用于读书机器人的声音变换方法

邓杰[1]，房宁[2]，赵群飞[1]

（1 上海交通大学自动化系，系统控制与信息处理教育部重点实验室，上海 200240；2 上海交通大学 国际教育学院 上海 200030）

摘　要：为了增加读书机器人(JoyTon)朗读声音的多样性,提出了基于单一语音库的声音变换方法，成功地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，把原始声音信号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相互干扰，取得了良好的声音变换效果，使读书机器人可用丰富多彩的感情和声调朗读。

关键词：读书机器人；声音变换；节奏变换；音调变换；音色变换

中图分类号：TP391.41　　**文献标识码**：A

Voice modification for book reading robot

DENG Jie[1], FANG Ning[2], ZHAO Qunfei[1]

（1 *Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China;*

2 *School of International Education, Shanghai Jiao Tong University, Shanghai 200030, China*）

Abstract: Voice modification based single speech database is proposed in order to increase the diversity of the book reading robot (JoyTon). Mutual interference between pitch modification and timbre modification is successfully resolved. 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 of voice modification using the proposed method show that pitch parameters and formants parameters can be effectively separated without interference between pitch and timbre, and thus the resultful modifications make JoyTon read a text with a variety of emotions and tones.

Keywords: book reading robot; voice modification; tempo modification; pitch modification; timbre modifcation

1 引言（Introduction）

世界卫生组织估计，目前全世界约有一亿八千万人患有视觉残疾，其中约四千五百万人完全失明。如何帮助盲人获取纸质出版物上的信息具有重要意义。作者所在实验室与日本早稻田合作研制了读书机器人（JoyTon）来帮助盲人阅读各种出版物。该读书机器人，如图 1 所示，具有自动翻书功能，利用视觉传感器获取文字图像信息并OCR（Optical Character Recognition）功能自动识别出文字，然后利用TTS（Text to Speech）文本语音转换与合成功能将识别出的文字内容转换成声音朗诵给用户。常用的语音合成功能依靠装载语音库，利用微软的TTS语音合成引擎来实现。然而由于受到语音库的限制，只能合成出语音库中既存的声音。这导致读书的声音比较单调，而且缺乏感情色彩。如果依靠切换不同的语音库来实现声音的多样性就要占用更多的系统软件和硬件资源，增加用户的经济负担。因此，研究通过单一语音库中的声音进行变换来合成出具有多种特色的声音来满足人们个性化需求的TTS技术和方法，越来越受到广泛的关注。

图 1 读书机器人(JoyTon)
Fig.1 Book reading robot (JoyTon)

基于语音库的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。近年来，研究人员提出了一系列声音变换的方法。例如，Portnoff提出了用短时傅立叶变换来进行节奏变换[1]，Griffin等人提出了利用修

通讯作者：邓杰，电子邮件：ddqre@163.com　收稿/录用/修回：yyyy-mm-dd/yyyy-mm-dd/yyyy-mm-dd

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[2]。为了提高声音变换的实时性，又相继提出了同步叠加法[3]，波形相似叠加法[4]，声码器以及各种改进方法[5-6]，峰值对齐叠加法[7]。在Griffin的基础上Xinglei等人又提出了实时语谱迭代转换法和超前实时语谱迭代转换法[8]。这些算法虽然在实时性方面有了较大改善，但却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。虽然能在一定程度上进行声音变换，但由于音调和音色不能分开调整，使得变换的多样性受到限制。

为了解决声音变换中音调变换和音色变换相互干扰的问题，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，使得音调和音色能分开调整。声音信号首先通过源滤波模型分解成激励信号和声道滤波器信号，然后利用短时傅里叶变换幅度谱信号重构法修改激励信号，并根据需要同时修改声道滤波器参数。最后将修改后的激励信号经修改后的声道滤波器滤波生成变换后的声音信号，有效地将基音参数和共振峰参数分开调整，避免音调变换和音色变换的混叠。

2 声音信号的分解（Decomposition of the voice signal）

利用源滤波模型，把声音信号可看作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 2 显示了源滤波模型的原理。

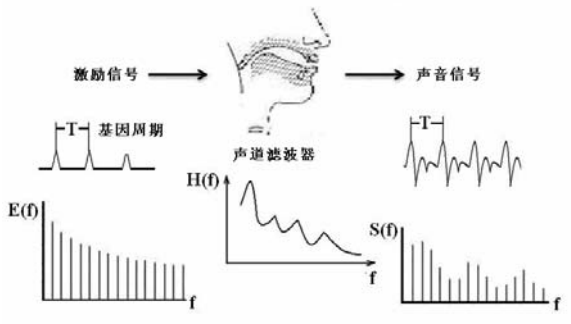

图 2 源滤波模型原理
Fig.2 Principle of source filter model

源滤波模型的基本原理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，$n=1,2,...$ 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，即 $s(n)$ 的预测值 $\hat{s}(n)$ 可表现为

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

其中 $\hat{s}(n)$ 和 $s(n)$ 均为离散实数序列，a_k（$k=1,2,...,p$）为加权系数，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 $\hat{s}(n)$ 将无限接近原始声音信号 $s(n)$。通常情况下，p 取 10 到 12。原始声音信号与其预测信号的误差称为激励信号，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，误差信号可由原始声音信号 $S(z)$ 和传递函数 $A(z)$ 相乘得到。$A(z)$ 代表了声道滤波器。

3 激励信号和声道滤波器的变换（Modification of excitation signal and vocal tract filter）

首先对激励信号进行变换。将上述（2）中的激励信号 $e(n)$ 经傅里叶变换到频域得到其频谱信号 $E(mS,\omega)$。于是，我们可以在频率域对 $E(mS,\omega)$ 进行修改，然后再将修改后的频谱信号 $E'(mS,\omega)$ 反变换回时域得到修改后的激励信号 $e'(n)$。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，所以很多时候需要直接利用幅度谱来重构信号。也就是说，需要首先得到激励信号 $e(n)$ 在频域的幅度谱 $|E(mS,\omega)|$，然后根据需要对幅度谱进行修改得到 $|E'(mS,\omega)|$，最后再利用修改后的幅度谱 $|E'(mS,\omega)|$ 来重构激励信号。利用修改过的激励信号幅度谱 $|E'(mS,\omega)|$ 来重构激励信号，使重构激励信号的短时傅里叶变换幅度谱尽可能地接近目标激励信号的短时傅里叶变换幅度谱，可通过一种目标激励信号 $e(n)$ 与重构激励信号 $e'(n)$ 相似性距离 $D[e(n),e'(n)]$ 来评价，其定义如下式所示

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

其中 $|E(mS,\omega)|$ 是目标激励信号的短时傅里叶变换幅度谱，$|E'(mS,\omega)|$ 为重构激励信号的短时傅里叶变换幅度谱。利用 $|E^i(mS,\omega)|$ 取代 $|E'(mS,\omega)|$，可

通讯作者：邓杰，电子邮件：ddqre@163.com
收稿/录用/修回：yyyy-mm-dd/yyyy-mm-dd/yyyy-mm-dd

推出(5)式所示的迭代方程[2]，其中 i 为迭代次数。

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

这里有

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

其中 $|E^{i}(mS,\omega)|$ 为 $e^{i}(n)$ 的短时傅里叶变换幅度谱，$w(mS-n)$ 为窗函数，本文取汉明窗。可以证明距离测度 $D[e(n),e^{i}(n)]$ 会随着迭代次数 i 的增加而逐步减小[2]。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。由于标准的叠加公式(7)能取得和(5)式相同的效果，但运算量较(7)式要小。因此，为了提高运算的实时性。本文采用标准叠加公式(7)来取代迭代式(5)。

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

利用(6)、(7)式，取重构激励信号 $e'(n)$ 为 $e^{4}(n)$ 或 $e^{5}(n)$ 。

接下来对声道滤波器进行变换。实际上，声道滤波器 $A(z)$ 为一个全零点滤波器。可根据需要来调整其零点的分布，改变其幅度谱的峰值的位置和大小，从而调节声音共振峰的频率和带宽。不同的滤波器具有不同的滤波特性，产生出不同的音色。

4 声音变换的步骤及结果（Steps and results of voice modification）

图 3 给出基于源滤波模型的声音变换的具体步骤，将节奏变换，音调变换，音色变换统一到一个流程当中。图 3 中的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。通过调整声音分析时的窗移长度和声音合成时的窗移长度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，通过调整每一帧声音的声道滤波器 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。

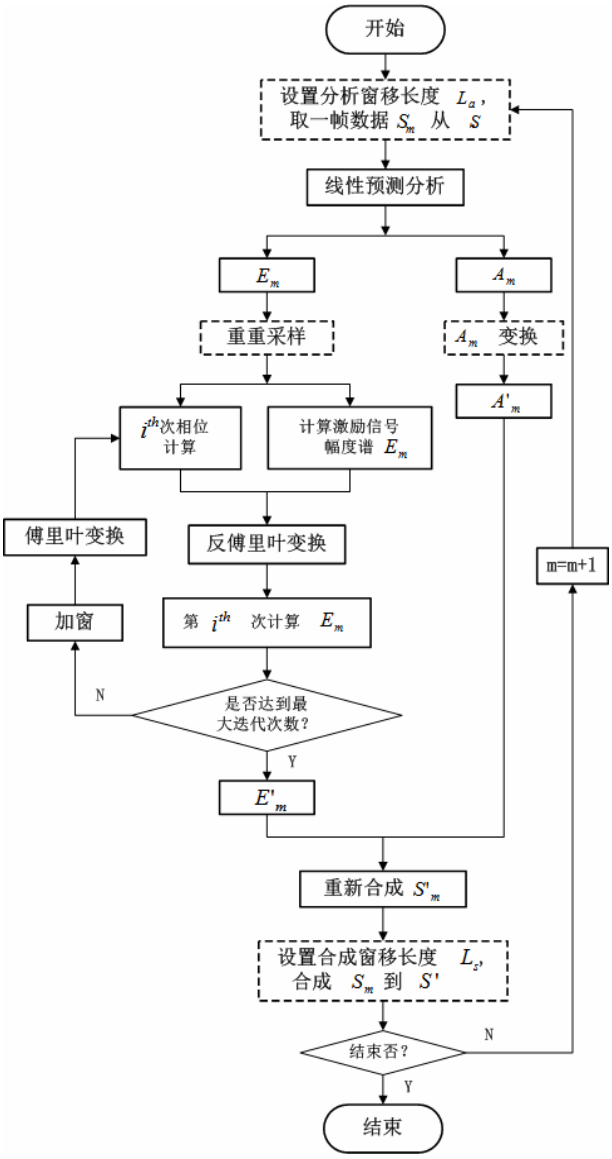

图 3 声音变换流程图

Fig.3 Diagram of voice modification

4.1 节奏变换（Tempo modification）

节奏变换的原理如图 4 所示。其中 L_{ma} 为声音分析时的窗移长度，L_{ms} 为声音合成时的窗移长度，L_{wa} 分析窗口长度，L_{ws} 为合成窗口长度。本文中的窗函数取汉明窗。节奏变换时 L_{wa} 等于 L_{ws}。通过改变 $L_{\mathrm{wa}}/L_{\mathrm{ws}}$ 的比值来调节声音的节奏。当 $L_{\mathrm{wa}}/L_{\mathrm{ws}}>1$ 可以加快原声音的节奏，$L_{\mathrm{wa}}/L_{\mathrm{ws}}<1$ 则放慢原声音的节奏。

通讯作者：邓杰，电子邮件：ddqre@163.com
收稿/录用/修回：yyyy-mm-dd/yyyy-mm-dd/yyyy-mm-dd

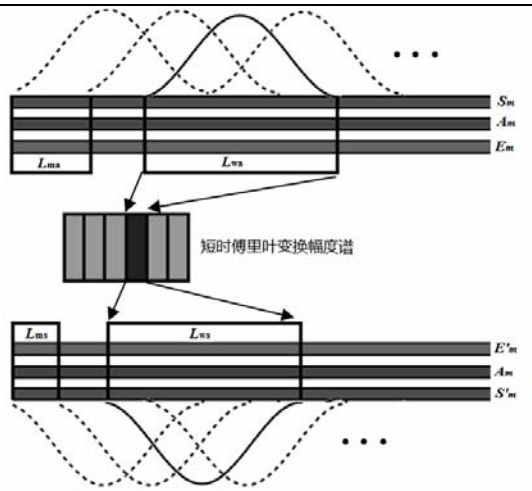

图 4 节奏变换原理图
Fig.4 Principle of tempo modification

图 5a 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 5b 是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 5 中可以看出，原始语音的节奏则加快了 1.5 倍，而声音的基音频率、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。

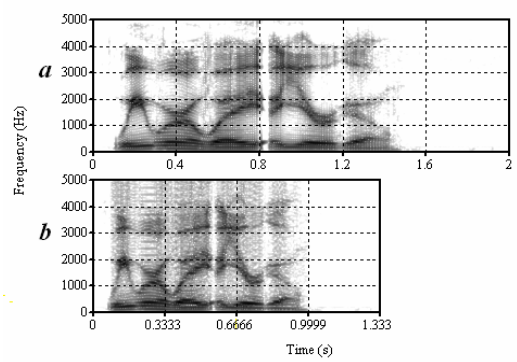

图 5 节奏变换结果图
Fig.5 Result of tempo modification

4.2 音调变换(Pitch modification)

音调变换原理如图 6 所示。其中 L_{ma}、L_{ms}、L_{wa} 和 L_{ws} 的定义与 4.1 中的相同。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_{ma} 等于 L_{ms}，通过改变 L_{wa}/L_{ws} 的比值来调节音调。当 $L_{wa} > L_{ws}$ 时可以提高原声音的音调，而 $L_{wa} < L_{ws}$ 则降低原声音的音调。

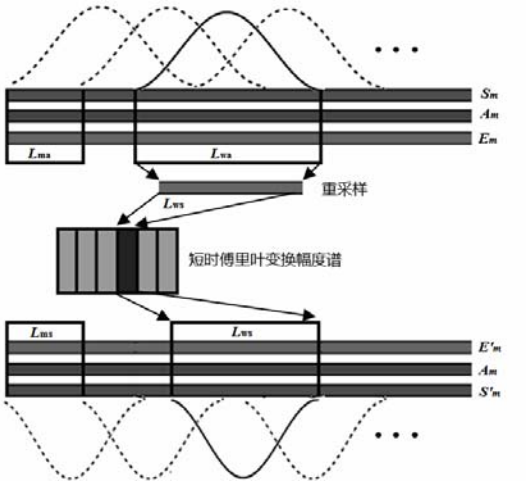

图 6 音调变换原理图
Fig.6 Principle of pitch modification

图 7a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 7b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 7c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 7 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[2]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。

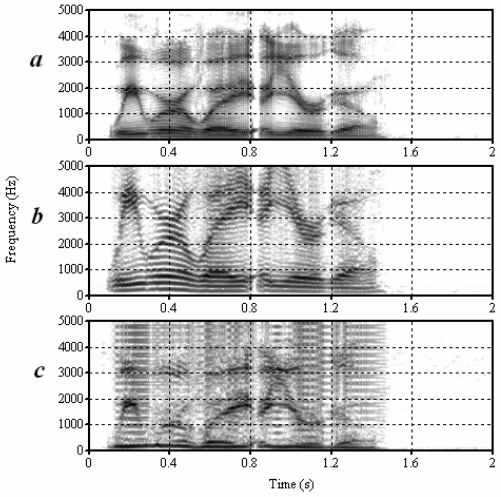

图 7 音调变换结果图
Fig.7 Result of pitch modification

通讯作者：邓杰，电子邮件：ddqre@163.com
收稿/录用/修回：yyyy-mm-dd/yyyy-mm-dd/yyyy-mm-dd

4.3 音色变换(Timbre modification)

音色变换的原理如图8所示。其中 L_{ma}、L_{ms}、L_{wa} 和 L_{ws} 的定义与4.1中的相同。在音色变换时，L_{ma} 等于 L_{ms}，L_{wa} 等于 L_{ws}。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。

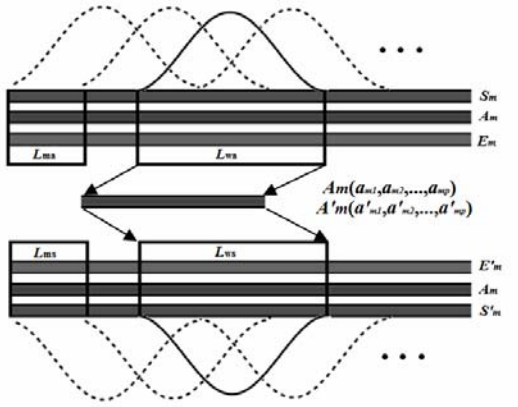

图 8 音色变换原理图
Fig.8 Principle of timbre modification

图 9a 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 9b 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 9 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。因此，在改变音调的同时会改变音色，反之在改变音色的同时也必然改变音调。图 7b 清楚地说明了这一点。

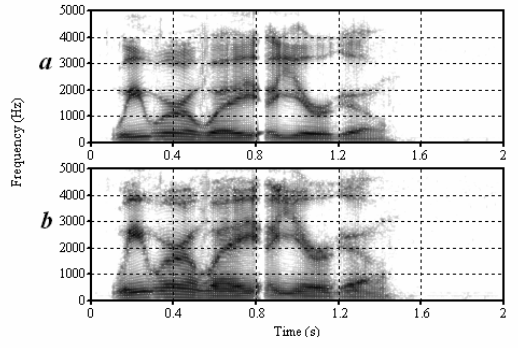

图 9 音色变换结果图
Fig.9 Result of timbre modification

4.4 JoyTon 的用户界面(User interface of JoyTon)

根据上述原理，我们设计了读书机器人 JoyTon 的用户界面如 10 所示。图 10 的右上角为采集到的图像数据，经 OCR 文字识别后显示在左上角的文本框里。配合对节奏、音调和音色等进行调整就能合成产生出各种声音效果。

图 10 JoyTon 的用户界面
Fig.10 User interface of JoyTon

5 结论（Conclusion）

本文提出了一种基于源滤波模型和短时傅立叶变换幅度谱信号重构的声音变换技术。成功地解决了依靠语音库不能满足读书机器人多样性需求的问题。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。音调和音色能在互不影响的情况下各自调整就使得声音变换更加灵活，在一定意义上又增加了声音的多样性。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，可使读书机器人朗读具备更加丰富多彩的声音效果。

参 考 文 献 (References)

[1] 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.

[2] 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

[3] 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]//Twenty-Third Asilomar Conference

通讯作者：邓杰，电子邮件：ddqre@163.com　收稿/录用/修回：yyyy-mm-dd/yyyy-mm-dd/yyyy-mm-dd

on Signals, Systems ands Computers: Vol.2. San Jose, CA, USA: Maple Press, 1989: 714-717

[4] 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: Vol.2. New York, NJ, USA: IEEE, 1993:554-557

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

[6] 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.

[7] Dorran D, Lawlor R, Coyle E. High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PAOLA)[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing: Vol.1. Piscataway, NJ, USA: IEEE, 2003:700-703

[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.

作者简介：

邓 杰（1986--），男，硕士研究生。研究领域：智能机器人控制，声音信号处理。

房 宁（1963--），女，语言学硕士。研究领域：对外汉语教学。

赵群飞（1960--），男，理学博士，教授，博士生导师。研究领域：特种机器人智能控制，机器视觉等。

通讯作者：邓杰，电子邮件：ddqre@163.com 收稿/录用/修回：yyyy-mm-dd/yyyy-mm-dd/yyyy-mm-dd