## 基于源滤波模型短时傅立叶变换幅度谱的声音变换

**摘 要：**声音变换的一个重要评价标准是在改变声音的目标参数的时候如何保持其他参数的恒定。比如，对声音节奏变换（Time Scale Modification, 简称TSM）、声音基频变换（Pitch Modification, 简称PM）和声音音色变换（Timbre Modification, 简称TM）而言，目标参数分别是声音节奏、基音频率和声音共振峰的位置和带宽。本文提出了一种基于源滤波模型（Source Filter Model, 简称SFM）短时傅立叶变换幅度谱（Short-time Fourier Transform Magnitude, 简称STFTM）的声音变换方法。实验表明，利用本文提出的方法能有效地将基音参数和共振峰参数分开调整，取得很好的声音变换效果。

**关键词**：声音变换，声音节奏，声音基频变换，声音音色变换，信号重构

**Abstract:** The key evaluation criterion 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 method for voice modification by short-time Fourier transform magnitude (STFTM) based on source filter model (SFM). Experiments using the proposed method show that pitch parameters and formants parameters are successfully separated. A pretty good performance is achieved.

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

1. **引言**

声音变换是一种用来改变声音特点的技术。这种技术广泛应用于娱乐产业以及用于增加声音合成数据库的多样性。例如，一个语言教学系统需要改变语音播讲的速度以使得发音更加地清楚；一个语音合成系统（TTS）需要改变原始语音数据库中的声音以使得一个男生听起来像一个女声，从而增加语音数据库的多样性。

声音变换一般包括四种类型：声音节奏变换，声音频率变换，声音音色变换，声音强度变换。声音节奏变换的难点在于如何只改变声音的播放速度而不改变基音频率和音色。声音基音频率变换的目标是压缩或扩展声音各次谐波间的空间距离而保持短时频谱包络以及声音节奏。声音音色变换的目标在于改变声音共振峰的位置和带宽的同时保持声音的节奏和基音频率。声音轻度变换可以通过简单地把信号乘上一个强度因子来得到。研究人员提出了一系列声音变换的方法。如同步叠加法（SOLA）[1]，波形相似叠加法（WSOLA）[2]，声码器以及各种改进方法[3-4]，峰值对齐叠加法（PAOLA）[5]。然而上述方法在改变声音基音频率的同时改变了声音共振峰的位置和带宽，反之亦然。结果使得你在改变音调的时候却改变了音色，从而导致一个男声听起来如同女声。声音的个性特点遭到了破坏。Portnoff提出了用短时傅立叶变换来进行声音节奏变换[6]。Griffin等人提出了利用修改短时傅立叶幅度谱来重构信号的方法处理声音节奏变换和声音基频变换[7]。在Griffin的基础上Xinglei等人又提出了实时语谱迭代转换法（RTISI）和超前实时语谱迭代转换法（RTISI-LA）[8]。这些算法在信号处理的实时性方面有了较大改善，但依然没有把声音基音频率参数和声音共振峰参数区分开来。导致在进行声音基音频率变换的时候改变的原始声音载有的个性特点，在进行声音音色变换时声音的音调也发生了改变。

为了解决上述存在的问题，本文引入了源滤波模型，将声音信号的基音频率参数和共振峰参数区分开来。原始的声音信号被分解成声音激励信号和声道滤波器信号。然后再分别通过在频域修改声音激励信号的频谱利用短时傅立叶幅度谱重构信号的方法来重构激励信号和修改滤波器参数的方法来修改声音的基音频率参数和共振峰参数。最后再将修改后的声音激励信号和声道滤波器信号重新合成回新的声音信号。

1. **源滤波模型**

源滤波模型认为声音信号是由声带振动产生的激励信号经过声道滤波产生的。因此，声音信号可以被分解成激励信号和声道滤波器两部分。激励信号携带了声音的基音频率，其大小决定这音调的高低。滤波器幅度谱的峰值被称为共振峰，其位置和带宽影响着声音的音色。下图显示了源滤波模型的原理。



图1 源滤波模型

源滤波模型的思想可以通过倒谱分析或是线性预测分析来实现。本文采用线性预测分析来分解语音信号。假设代表一段离散语音序列，n=1.2.3..为序列号。可认为是由之前p个信号的加权值来预测。记的预测值为

 (1)

其中和均为离散实数序列，加权参数 (k=1, 2…p)可由莱文森-杜宾算法求解得到。p为线性预测分析的阶数。理论上，当p趋近正无穷时，预测语音无线接近原始语音信号。通常情况下，p取10到12。相应的，原始声音信号和其预测信号的误差定义为

 (2)

对（2）式两边取z变换得到



 (3)

于是在z域里头，误差信号由原始语音信号和传递函数得到。这里的是一个全零点数字滤波器，代表了声道滤波器。通过调整零点，就可以调节声音共振峰的频率和带宽。误差信号携带了声音的基音频率信息。

1. **基于短时傅立叶变换幅度谱的信号重构**

短时离散傅立叶变换能够将一个离散时间信号转化到频率域得到频谱信号。于是，我们可以在频率域对进行修改，然后再将修改后的频谱信号反变换回时域得到修改后的声音信号。然而频谱包括幅度谱和相位谱。而相位在实际操作中很不方便，所以很多时候我们需要利用幅度谱来重构新的信号。也就是说我们需要首先将时域信号转换到频域得到频域的幅度谱，然后根据需要对幅度谱进行修改得到，最后再利用修改后的幅度谱来重构信号。

Griffin等人提出了一种算法[7]从信号的幅度谱来重构信号。定义原始信号和目标信号

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]. However, above approaches change the formants of a voice as PM is performed. 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. 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]. 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 voice signals to realize PM, which of course results the shift of the location and bandwidths of formants. When the location and bandwidths of formants are changed, features of the voice are also changed.

In order to solve above problems, we combine the MSTFTM with source filter model (SFM). 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. 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. We first use SFM to divide the voice signals into excitation part and vocal tract part then process the two parts respectively.

1. **Signal estimation from 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 .

Griffin and Lim proposed 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 , Griffin and Lim gave the following update equation

(5)

where

 (6)

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

To reduce the computational load, we use the standard overlap-add form which is defined in (7)

 (7)

to replace iterative formula (5).

1. **Source filter model based voice modification**

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



Fig.1. Source filter model

Linear predictive analysis (LPA) is a powerful voice analysis technique which can be used to put SFM into practice. denotes the nth sample in a sequence of voice samples. LPA predicts that is approximated 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. is an estimate of the true value . 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

 (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. We get the two parts of voice depicted in SFM, i.e., the excitation represented by and the vocal tract filter by .

When implementing voice modification, we use excitation signal to replace in formula (4), (6), (7) to get modified excitation , modify the vocal tract to . Finally we filter by to get modified voice signal .

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

In this section, we use the SFM-based MSTFTM strategy to implement 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.

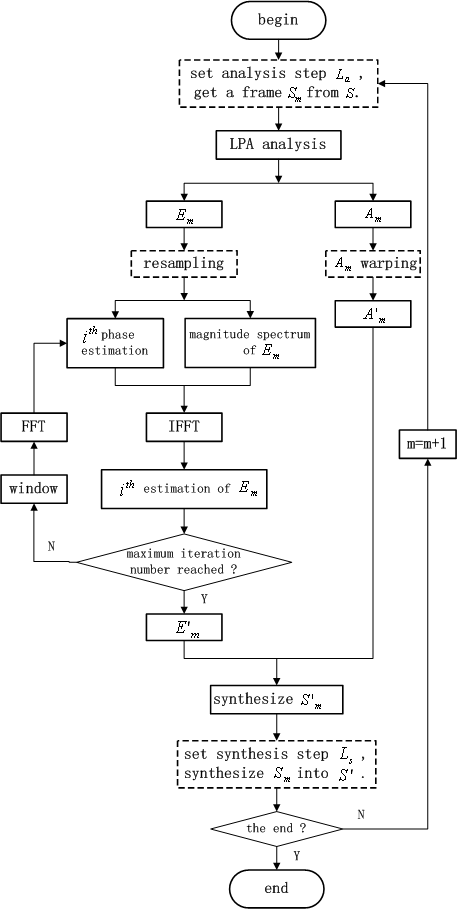
****

Fig.2. Procedures of voice modification. Solid-line blocks illustrate common processes mainly including LPA and voice estimation using MSTFTM. Dashed-line blocks stand for the differences in TSM, PM and TM.

Solid-line blocks represent common processing procedures of all kinds of these voice modifications. Dashed-line blocks 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 changing the rate of , we can change the speed of the voice while have no influence on its pitch. If , voice speeds up, else if , voice slows down.

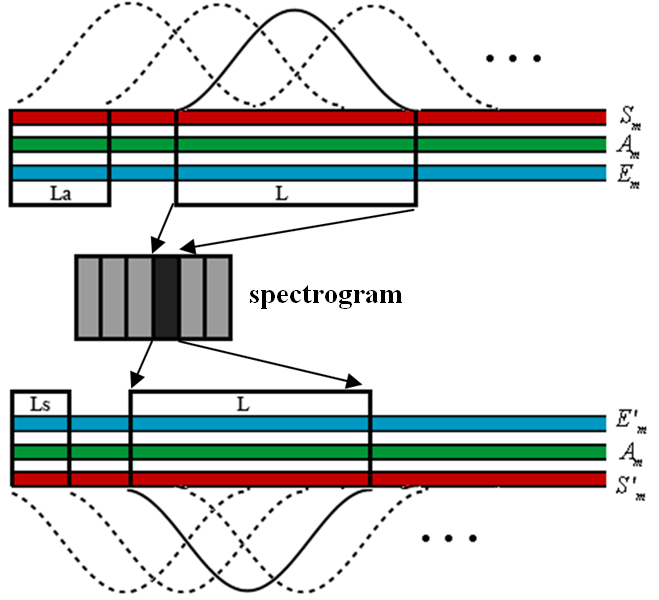


Fig.3. Process of TSM using SFM-based MSTFTM. , a more rapid voice will be produced.

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.

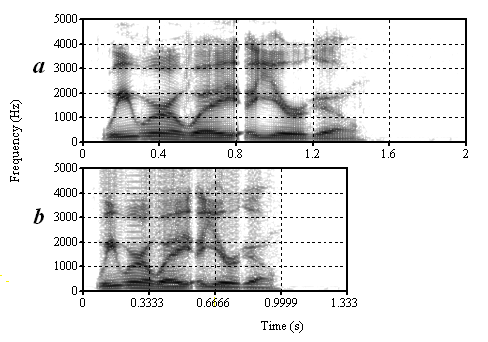


Fig.4. (a) The spectrogram of the sentence “We were away a year ago.” (b) The spectrogram of TSM using SFM-based MSTFTM, the modified voice is 1.5 times faster than the original voice.

1. **Pitch Modification (PM)**

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

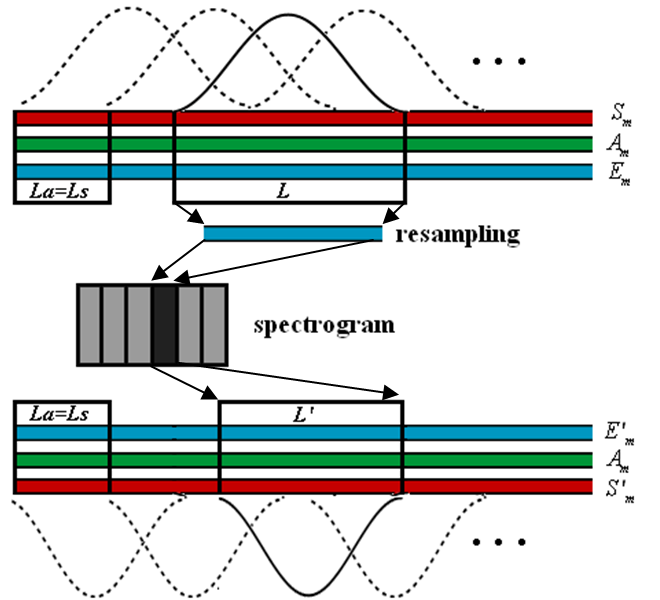


Fig.5. Process of PM using SFM-based MSTFTM. means the size of the audio file will keep the same when PM finished. , a voice with higher pitch will be produced.

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.

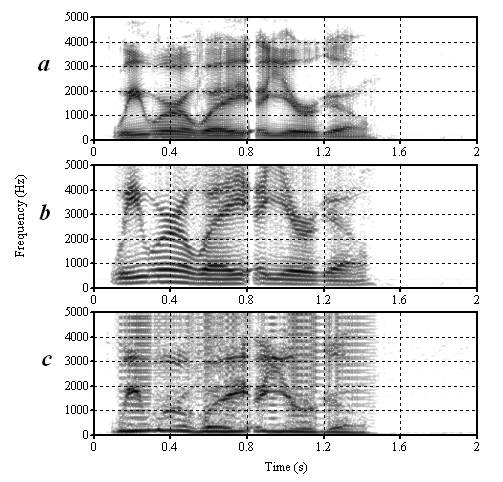


Fig.6. (a) The spectrogram of the sentence “We were away a year ago.” (b) The spectrogram of PM using Griffin and Lim’s MSTFTM. (c) The spectrogram of PM using SFM-based MSTFTM. Pitch of (b) and (c) is twice higher than that of (a).

1. **Timbre Modification (TM)**

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

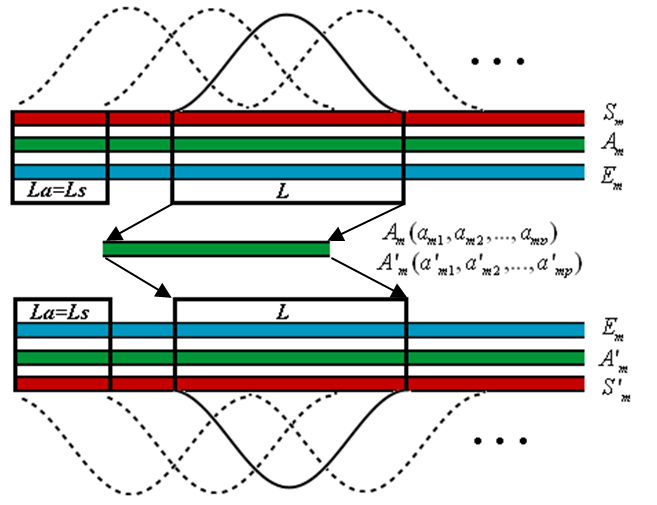


Fig.7. Process of TM using SFM-based MSTFTM. means the size of the audio file will keep the same when TM finished.

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.

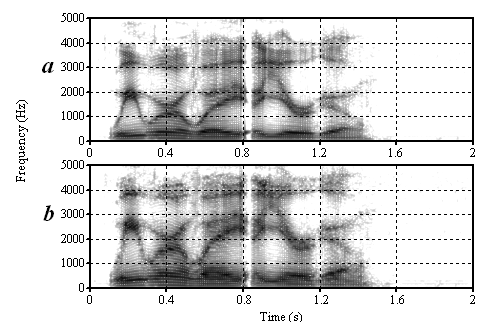


Fig.8. (a) The spectrogram of the sentence “We were away a year ago.” (b) The spectrogram of TM using SFM-based MSTFTM.

1. **Conclusion**

This paper has proposed a SFM-based MSTFTM algorithm for voice modification. The combination of SFM and MSTFTM estimation is feasible and flexible to modify voices. On the one hand, SFM-based MSTFTM method can effectively change the pitch without shifting the location and bandwidth of formants, which can maintain the personality characteristics of original voice when modifying the pitch of the voice. On the other hand, it can change the location and bandwidth of formants without changing the pitch, which can maintain the pitch when modifying the personality characteristics of a voice. 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.