基于多时间尺度韵律特征分析的语音转换研究

李燕萍，张玲华
（南京邮电大学 通信与信息工程学院，江苏 南京 210003）

摘 要：为了提高转换语音的可懂度与自然度，文中在语音信号的特征抽取方面，注重对语音信号韵律特性的研究，提出了一种多时间尺度的韵律特性抽取方法及其参数化表示，基于逐级细化的策略实现语音信号在多时间尺度下的韵律特征分析与提取，实现对韵律特性从整体到局部细致完整地刻画，克服了韵律信息表述的模糊性和复杂性。实验结果表明，文中提出的语音转换系统在四种测试类型中性能良好，与现有的高斯混合模型相比，ABX 测试结果提高了 10.88%，同时 MOS 得分平均提高了 18.59%。

关键词：语音转换；韵律；多时间尺度；高斯混合模型
中图分类号：TP31 **文献标识码**：A **文章编号**：1673-629X(2012)12-0067-04

Voice Conversion Research Based on Multi-time Scale Prosodic Feature Analysis

LI Yan-ping, ZHANG Ling-hua
(College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In order to improve the conversion speech intelligibility and natural degrees, based on speech signal feature extraction, pay great attention to the research of speech signal prosody characteristics, put forward a prosody characteristics extraction method based on multi-time scale and parameterized representation. Based on stepwise refinement strategy, achieve the implementation of prosodic feature extraction on different time scales, which can enable detailed full description for prosodic information from global to local, overcome the ambiguity and complexity of prosody characterization. The experimental results show that the performance of proposed voice conversion system in four test type is good, and compared with existing Gaussian mixture model, ABX test results increased by 10.88%, and at the same time, MOS scoring average is improved by 18.59%.

Key words: voice conversion; prosody; multi-time scale; Gaussian mixture model

1 概 述

语音转换（Voice Conversion，VC）的研究目标是将源说话人语音中的个性特征转变为目标说话人的个性特征，使转换后的语音听起来就像是目标说话人的声音，而其中的语义信息保持不变[1~3]。该技术是在说话人识别和语音合成的研究基础上发展起来的，是对两个分支内涵的丰富和延拓，在语音翻译（Speech-to-Speech Translation，SST）系统、文-语转换（Text-to-Speech，TTS）系统、极低速率的语音编码、受损语音恢复和语音增强等方面都有极其重要的用途[4]。

性能良好的语音转换系统，既要保持重构语音的听觉质量，又要兼顾转换后的目标说话人个性特征是否准确，现有的语音转换算法大多单纯关注频谱特征的准确转换，重建的语音尽管与目标说话人声音个性相似，但存在发音模糊和不连续现象，自然度较差。目前韵律特性的研究已经引起语音转换、情感语音合成和情感识别等情感信息处理领域的广泛重视[5~8]。然而由于韵律信息的表述具有不稳定性，对其建立有效的数学模型和参数提取仍存在一定的困难与挑战。文献[9]提出一种使用声调映射码本的汉语声音转换方法，从源语音和目标语音分别提取汉语单音节的基频曲线作为基频变换单元，依据时间对准原则建立声调模式映射码本，有效改善了声音转换性能，但在训练阶段需要进行准确的音节标注且只适合于汉语。文献[10]提出基于基元段特征的 $F_0 \sim t$ 转换，由连续几个短时帧组成一个基元段，从基元段的整体考虑提取的

收稿日期：2012-04-19；**修回日期**：2012-07-26
基金项目：国家自然科学基金资助项目（60902065，61001152，61172118）；浙江省自然科学基金（Y1090649）；南京邮电大学引进人才基金（NY209004）
作者简介：李燕萍（1983- ），女，陕西渭南人，博士，讲师，研究方向为语音转换、说话人识别；张玲华，教授，博士生导师，研究方向为语音增强、多媒体通信。

特征矢量能够有效反映基频轨迹的局部时间演变特征,但只适合于文本相关的情形,在文本无关的测试环境中就产生了失配现象。文献[11]提出一种基于基音同步分析方法提取语音信号中不同层次的特征,利用神经网络来建立映射关系,实现高质量的语音转换。由以上分析可以看出,如何提取韵律特征以及如何参数化表示成为提高转换质量的关键。

文中将基于逐级细化的策略实现语音信号在不同时间尺度下的韵律特征抽取,然后采用矢量量化算法实现韵律特征的码本映射,在频谱特征转换方面,基于高斯混合模型经典算法实现线谱对参数的映射。

2 语音转换系统中的说话人特征提取

语音信号中包含不同层次的信息,这些信息以复杂的形式混合在一起共同传达语言信息、语义信息和说话人个性信息等。频谱特征主要包括基音频率,线谱对参数(Linear Spectral Frequency, LSF)、倒谱参数、共振峰频率(Formant Frequency)以及频谱倾斜(Spectral Tilt)等[1],这些特征反映激励源和声道特性,是短时频谱特征,属于低层(Low Level)信息。虽然提取简单,但无法反映自然语流中的语调、节奏和情感等。韵律特征包括基频轨迹、能量轨迹和语速等,反映声学特征的时变演化特性,是超音段特征,属于高层(High Level)信息[12],不能依据音段特征的处理方法在短时帧(Frame)的层面上提取固定单一的特征参数。

2.1 多时间尺度韵律特征提取

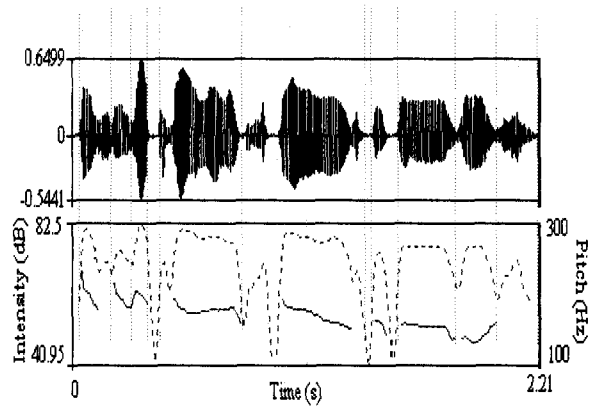
韵律特性包括语调、时长、轻重等,这些不同的特性在不同的时间尺度下得到最适合的体现。鉴于此,文中将深入研究语音信号的语法规则和人耳的听觉感知特性,一个语句可以分解为若干短语,这些短语能够完整独立地表达一个语义。一个短语可以划分为若干音节,每个音节是发音的基本单元。将语音划分为语句(Utterance)、短语(Phrase)和音节(Syllable)三个时间尺度,在多时间尺度上分析语音的韵律特性,分别抽取对应的韵律特征,其中语句尺度下的特征描述韵律的全局长时信息,音节尺度下的特征描述韵律的局部细节信息,短语介于全局和局部之间,描述音节构成语句时的语法规则约束与信息补充。这种抽取方法不仅可以降低韵律特性表述的模糊性和复杂性,也符合语音信号表达高层次信息的方式,提取的韵律特征必将作为音段特征的重要补充。

2.2 韵律特征的参数化描述

图1是音节划分和韵律特性示意图。节奏和重音通过每个音节的持续时长(Duration)和能量大小(Energy)来体现,属于音节尺度下的局部特性。选用时长结合短时均方(Root Mean Square, RMS)能量等参数来描述,对于语音信号$\{x(n)\}$,其中时长就直接采用标注音节的持续采样点数$\{D_t, t=1,2,\cdots,T\}$来描述,其中t表示语句中的第t个音节。标注音节的短时均方能量RMS_t表示为:

$$\mathrm{RMS}_t = \sqrt{\frac{\sum_{i=1}^{D_t} (x(n))^2}{D_t}} \tag{1}$$

考虑到相邻音节之间的动态特性,将当前音节与下一音节参数之间的差值作为补充特征,并且加入整句中音节持续时长和能量的平均值构成联合特征。



(上图为时域波形,下图为韵律特性示意图,其中实线的非连续包络为基率轨迹(Pitch);虚线的连续包络用有效声压级轨迹反映强度(Intensity))

图1 音节划分示意图

响度作为人耳对声音强弱的主观评价,主要决定于声压,而且与频率和频谱有一定的关系。声压越大,引起人耳主观感觉到的响度也愈大[12]。属于语句尺度下的特性,可以用有效声压级曲线(effective sound pressure level contour)的量化码本来表征,其中逐帧的有效声压级表示如下:

$$\mathrm{SPL}_i = 20 \times \log(\mathrm{RMS}_i/(2 \times 10^{-5})) \tag{2}$$

其中RMS_i表示语句中第i帧的短时均方能量。

基音F_0是指发浊音时声带振动的周期性,是短时音段层面的特征,而基频轨迹(Pitch Contour)$F_0 \sim t$的变化则可以体现音高信息,反映说话人声音音调随时间的起伏,当多个音节在上下文语法环境下构成短语时,短语内部的相邻音节以及短语之间会存在协同发音的现象,因此是介于全局特性和局部特性之间,属于短语尺度下的特性,可以用基频轨迹的量化码本及统计参数来表征,包括整句中的最大基频MAX_{F_0},最小基频值MIN_{F_0}以及平均值AVG_{F_0}。

2.3 频谱特征的参数化描述

文中对频谱特征的转换选择目前语音转换系统中使用最为广泛的线谱对LSF参数,主要有三个优点:

(1) 能够很好地表征共振峰的位置和带宽;

(2) 具有良好的插值性能;

(3) 特征参数中某一维系数的失真只会影响重构谱参数中的一部分区域,而不会导致整个谱参数估计错误。

3 语音转换系统的构建

完整的语音转换系统流程图如图 2 所示,在训练阶段,首先分别提取源和目标说话人的韵律特征和频谱特征,包括三个时间尺度下的韵律特征以及线谱对参数,然后分别基于矢量量化和高斯混合模型进行映射规则的获取,完成高质量的语音转换。

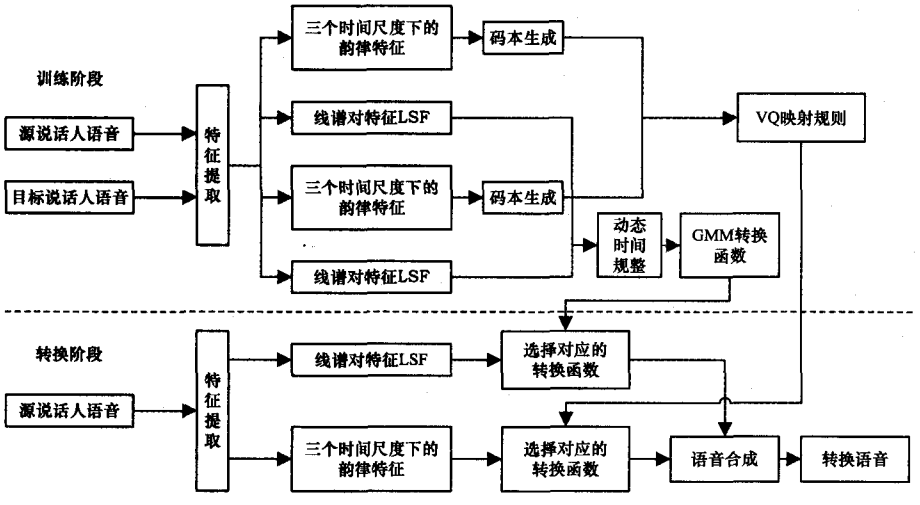


图 2　基于多时间尺度韵律特征的高质量语音转换流程图

3.1 基于 VQ 码本映射的韵律特征转换

码本映射算法的主要思想是先用矢量量化(Vector Quantization, VQ)算法对源说话人和目标说话人的特征矢量空间进行划分,对于在三个时间尺度下提取的韵律特征,分别经过矢量量化生成码本,对源和目标说话人对应时间尺度下的韵律特征进行动态时间规整(DTW),并计算统计直方图,可以得到源说话人的所有码字经对齐后对应目标码字的统计直方图,进一步将该直方图作为加权因子,建立基于矢量量化的映射规则,使得转换后的映射码本是所有目标码字的线性加权平均。

采用矢量量化将经过动态时间规整后的源和目标说话人的韵律特征矢量分别聚类为各自的码本,其中每个码本是由有限个码字组成的序列。以短语尺度下的基频轨迹的量化码本为例来简要说明映射码本的产生过程。首先分析计算源说话人码本中的第 i 个码字 $C_i(S)$ 对应到目标说话人的所有可能的码字 $C_j(T)$ 的出现次数,即训练数据上 $C_i(S) \rightarrow C_j(T)$ 的频度,用权值 h_{ij} 表示该值。基于最小均方误差(Minimum Mean Square Error, MMSE)准则,第 i 个映射码字 X_i 可表示为目标说话人的码字 $C_j(T)$ 的线性加权平均[10],如下所示:

$$X_i = \sum_{j=1}^{N_i} h_{ij} C_j(T) / \sum_{j=1}^{N_i} h_{ij} \tag{3}$$

其中 N_i 表示经过矢量量化后与 $C_i(S)$ 存在对应关系的目标说话人的码字个数。在转换阶段,将输入的源说话人的待转换语音的韵律特征量化成码本,通过训练阶段建立的映射规则进行转换,实现源说话人的韵律特性向目标说话人的转换。

3.2 基于 GMM 的频谱特征转换

基于 GMM 的频谱转换算法对提取的源和目标说话人的线谱对 LSF 参数进行转换,设分别为 L 维 X 和 Y,用动态时间规整算法对 X 和 Y 的频谱特征序列进行对齐,构成两组数目相同并且一一对应的特征参数序列。

设分别为:

$X = \{x_t, t = 1,2,\cdots,T\}$、

$Y = \{y_t, t = 1,2,\cdots,T\}$,

将对应的 x_t 和 y_t 拼接成一个 $2L$ 维的特征向量 $\boldsymbol{Z} = \left\{\begin{bmatrix} x_t \\ y_t \end{bmatrix}, t = 1,2,\cdots,T\right\}$,

形成一个新的向量空间 $\boldsymbol{Z} = \{z_t, t = 1,2,\cdots,T\}$。

基于高斯混合模型对新的向量空间 $\boldsymbol{Z}$ 拟合其概率分布,设 z_t 的概率密度函数为:

$$p(z_t) = \sum_{i=1}^{M} \omega_i \cdot N(z_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), t = 1,2,\cdots,T \tag{4}$$

式中 μ_i 和 Σ_i 分别是第 i 个分量的均值和协方差矩阵,其值分别为:

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{iX} \\ \mu_{iY} \end{bmatrix}, \boldsymbol{\Sigma}_i = \begin{bmatrix} \Sigma_{iXX} & \Sigma_{iXY} \\ \Sigma_{iYX} & \Sigma_{iYY} \end{bmatrix} \tag{5}$$

与此同时也相应地获得了 X 和 Y 的概率密度函数。在 X 服从高斯分布和 X 与 Y 服从联合高斯分布的前提下,如果 X 已知,则在最小均方误差估计准则下,对 Y 的估计为:

$$E[Y \mid X] = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) \tag{6}$$

其中 μ_X 和 μ_Y 分别是 X 和 Y 的均值,Σ_{YX} 是 Y 和 X 的互协方差矩阵,Σ_{XX} 是 X 的协方差矩阵。因此根据(6)式可以推导出在高斯混合模型下,对 $y_t(t = 1,2,\cdots,T)$ 的 MMSE 估计为:

$$y_t = E[y_t \mid x_t] = \sum_{i=1}^{M} \omega_i [\mu_{iY} + \Sigma_{iYX}\Sigma_{iXX}^{-1}(x_t - \mu_{iX})] \tag{7}$$

上式即为源说话人的频谱特征参数映射到目标说话人的频谱特征参数的转换函数，在转换阶段，使用该函数对源说话人的频谱特征进行转换，最终得到转换后的频谱特征。

4 实验结果与分析

4.1 语音库

文中实验选用的对称语料来自 CMU ARCTIC 语料库，采样频率为 16kHz，量化采用 16bit，单声道录音。该语料库包括 7 位说话人(5 男 2 女)，每个说话人的语音内容包括 1132 个音素均衡的语句[13]。在实验中选择其中的两位男性和两位女性(分别为 BDL, RMS 和 SLT, CLB)的语音，实验分为四种测试类型，分别为女声到男声(SLT-BDL)、女声到女声(SLT-CLB)、男声到女声(RMS-CLB)、男声到男声(BDL-RMS)。

从每个说话人的语句中选择 300 个语句用于系统训练，10 个语句用于系统性能测试。预处理包括使用汉明窗进行处理，512 点进行分帧，帧重叠 50%，预加重参数为 0.96，实验重选取 16 维的线谱对 LSF 参数作为频谱信息的特征矢量。对于韵律特征参数，不同语句的长度之间存在一定的差别，对于音节尺度下的韵律特征，持续时长、短时均方能量参数及其一阶差分特征，结合整句中音节持续时长和能量的平均值构成联合特征，特征维数根据较长语句确定为 30 维，对于较短语句的特征维数不足 30 维时，在其后补充零。对于短语尺度下的韵律特征，每个语句用基频轨迹的量化码本及统计参数来表征，包括整句中的最大基频 MAX_{F_0}、最小基频值 MIN_{F_0} 以及平均值 AVG_{F_0}，特征维数确定为 15 维。对于语句尺度下的韵律特征，每个语句用有效声压级曲线的 15 维量化码本来表征。

4.2 ABX 测试

文中主要分析韵律特性的研究对转换后语音质量的改善程度。一般而言，对转换语音的质量评估从可懂度、自然度和说话人识别度三个方面进行考察。其中 ABX 测试方法是用来测试说话人识别性能的一种方法，其中的 A，B 分别表示转换后的语音更接近于 A 或者更接近于 B，X 则表示两者都不是[13]。

文中提出的系统是基于 GMM 进行频谱特征转换的同时对从不同时间尺度提取的韵律特征分别进行码本映射，实现韵律特性的转换。进行性能测试对比的基准系统是基于 GMM 进行频谱特征转换，对基音周期进行单高斯建模，实现均值线性变换。

在四种测试类型中将共产生转换语句为 4 * 10 * 20=800 个，表示语料库中的四位说话人的 10 个转换语音在四种测试类型中，分别由 20 位志愿者进行测试，这些志愿者都经过一定的语音学知识培训。测试结果如表 1 所示，表中的“43(5.37)”表示有 43 个转换语句被判断为源说话人，占全部转换语句的 5.37%。从实验结果分析可得，基准系统中存在大量的语句被判定为既不是源说话人又不是目标说话人，这也是目前语音转换系统面临的主要问题之一，而结合对韵律特征的码本映射转换，可以一定程度上增强转换语音中的说话人个性特征，提高其说话人倾向性，也再次证明了韵律特征是对音段特征的重要补充。在对不同时间尺度提取的韵律特征进行转换的系统中，有 93.63% 的转换语句更接近于目标说话人，而基准方法中只有 82.75% 的转换语句更接近于目标说话人，相比提高了 10.88%。

表 1 ABX 测试性能对比

	基准方法	文中方法
A	43(5.37)	30(3.75)
B	662(82.75)	749 (93.63)
X	95(11.88)	21(2.62)

4.3 MOS 得分实验

平均意见得分法(Mean Opinion Score, MOS)是用来测试转换语音的自然度和可懂度，在其打分机制中，对转换语音的得分划分为五种等级，分别为 1 ~ 5，其中 1 为最差，5 为最优。20 位志愿者对四种测试类型中的转换语音进行打分，实验结果如表 2 所示，由测试结果观察可得，文中提出系统的 MOS 得分性能要明显高于原基准系统的得分，相比平均提高了 18.59%。分析认为，文中提出的转换方法不仅可以获取更多的说话人个性信息，而且可以通过调节转换合成语音的韵律特性，显著提高转换语音的自然度和可懂度。

表 2 四种测试类型中的 MOS 得分

	基准方法	文中方法
SLT-BDL	3.65	4.20
RMS-CLB	3.74	4.15
BDL-RMS	3.58	4.38
SLT-CLB	3.61	4.56

5 结束语

文中提出了一种基于多时间尺度韵律特征的语音转换系统，在语音信号的特征抽取方面，注重对语音信号韵律特性的研究，基于逐级细化的策略实现语音信号在多时间尺度下的韵律特征分析与提取，实现对韵律特性从整体到局部细致完整地刻画，克服了韵律信息表述的模糊性和复杂性，为转换语音质量的提高和自然度的改善奠定了良好的基础。

实验结果表明，文中提出的算法与基准算法相比

(下转第 74 页)

Effort = UCP×PF = 88×36 = 3168h (PF = 36)

通过上述的计算可知该项目的工作量在1760h到3168h之间。假设每个工作人员一周工作35h,且共有8个人员参与,则每周的工作量为35×8 = 280h,由1760÷280 = 6.29 week,2464÷280 = 8.8和3168÷280 = 11.34 week可知,完成该项目所需的时间在7周到12周之间。但一般估算的时候往往只计算PF = 20和PF = 28时的项目工作量,简而言之,完成该项目所需的时间在7周到9周之间。

4 结束语

现有的国内外的软件估算方法的确很多,但是在这众多的估算方法中,能将技术复杂度因素和环境复杂度因素充分考虑进去的并不是很多。而文中讨论的用例点估算方法就充分考虑这些主观因素,并且对这些主观因素的影响程度都用具体的表格表示出来,相关的计算也用对应的公式表示出来,方便计算。通过对UCP算法的具体介绍以及详细的估算过程步骤,最后又通过一个小的案例使人们进一步了解该方法的特色及其方便性,并且随着项目开发的进一步深入,将会发现使用用例点数得到的估算可靠性[13]。当然文中介绍的用例点算法可能还存在一定的缺陷有待进一步的改进。

参考文献:

[1] 任永昌,邢 涛,刘大成.基于网络图的软件项目进度计划编制[J].吉林大学学报,2011,29(2):128-134.

[2] Adolph S,Bramble P,Cockburn A,et al. Patterns for Effective Use Cases[R]. [s. l.]:Addison-Wesley,2002.

[3] Cockburn A. Writing Effective Use Cases[R]. Boston:Addison-Wesley,2001.

[4] Ochodek M,Nawrocki J,Kwarciak K. Simplifying effort estimation based on use case points [J]. Information and Software Technology,2011,53(3):200-213.

[5] 周 杨,吴海涛,张栋伟.扩张的用例点估算[J].计算机技术与发展,2006,16(12):64-66.

[6] Albrecht A. Measuring application development productivity [C]//Proceedings of the Joint SHARE/GUIDE/IBM Application Development Symposium. [s. l.]:[s. n.],1979:83-92.

[7] Symons C R. Software Sizing and Estimating:Mk II FPA (Function Point Analysis) [R]. New York,NY,USA:John Wiley & Sons,Inc. ,1991.

[8] 薛 丹,杨 宸,周 健.一种基于区间值的模糊访问控制策略研究[J].计算机技术与发展,2012,22(1):246-249.

[9] 余秋冬,徐 辉.用例点估算方法在电信行业中的应用[J].计算机工程,2009,35(24):276-277.

[10] Schneider G,Winters J. Applying Use Cases:A Practical Guide[R]. Boston,MA,USA:Addison-Wesley Longman Publishing Co. ,Inc. ,1998.

[11] 赵文杰,刘俊萍,南振岐.改进的用例点估算方法[J].电脑知识与技术:学术交流,2010(12):9917-9919.

[12] 王 悠,罗燕京,易福华,等.基于用例的软件复杂度估算及应用[J].计算机技术与发展,2007,17(7):196-199.

[13] 陶 强,刘 莉,薛振清.基于UML的软件规模估算服务模型研究[J].信息技术与信息化,2010,7(2):70-73.

(上接第70页)

具有良好的性能,在四种测试类型中,ABX测试结果提高了10.88%,MOS得分平均提高了18.59%。

参考文献:

[1] Stylianou Y. Voice transformation:a survey[C]//International Conference on Acoustics,Speech and Signal Processing. [s. l.]:[s. n.],2009:3585-3588.

[2] 左国玉.声音转换技术的研究与进展[J].电子学报,2004,26(7):1165-1172.

[3] 李 波,王成友,蔡宣平,等.语音转换及相关技术综述[J].通信学报,2004,25(5):109-118.

[4] Nakamura K,Toda T,Saruwatari H,et al. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech[J]. Speech Communication,2012,54(1):134-146.

[5] 陈 芝,张玲华.基频轨迹转换算法及在语音转换系统中的应用研究[J].南京邮电大学学报(自然科学版),2010,30(5):83-87.

[6] Laskar R H,Talukdar F A,Bhattacharjee R,et al. Voice conversion by mapping the spectral and prosodic features using support vector machine[J]. Applications of Soft Computing,2009,58:519-528.

[7] 尹 伟,易本顺.一种基于正弦激励的线性预测模型的语音转换方法[J].数据采集与处理,2010,25(2):218-222.

[8] Kunikoshi A,Qian Yao,Soong F,et al. Improve F0 modeling and generation in voice conversion[C]//IEEE International Conference on Acoustics,Speech and Signal Processing. [s. l.]:[s. n.],2011:4568-4571.

[9] 左国玉,刘文举,阮晓钢.一种使用声调映射码本的汉语声音转换方法[J].数据采集与处理,2005,20(2):144-149.

[10] 孙 俊.基于激励源及其韵律特征的源-目标说话人声音转换研究[D].合肥:中国科学技术大学,2006.

[11] Rao K S. Voice conversion by mapping the speaker-specific features using pitch synchronous approach [J]. Computer Speech and Language,2010,24(3):474-494.

[12] 赵 力.语音信号处理[M].北京:机械工业出版社,2008.

[13] 李燕萍,张玲华,丁 辉.基于音素分类的汉语语声转换算法[J].南京邮电大学学报:自然科学版,2011,31(1):10-15.