用于读书机器人的声音变换方法

邓杰，赵群飞

(上海交通大学 图像处理与模式识别研究室,上海 200240)

摘 要: 自动读书机器人主要通过安装语音库来实现自动朗读，朗读的声音比较单调。若希望增加朗读声音的多样性就需要安装不同的语音库。这使得语音库容量过大，浪费了存储资源。本文提出了一种声音变换技术，通过对单一语音库中的声音进行变换来实现朗读声音的多样性。重点解决了声音变换中音调变换和音色变换相互干扰的问题。首先，原始声音信号被分解成声音激励信号和声道滤波器信号。然后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激励信号的方法以及通过修改声道滤波器参数的方法来分别进行音调变换和音色变换。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。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，避免了音调变换和音色变换的混叠，取得了良好的声音变换效果。

关键词: 自动读书机器人; 声音变换; 节奏变换; 音调变换; 音色变换
中图法分类号: TP391.41 **文献标识码:** A

Voice modification for reading robot

DENG Jie, ZHAO Qun-fei

(*Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai 200240 , China*)

Abstract: Automatic reading machines implement reading mainly through installing speech library. However, the voice is rather monotonous. Different speech libraries are needed to increase the diversity of the reading voice, which increases the capacity of the whole speech library, as a result, wastes the storage resources. In order to increase the diversity of the reading voice without increasing the size of speech library, a voice modification technology is proposed. The focus is on how to solve the mutual interference between pitch modification and timbre modification. Firstly, the voice is broken down into excitation and vocal tract filter. Secondly, excitation reconstruction through short-time Fourier transform magnitude and parameters modification of vocal tract filter are respectively used to achieve pitch modification and timbre modification. Finally, modified excitation and modified vocal tract filter are synthesized back to voice signals. Experiments using the proposed method show that pitch parameters and formants parameters are successfully separated. Modification interference between pitch and timbre is eliminated. A pretty good performance is achieved.

Key words: automatic reading robot; voice modification; tempo modification; pitch modification; timbre modification

0 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清楚；一个拥有语音合成系统的机器人需要改变原始语音数据库中的声音以使得一个男声听起来像一个女声，从而在不增加语音数据库大小的情况下增加语音数据库的多样性。

声音变换包括四种类型：节奏变换、音调变换、音色变换以及强度变换。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音调和音色。声音音调变换的难点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。声音音色变换的难点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。声音强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。例如，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节奏变换和音调变换[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，但依然没有把声音的基音频率参数和共振峰参数区分开来。导致在进行音调变换的时候改变了原始声音载有的个性特点，在进行音色变换时，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。从而避免了音调变换和音色变换相互干扰的情况。大大增加了机器人朗读声音的多样性。

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，其大小决定着音调的高低。声道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。图 1 显示了源滤波模型的原理。

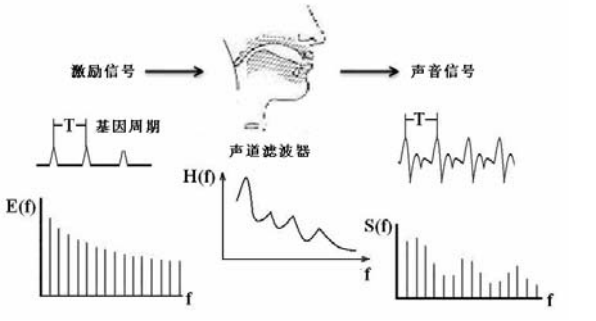

图 1 源滤波模型原理
Fig.1 Principle of source filter model

源滤波模型的思想可以通过倒谱分析或线性预测分析来实现。本文采用线性预测分析来分解声音信号。假设 $s(n)$ 代表一段离散声音信号序列，n=1,2,3…为序列号。根据线性预测的原理， $s(n)$ 可由之前的 p 个信号的加权值来预测。则 $s(n)$ 的预测值 $\hat{s}(n)$ 记为

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

其中 $\hat{s}(n)$ 和 $s(n)$ 均为离散实数序列，加权参数 a_k (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，预测声音信号无限接近原始声音信号。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

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

对(2)式两边取 z 变换得到

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

于是在 z 域里，误差信号由原始声音信号 $S(z)$ 和传递函数 $A(z)$ 相乘得到。这里的 $A(z)$ 是一个全零点数字滤波器，代表了声道滤波器。通过调整 $A(z)$ 的零点，就可以调节声音共振峰的频率和带宽。误差信号 $E(z)$ 则携带了声音的基音频率信息。

2 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 $x(n)$ 转化到频率域得到频谱信号 $X(mS,\omega)$ 。于是，我们可以在频率域对 $X(mS,\omega)$ 进行修改，然后再将修改后的频谱信号 $X'(mS,\omega)$ 反变换回时域得到修改后的时间信号。然而频谱参数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直接利用幅度谱来重构信号。也就是说，需要首先将时域信号转换到频域得到频域的幅度谱 $|X(mS,\omega)|$ ，然后根据需要对幅度谱进行修改得到 $|X'(mS,\omega)|$ ，最后再利用修改后的幅度谱 $|X'(mS,\omega)|$ 来重构信号。

Griffin 等人提出了一种算法[7]从修改过的信号幅度谱 $|X'(mS,\omega)|$ 来重构信号，使重构的信号的短时傅里叶变换幅度谱尽可能地接近目标信号的短时傅里叶变换幅度谱。定义目标信号 $x(n)$ 与重构信号 $x'(n)$ 相似性距离 $D_M[x(n),x'(n)]$ 如(4)式所示

$$D_M[x(n),x'(n)]=\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{-\pi}^{\pi}[|X(mS,\omega)|-|X'(mS,\omega)|]^2 d\omega \tag{4}$$

其中 $|X(mS,\omega)|$ 是目标信号的短时傅里叶变换幅度谱，$|X'(mS,\omega)|$ 为重构信号的短时傅里叶变换幅度谱。利用 $|X^i(mS,\omega)|$ 取代 $|X'(mS,\omega)|$，Griffin 等给出了如下迭代方程

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

这里有

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

其中 $|X^i(mS,\omega)|$ 为 $x^i(n)$ 的短时傅里叶变换幅度谱。从数学上可以证明距离测度 $D_M[x(n),x^i(n)]$ 会随着迭代次数的增加而逐步减小[7]。一般情况下迭代次数取 4 到 5 就能使得距离测度减小到可以接受的程度。在本文的声音变换算法中，为了进一步减轻计算负担，提高运算的实时性。我们采用了标准叠加公式(7)来取代 Griffin 的迭代公式(5)。

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

3 声音变换原理及实验结果

图 2 显示了声音变换的原理，将节奏变换，音调变换，音色变换统一到一个流程当中。

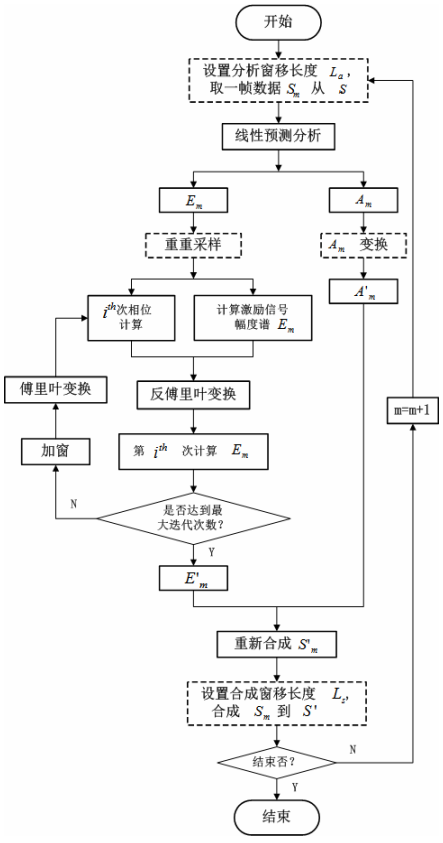

图 2 声音变换流程图
Fig.2 Chart of voice modification

图 2 中实线框表示节奏变换、音调变换和音色变换都需要进行的处理步骤。虚线框表示三种变换在这些地方的处理有所不同。声音信号首先通过线性预测分析得到激励信号和声道滤波器参数。然后根据不同变换的需要利用短时傅里叶变换幅度重构信号法修改激励信号，并修改声道滤波器参数。最后将修改的激励信号和声道滤波器重新合成回时域信号就得到了变换后的声音信号。通过调整 L_a/L_s 的比值可实现声音节奏变换，通过调整重采样率可实现音调变换，通过调整 A_m 的零点参数可实现音色变换。当然也可以同时调整这些参数以实现特定的混合变换效果。

3.1 节奏变换

节奏变换的原理如图 3 所示。其中 L_a 为声音分析时的窗移长度，L_s 为声音合成时的窗移长度，L 为窗口长度。本文中的窗函数取汉明窗。通过改变 L_a/L_s 的比值来调节声音的节奏。当 $L_a/L_s>1$ 可以加快原声音的节奏，$L_a/L_s<1$ 则放慢原声音的节奏。

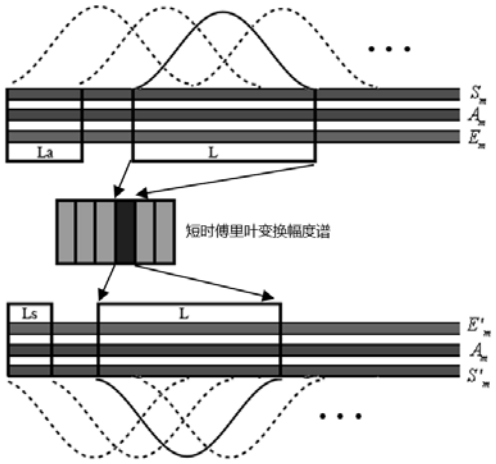

图 3 节奏变换原理图
Fig.3 Principle of tempo modification

图 4a 显示了英文句子“We were away a year ago.”的短时傅里叶变换幅度谱，图 4b 是用本文提出的方法进行节奏加快处理后的声音的短时傅里叶变换幅度谱。从图 4 中可以看出，声音的基音频率、共振峰的位置和带宽都几乎没有改变，因此声音的音调和音色得到了很好的保持。原始语音的节奏则加快了 1.5 倍。

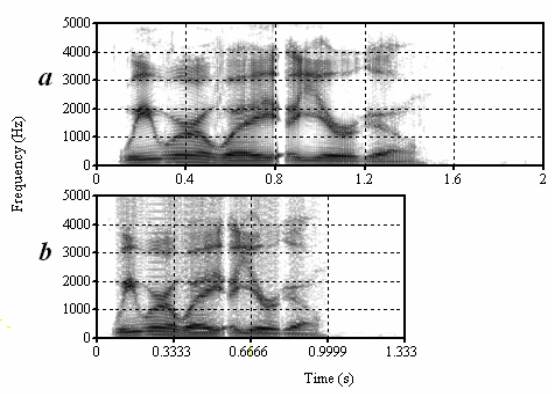

图 4 节奏变换实验结果
Fig.4 Experiment result of tempo modification

3.2 音调变换

音调变换原理如图 5 所示。其中 L_a 为声音分析时的窗移长度，L_s 为声音合成时的窗移长度，L 为声音分析时的窗口长度，L' 为声音合成时的窗口长度。声音信号首先通过线性预测分析被分解成激励信号和声道滤波器。然后对激励信号进行重采样后变换到频率域进行处理。处理完成后再将其通过声道滤波器滤波生成目标声音。在音调变换时，L_a 等于 L_s，通过改变 L/L' 的比值来调节音调。当 $L>L'$ 可以提高原声音的音调，而 $L<L'$ 则降低原声音的音调。

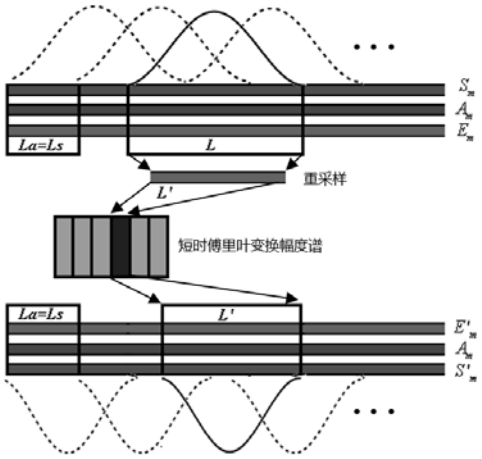

图 5 声调变换原理图
Fig.5 Principle of pitch modification

图 6a 显示了英文句子“We were away a year ago.”的短时傅里叶变换幅度谱，图 6b 是利用 Griffin 提出的幅度谱重构信号法进行音调升高处理后的声音的短时傅里叶变换幅度谱，图 6c 是用本文提出的方法进行音调升高处理后的声音的短时傅里叶变换幅度谱。从图 6 中可以看出，Griffin 的方法虽然升高了音调，却同时改变了声音共振峰的位置和带宽，在改变音调的同时损坏了原声音的音色。这是因为 Griffin 的方法没有将基音频率参数和共振峰参数区分开来造成的。本文提出的方法解决了 Griffin 方法中的这一不足，有效地将基音频率参数和共振峰参数区分了开来，因此不会在升高音调的同时改变声音的音色。

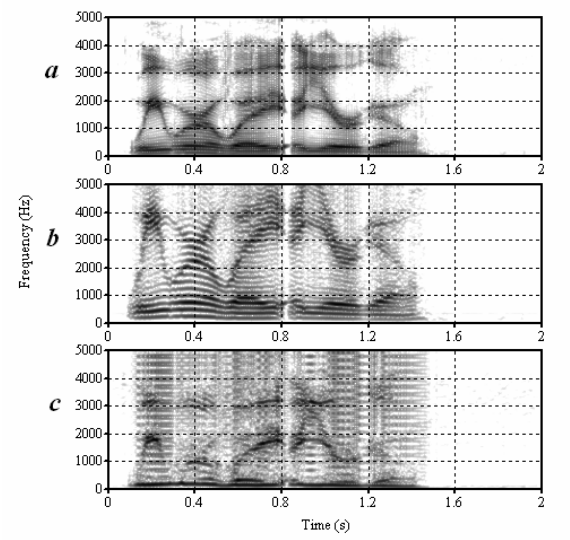

图 6 声调变换实验结果
Fig.6 Experiment result of pitch modification

3.3 音色变换

音色变换的原理如图7所示。其中 L_a 为声音分析时的窗移长度，L_s 为声音合成时的窗移长度，L 为窗口长度。在音色变换时，L_a 等于 L_s，声音分析时的窗口长度和声音合成时的窗口长度均等于 L。通过线性预测分析得到声道滤波器后，可对其零点进行修改来改变滤波器的滤波特性。

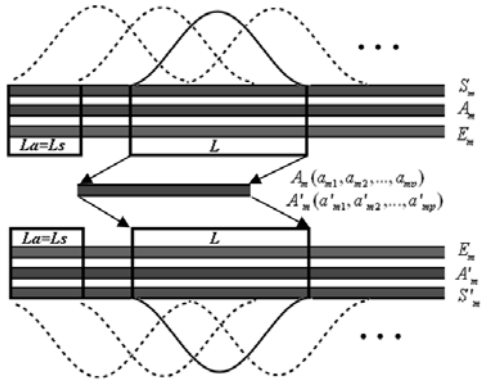

图7 音色变换原理图
Fig.7 Principle of timbre modification

图 8a 显示了英文句子“We were away a year ago.”的短时傅里叶变换幅度谱，图 8b 是利用本文方法对声音音色进行修改后的声音的短时傅里叶变换幅度谱。从图 8 中可以看出，利用本文的方法在对声音共振峰进行修改后并没有改变声音的基音频率，因此在改变音色的时候不会改变音调。而在 Griffin 的方法中，由于声音的基音频率参数和共振峰参数没有区分开来，音调和音色是混合在一起的，在改变音调的同时必然改变音色，反之亦然，图 6b 清楚地说明了这一点。

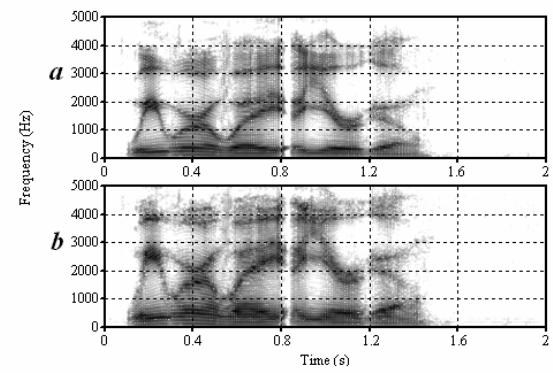

图8 音调变换原理图
Fig.8 Experiment result of timbre modification

4 结束语

本文提出了一种基于源滤波模型短时傅立叶变换幅度谱的声音变换技术。成功解决了读书机器人语音库不能满足多样性需求的问题。利用该方法可以很好地处理声音的节奏变换、音调变换和音色变换。该法将声音的基音频率参数和共振峰参数区分开来，使得音调变换和音色变换不至相互影响。因此能够在改变音调的时候不改变音色，这就保持了原声音的个性特征；在改变音色的时候不改变音调，这就保持了原声音的音调特点。当然也可以利用本文提出的方法将声音的节奏变换、音调变换和音色变换结合在一起以获得一个具有特定效果的混合变换声音。

参考文献

[1] Wayman, J.L., Reinke, R.E., Wilson, D.L. High quality speech expansion, compression, and noise filtering using the sola method of time scale modification [C] // IEEE Cat. 1989 714-717

[2] Verhelst W, Roelands M. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech[C]// Proceedings of ICASSP '93. 1993 554-557

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

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

[5] Dorran D, Lawlor R, Coyle E. High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PAOLA)[C].//ICASSP'03. 2003 700-703

[6] Portnoff M. Time-scale modification of speech based on short-time Fourier analysis[J]. **IEEE Transactions on Acoustics, Speech and Signal Processing**, 1981, **29**(3): 374-390.

[7] Griffin D, Jae L. Signal estimation from modified short-time Fourier transform[J]. **IEEE Transactions on Acoustics, Speech and Signal Processing**, 1984, **32**(2):236-243

[8] Xinglei Z, Beauregard G, Wyse L. Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra[J]. **IEEE Transactions on Audio, Speech, and Language Processing**, 2007, **15**(5): 1645-1653.

收稿日期：2011-09-04；**修返日期：** **基金项目：**金项目 1 全称（基金项目号）；基金项目 2 全称（基金项目号）；……
作者简介：邓杰(1986－)，男，湖南武冈人，硕士研究生，主要研究方向为声音信号处理、声音转换(ddqre@163.com)；赵群飞(1960－)，男，教授，博士，主要研究领域为主要研究方向为机器视觉与基于图像的测控系统理论和方法、两足步行机器人、医疗服务等特种机器人的智能控制理论与设计.