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

Abstract-The key issue in a specific voice modification is how to change target parameters while keeping other parameters constant. For time-scale modification (TSM), pitch modification (PM) and timbre modification (TM), the target parameters are respectively the tempo, pitch frequency and location and bandwidth of formants. This paper proposes a new method for voice modification by modified short-time Fourier transform magnitude (MSTFTM) based on source filter model (SFM). Different types of voice modification experiments using proposed method are performed. The results show that using SFM-based MSTFTM, a pretty good performance is achieved. Comparisons with another classic MSTFTM method in PM have verified that the method can modify the pitch without affecting the formants.

Keywords- voice modification, time scale modification (TSM), pitch modification (PM), timbre modification (TM), signal estimation

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 will sound 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 change 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. The aim of timbre modification (TM) is to change the locations and bandwidths of formants while keeping the same pitch. 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. However, when in PM, above methods change the formants of a voice. PM and TM are mixed together. 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 voice 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 voice production is linear and assumes superposition holds. TSM using short time Fourier Transform (STFT) has been proposed by Portnoff[6]. Griffin and Lim developed an algorithm for signal estimation from modified short-time Fourier transform (MSTFT) and modified short-time Fourier transform magnitude (MSTFTM) [7]. Xinglei, et al improved the real-time performance of Griffin and Lim's method [8].

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.

2. Source filter model

The principle of SFM is shown in fig.1. The source provides the excitation, which is shaped spectrally by the vocal tract filter. The key effect of SFM is that it considers a voice 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.

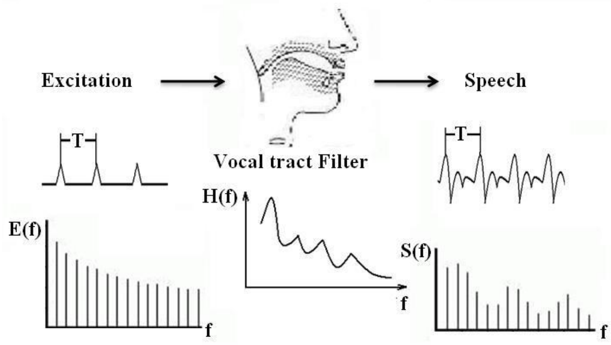


Fig.1. Source filter model

Linear predictive analysis (LPA) is a powerful voice analysis technique which can be used to put SFM into practice. LPA predicts that nth sample in a sequence of voice samples is approximated by the weighted sum of the p previous samples

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \qquad (1)$$

Where $\hat{s}(n)$ and $s(n)$ are real discrete sequences, and a_k (k=1, 2…p) are parameters which can be estimated by Levinson-Durbin algorithm. The term $\hat{s}(n)$ is an estimate of the true value $s(n)$. The number of samples p is referred to as the order of LPA. As p approaches infinity, we are 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

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad (2)$$

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

$$E(z) = S(z) - \sum_{k=1}^{p} a_k S(z) z^{-k}$$

$$= (1 - \sum_{k=1}^{p} a_k z^{-k}) S(z) = A(z)S(z) \qquad (3)$$

Thus, we can denote the error signal $E(z)$ as the product of original speech signal $S(z)$ and the transfer function $A(z)$. Here, $A(z)$ 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 $E(z)$ and the vocal tract filter by $A(z)$.

3. Signal estimation from improved modified short-time Fourier transform

A discrete signal x(n) 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 $|X(mS,\omega)|$, or a modified version $|X'(mS,\omega)|$.

Griffin and Lim developed an algorithm to estimate the signal form $|X(mS,\omega)|$ or $|X'(mS,\omega)|$ by monotonically decreasing the distance measure function $D_M[x(n), x'(n)]$ which is defined as

$$D_M[x(n), x'(n)] = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi}$$

$$\int_{-\pi}^{\pi} [|X(mS,\omega)| - |X^{'}(mS,\omega)|]^2 d\omega \qquad (4)$$

where $|X(mS,\omega)|$ is the STFTM of original signal $x(n)$ and $|X'(mS,\omega)|$ is the corresponding MSTFTM.

Using $|X^i(mS,\omega)|$ in place of $|X'(mS,\omega)|$, the iterative algorithm results in the following update equation

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n) \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}^i(mS,\omega) e^{j\omega n} d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS-n)} \qquad (5)$$

where

$$\hat{X}^i(mS,\omega) = X^i(mS,\omega) \frac{|X(mS,\omega)|}{|X^i(mS,\omega)|} \qquad (6)$$

It can be proved by mathematical justification that the algorithm decreases in each iteration the distance $D_M[x(n), x^i(n)]$[7].

Based on Griffin and Lim's method, Xinglei et al 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 Griffin and Lim's algorithm by employing a Griffin and Lim's 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, Xinglei's algorithm is directly imposed on $S(z)$ to realize TSM and PM, which of course results the shift of the location and bandwidths of formants.

Instead of processing the $S(z)$ directly, we firstly utilize the LPA to divide the S(z) into two parts, i.e., E(z) and A(z), 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)

$$x^{i+1}(n)=\frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{-\pi}^{\pi}\hat{X}^{i}(mS,\omega)e^{j\omega n}d\omega}{\sum_{m=-\infty}^{\infty}w(mS-n)} \qquad (7)$$

to replace iterative formula (5).

4. 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.

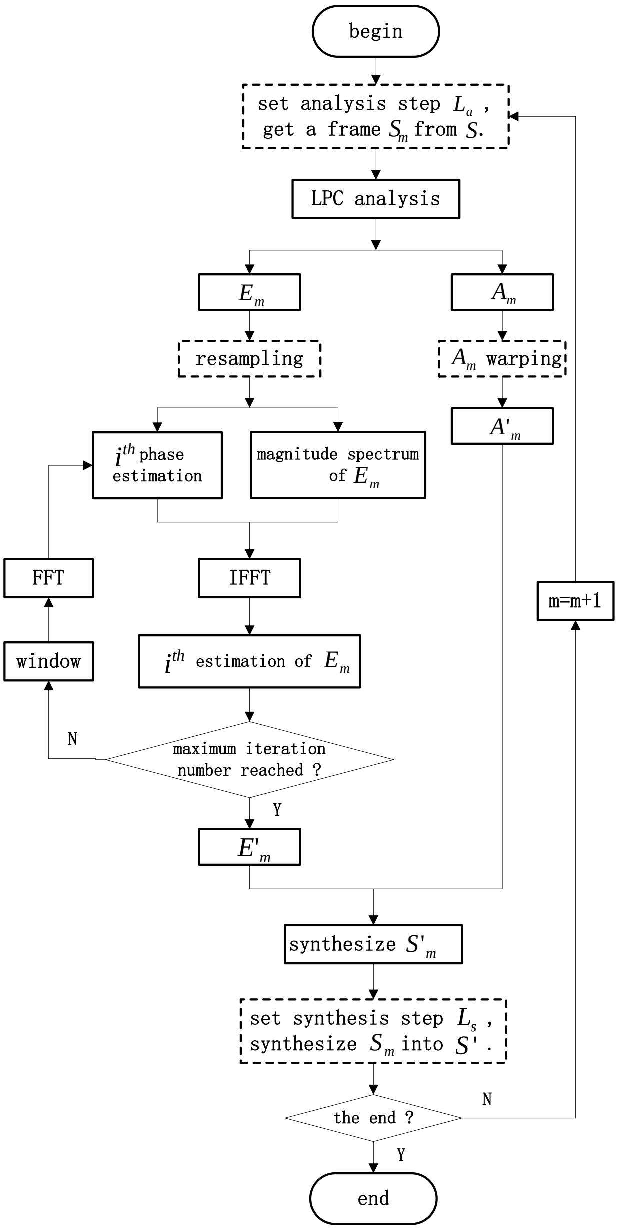

begin
set analysis step L_a ,
get a frame S_m from S.
LPC analysis
E_m
A_m
resampling
A_m warping
A'_m
i^{th} phase estimation
magnitude spectrum of E_m
FFT
IFFT
m=m+1
window
i^{th} estimation of E_m
N
maximum iteration number reached ?
Y
E'_m
synthesize S'_m
set synthesis step L_s ,
synthesize S_m into S' .
the end ?
N
Y
end

Fig.2. Procedures of voice modification. Solid boxes illustrate common processes mainly including LPA 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 L_a/L_s to accomplish TSM, adjust the resampling rate, L_a and L_s to realize PM, warp the A_m to implement TM. Certainly, TSM, PM and TM can be integrated in one modification to achieve new voice features.

A. Time Scale Modification (TSM)

The process of TSM is shown in Fig.3. By changing the rate of L_a/L_s , we can change the speed of the voice while have no influence on its pitch. If $L_a/L_s > 1$, voice speeds up, else if $L_a/L_s < 1$, voice slows down.

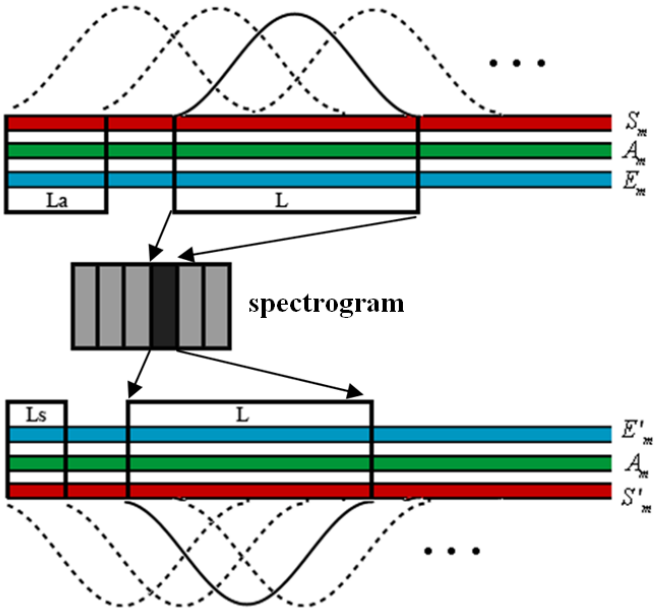


Fig.3. Process of TSM using SFM-based MSTFTM. It is shown in this figure that $L_a > L_s$, which generates a more rapid voice.

The spectrogram of the sentence “We were away a year ago.” is shown in Fig.4a. The result of TSM using SFM-based MSTFTM is shown in Fig.4b. The pitch, the location and bandwidths of formants are successfully kept. The tempo of the modified voice is 1.5 times faster than that of the original voice.

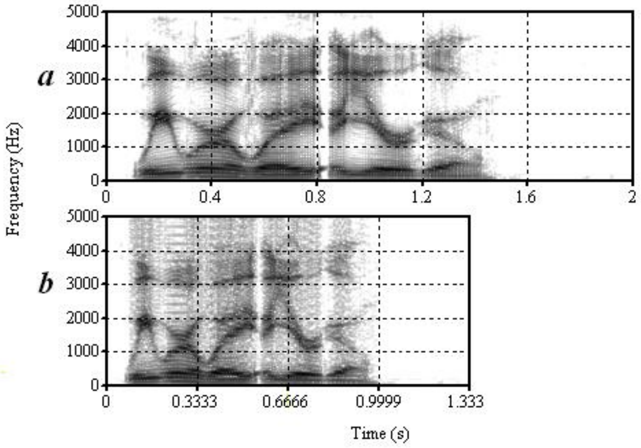


Fig.4. (a) The spectrogram of the sentence “We were away a year ago.” (b) The spectrogram of TSM using SFM-based MSTFTM, the modified voice is 1.5 times faster than the original voice.

B. Pitch Modification (PM)

The process of PM is shown in Fig.5. The analysis size L_a is equal to the synthesis size L_s. First, a frame of signal is extracted from the original voice. Then, LPA is used to divide the frame into two parts. Next, E_m is resampled and estimated through improved MSTFTM introduced in section 3. Finally, a re-synthesis processing is implemented. The rate L/L' defines the new pitch of the modified voice. $L > L'$ results to a voice with higher pitch. Inversely $L < L'$ results to a voice with lower pitch.

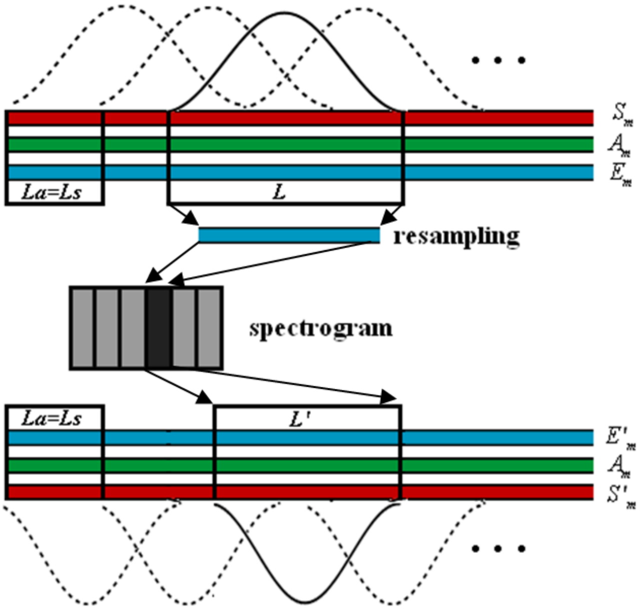


Fig.5. Process of PM using SFM-based MSTFTM. $L_a = L_s$ means the size of the audio file will keep the same when PM finished. It is shown in this figure $L > L'$, which generates a voice with higher pitch.

The spectrogram of the sentence “We were away a year ago.” is shown in Fig.6a. The result of PM using Griffin and Lim’s MSTFTM and SFM-based MSTFTM is respectively shown in Fig.6b and Fig.6c. It is clear that pitch of Fig.6b and Fig.6c is twice higher than that of Fig.6a. However, in Fig.6b, the location and bandwidth of formants shifts along with the pitch. In Fig.6c, formants are kept in substance the same.

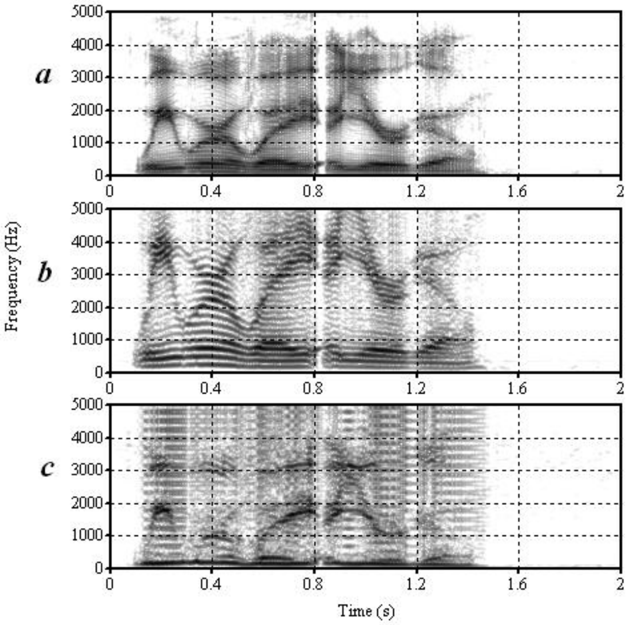
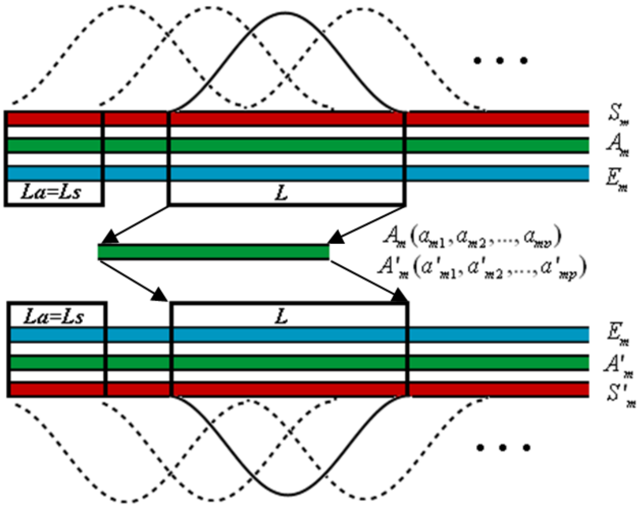


Fig.6. (a) The spectrogram of the sentence “We were away a year ago.” (b) The spectrogram of PM using Griffin and Lim’s MSTFTM. (c) The spectrogram of PM using SFM-based MSTFTM. Pitch of (b) and (c) is twice higher than that of (a).

C. Timbre Modification (TM)

The process of PM is shown in Fig.7. The analysis size L_a is equal to the synthesis size L_s. The Length of analysis window L is also equal to that of synthesis window L'. The vocal tract filter A_m is warped before re-synthesis process.



• • •
S_m
A_m
E_m
La=Ls
L
$A_m(a_{m1}, a_{m2}, ..., a_{mp})$
$A'_m(a'_{m1}, a'_{m2}, ..., a'_{mp})$
La=Ls
L
E_m
A'_m
S'_m
• • •

Fig.7. Process of TM using SFM-based MSTFTM. $L_a = L_s$ means the size of the audio file will keep the same when TM finished.

The spectrogram of the sentence “We were away a year ago.” is shown in Fig.8a. The result of TM using SFM-based MSTFTM is shown in Fig.8b. The location of the formants are moved to higher frequency, however, the tempo and the pitch of the voice are kept the same.

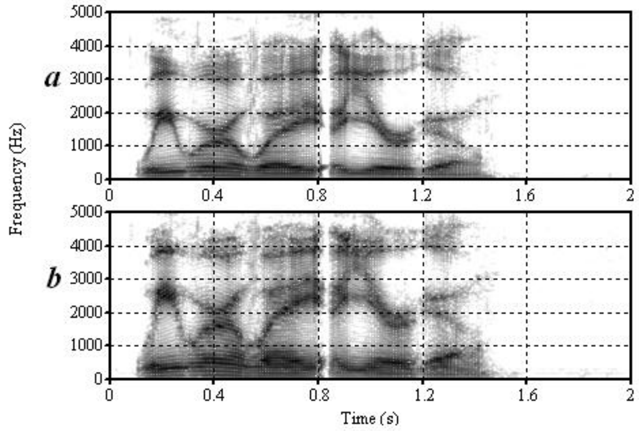


Fig.8. (a) The spectrogram of the sentence “We were away a year ago.” (b) The spectrogram of TM using SFM-based MSTFTM.

5. Conclusion

This paper has proposed a new SFM-based MSTFTM algorithm for voice modification. SFM is used to divide the original signal into two parts, i.e., excitation and vocal tract filter. Improved MSTFTM is used to estimate signal. The combination of SFM and MSTFTM estimation is feasible and flexible to modify voices. When doing PM, Griffin and Lim’s algorithm and its refinement cannot change the pitch without shifting the location and bandwidth of formants, while SFM-based MSTFTM method overcome these difficulties. SFM-based MSTFTM algorithm separates the excitation and vocal tract filter, which makes the control of the parameters of pitch and formants more feasible. TSM, PM and TM can be assembled to synthesize a voice with new features.

References

[1] J. L. Wayman, R. E. Reinke and D. L. Wilson, "High quality speech expansion, compression, and noise filtering using the sola method of time scale modification," in *Proc. 1989 Twenty-Third Asilomar Conference on Signals, Systems and Computers, 1989.*, pp. 714-717.

[2] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993. ICASSP-93.*, pp. 554-557.

[3] M. Dolson, "The Phase Vocoder: A Tutorial," *Computer Music Journal*, vol.10, pp. 14-27, 1986.

[4] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol.7, pp. 323-332, 1999.

[5] D. Dorran, R. Lawlor and E. Coyle, "High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PAOLA)," in *Proc. 2003 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. (ICASSP '03).*, pp. 700-703.

[6] M. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.29, pp. 374-390, 1981.

[7] D. Griffin and L. Jae, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.32, pp. 236-243, 1984.

[8] Z. Xinglei, G. Beauregard and L. Wyse, "Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.15, pp. 1645-1653, 2007.