which is flexible to modify the speech spectral envelope, and does not produce artifacts.
4. Apply phoneme constraints to the GMM-based spectral voice conversion method to improve the quality of transformation functions, which leads to enhancement of converted speech quality.

Key Words: **spectral modification, voice transformation, temporal decomposition, spectral-GMM**