## Voice modification by source filter model-based modified short-time Fourier transform magnitude

**Abstraction:**

1. **Introduction**

Voice modification is a technique which can be used to change the characteristics of various sound produced by a person. This field of speech technology can contribute greatly to the Entertainment industry as well as increase diversity of the voice database for multiple speaker Text to Speech (TTS) systems. For example, a language learning system may need to reduce speaking rate so that the pronunciation is much clear, or a TTS system can change the original voice pitch so that a synthetical speech uttered by a man is heard as if spoken by a woman.

Voice modifications are usually referred to as prosodic modifications including four main types: time scale modification (TSM), pitch modification (PM), timbre modification (TM) and intensity modification (IM). The greatest challenge in TSM is to change the audio rate, while preserving other characteristics such as pitch and timbre. The goal of PM is to alter the fundamental frequency in order to compress or expand the spacing between the harmonic components in the spectrum while preserving the short time spectral envelope as well as the time evolution. There are circumstances where we want to change the locations and bandwidths of formants while keeping the same pitch, which is defined as timbre modification (TM). IM can be easily achieved by associating an intensity scale factor at each analysis time instant of a signal. Several approaches have been proposed for voice modification. Such approaches include synchronized overlap and add algorithm (SOLA)[1], overlap-add technique based on waveform similarity(WSOLA)[2], phase vocoder method and its refinement[3, 4], peak alignment overlap-add algorithm(PAOLA)[5] , etc.

In the process of changing the pitch of a signal to sharp or flat, either with or without keeping the original audio file length, the sample rate of the audio signal is altered thus changing the fundamental frequency along with all harmonics and spectral envelope. As a result, pitch is changed as well as the locations and bandwidths of formants, which we need to avoid in some applications. Similar cases also happen in TM.

The source filter model (SFM) is a model of speech where the spoken word is comprised of a source component originating from the vocal cords which is then shaped by a filter imitating the effect of the vocal tract. This model of sound production is linear and assumes superposition holds. TSM using STFT has been proposed by Portnoff[6]. Griffin and Lim (GL) developed an algorithm for signal estimation from modified short-time Fourier transform (MSTFT) and modified short-time Fourier transform magnitude (MSTFTM) [7]. In this paper, an improved GL algorithm which is more suitable for real-time environment is adopted to modify the voice. SFM-based MSTFTM for voice modification, including TSM, PM and TM, is proposed in this paper.

The rest of the paper is organized as follows. Section 2 briefly reviews the source filter model. Section 3 introduces an improved MSTFTM. TSM, PM and TM using SFM-based MSTFTM respectively proposed in section 4. Finally, conclusions are drawn in the last section.

1. **Source filter model**

The principle of SFM is shown in fig.1. The source provides the excitation, which is shaped spectrally by 

Fig.1. Source filter model

the vocal tract filter. The key effect of SFM is that it looks a speech signal as two parts: the transfer function of vocal tract filter which contains the vocal quality and the excitation which contains the pitch and the sound.

Linear predictive analysis can help us to put SFM into practice. We can predict that nth sample in a sequence of voice samples is represented by the weighted sum of the p previous samples

 (1)

Where and are real discrete sequences, and (k=1, 2…p) are parameters which can be estimated by Levinson-Durbin algorithm. The term is an estimate of the true value . The number of samples p is referred to as the order of LPC. As p approaches infinity, we should be able to predict the nth sample exactly. However, p is usually on the order of ten to twenty, where it can provide an accurate enough representation with a limited cost of computation. Consequently, we have an error, defined as

 (2)

Then we can take the z-transform of the above equation



 (3)

Thus, we can denote the error signal as the product of original speech signal and the transfer function . Here, is an all-zero digital filter which represents the effect of vocal tract in SFM. Through above analysis, we get the two parts of voice depicted in SFM, i.e., the excitation represented by and the vocal tract filter by .

1. **Signal estimation from improved modified short-time Fourier transform**

A discrete signal can be represented as a STFT sequence. This means we can recover the signal from its original or modified STFT form. However in many applications, we need to recover the time domain from the magnitude spectrum , or a modified version .

GL developed an algorithm to estimate the signal form or by monotonically decreasing the distance measure function which is defined as



 (4)

where is the STFTM of original signal and is the corresponding MSTFTM.

Using in place of , the iterative algorithm results in the following update equation

(5)

where

 (6)

It can be proved by mathematical justification that the algorithm decreases in each iteration the distance [7].

Based on GL iteration strategy, Xinglei, Gerald and Lonce (XGL) proposed a real-time iterative spectrogram inversion (RTISI) algorithm and the RTISI with look-ahead (RTISI-LA) [8]. These refined methods are mainly aimed to improve the real-time performance of classic GL algorithm by employing a GL iteration strategy on the current frame alone, using information from the audio frames already reconstructed that overlap with the current frame to construct an initial current frame phase estimate. However, XGL algorithm is directly imposed on to realize TSM and PM, which of course results the shift of the location and bandwidths of formants.

In this paper, we make use of some of the XGL strategy. Instead of processing the directly, we first utilize the LPC analysis to divide the into two parts, i.e., and , then process the two parts respectively, finally synthesize them back to new voice. To reduce the computational load, we use the standard overlap-add form which is defined in (7)

 (7)

to replace iterative formula (5).

1. **TSM, PM and TM using SFM-based MSTFTM**

In this section, we impose the SFM-based improved MSTFTM strategy on voice modification. Under the main technical framework of SFM and MSTFTM, three types of voice modifications including TSM, PM and TM are combined together. The procedures are showed in Fig.2.



Fig.2. Procedures of voice modification. Solid boxes illustrate common processes mainly including LPC analysis and voice estimation using MSTFTM. Dashed boxes stand for the differences in TSM, PM and TM.

Solid boxes represent common processing procedures of all kinds of these voice modifications. Dashed boxes depict the different processing techniques which is optional according to each voice modification. For example, we can regulate the rate to accomplish TSM, adjust the resampling rate, and to realize PM, warp the to implement TM. Certainly, TSM, PM and TM can be integrated in one modification to achieve new voice features.

1. **Time Scale Modification (TSM)**

The process of TSM is shown in Fig.3. By altering the rate of , we can change the speed of the voice while have no influence on pitch.

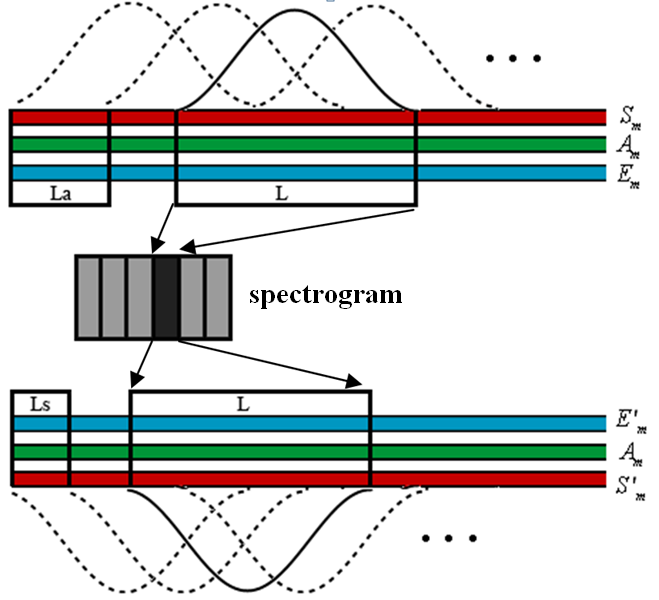
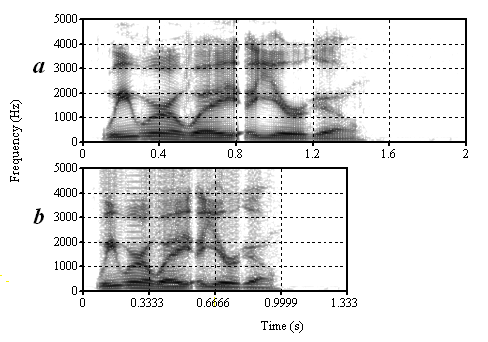
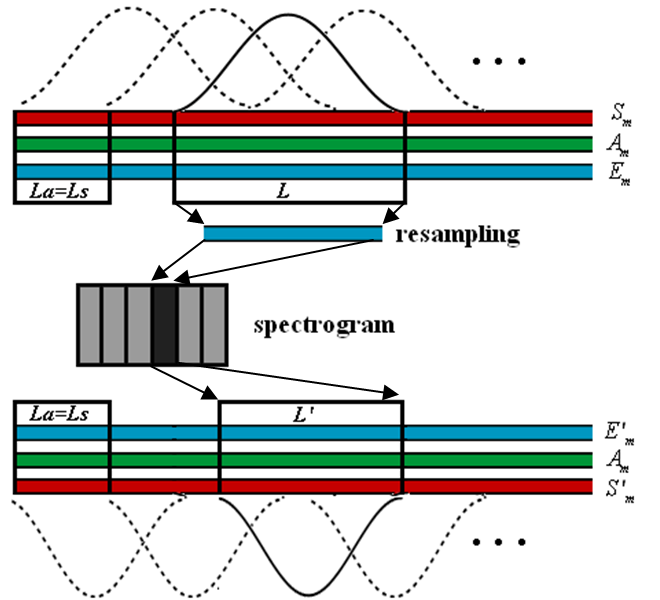
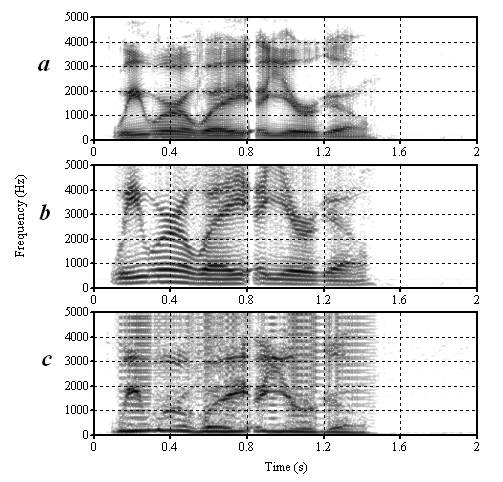


Fig.3. TSM using SFM-based MSTFTM.

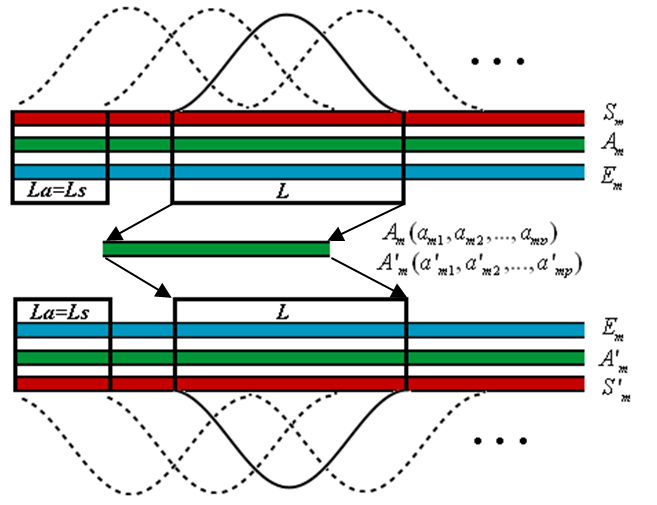


1. **Pitch Modification (PM)**





1. **Timbre Modification (TM)**



**References:**

[1]. Wayman, J.L., R.E. Reinke and D.L. Wilson. High quality speech expansion, compression, and noise filtering using the sola method of time scale modification. in Twenty-Third Asilomar Conference on Signals, Systems and Computers, 1989. 1989.

[2]. Verhelst, W. and M. Roelands. An overlap-add technique based on waveform similarity (WSOLA) for

high quality time-scale modification of speech. in IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993. ICASSP-93. 1993.

[3]. Dolson, M., The Phase Vocoder: A Tutorial. Computer Music Journal, 1986. 10(4): p. 14-27.

[4]. Laroche, J. and M. Dolson, Improved phase vocoder time-scale modification of audio. IEEE Transactions on Speech and Audio Processing, 1999. 7(3): p. 323-332.

[5]. Dorran, D., R. Lawlor and E. Coyle. High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PAOLA). in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. (ICASSP '03). 2003.

[6]. Portnoff, M., Time-scale modification of speech based on short-time Fourier analysis. Acoustics, Speech and Signal Processing, IEEE Transactions on

Acoustics, Speech and Signal Processing, IEEE Transactions on

DOI - 10.1109/TASSP.1981.1163581

Acoustics, Speech and Signal Processing, IEEE Transactions on, 1981. 29(3): p. 374- 390.

[7]. Griffin, D. and L. Jae, Signal estimation from modified short-time Fourier transform. Acoustics, Speech and Signal Processing, IEEE Transactions on

Acoustics, Speech and Signal Processing, IEEE Transactions on

DOI - 10.1109/TASSP.1984.1164317

Acoustics, Speech and Signal Processing, IEEE Transactions on, 1984. 32(2): p. 236- 243.

[8]. Xinglei, Z., G. Beauregard and L. Wyse, Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra. Audio, Speech, and Language Processing, IEEE Transactions on

Audio, Speech, and Language Processing, IEEE Transactions on

DOI - 10.1109/TASL.2007.899236

Audio, Speech, and Language Processing, IEEE Transactions on, 2007. 15(5): p. 1645-1653.