一种面向声音变换的参数化模型*

黄德智　蔡莲红

（清华大学计算机科学与技术系　北京　100084）

2006 年 1 月 13 日收到

2006 年 6 月 24 日定稿

摘要　在源滤波器模型的基础上，利用统计学习方法，建立了一种面向声音变换的混合参数化模型。该模型包括浊音声学模型、清音声学模型和韵律补偿模型三部分。基于线性预测分析和 mel 倒谱分析的浊音声学模型，刻画了说话人声腔的共振特性。基于线性预测分析和噪声源分析的清音声学模型，反映了说话人发清音的特点。基于统计学习方法的韵律补偿模型描述了音高、能量与时长等分布特性。在该混合参数化模型的基础上，提出了一个声音变换算法，并将其应用到汉语音节的变换问题上。实验结果表明，对清浊音和韵律特性分别建模的变换算法能够提高重建语音的清晰度和可懂度，缩小重建语音与目标语音之间的感知距离，使重建语音具有目标说话人的韵律特征。

PACS 数：43.70

A parametric model for voice conversion

HUANG Dezhi　CAI Lianhong

(*Department of Computer Science and Technology, Tsinghua University,* Beijing 100084)

Received

Revised

Abstract　On the basis of the source-filter model, a hybrid parametric model, consisting of a voiced acoustic model, an unvoiced acoustic model and a compensation model of prosody, is presented for voice conversion and built by statistical learning. The voiced acoustic model is built on linear prediction analysis and mel cepstrum analysis to characterize the resonance of the vocal tract of speakers. The unvoiced acoustic model is adopted by linear prediction and noise-source modeling, to reflect the characteristics of the unvoiced speech of speakers. Statistical learning is involved to train the compensation model of prosody, which characterizes the distributions of pitch, energy, and duration respectively. An algorithm on the basis of the hybrid parametric model is proposed and applied to voice conversion of Mandarin syllables. The experiments demonstrate that the proposed algorithm not only improves the articulation and intelligibility of the converted speech, but also reduces the perceptual distance between the target and converted speech significantly. The formal listening tests also show that the prosodic features of target speakers are presented in the converted speech.

引言

语音是人类生活中最自然和最常用的沟通工具。它不仅包含了说话人所要表达的内容（文本信息），携带了说话人的情绪（情感信息），还传递了说话人的个性化特征（身份信息）。其中，个性化特征标识了说话人，在日常交流中发挥着重要作用，使得听音人能够“闻其声而知其人”。

声音变换（Voice Conversion）的目标就是变换语音中的个性化特征。它将输入的源说话人的声音，变换为听感上接近目标说话人的重建语音。我们知道，说话人辨识的研究目标是从个性化特征集内检索和匹配输入语音，从而标识身份信息。而声音变换则是要保留文本信息、情感信息，替换语音中的身份信息。因此声音变换能够应用于多媒体、个性化人机交互、虚拟现实等领域。

研究表明，个性化特征中最主要的成分就是语音音色。它被定义为“听觉的属性，听话人据此属性就能判断音调、响度、音长相同的语音之间的不

* 国家自然科学基金面上项目（60275014）和基金重点项目（60433030）

同点”[1]。语音音色反映了人对超音段特性相同的语音之间的知觉差异。产生这种差异的根源，在于说话人的发声器官的生理构造、社会属性、说话方式、所要表达的情绪和意图等诸多因素。虽然从定义上看，音色、音高和响度是不相关的听觉属性，但是感知实验却发现，音高和响度影响了音色的感知过程。进一步的研究表明，音色感知是多种听觉属性的混合作用的结果。另外，语音的个性化特征还表现在音高、时长和能量等韵律特性上。例如，男声和女声具有不同的音域。而且，超音段特征之间相互影响[2]，并且这些影响还会反映到频谱上。因此声音变换不仅仅要变换说话人频谱特性上的差异，还要变换韵律特性上的差异。

声音变换的研究，涉及到语音生成机理、言语的听觉感知等一系列问题，特别是音色参数化、发音机理中声源与声道的关系、音色与韵律的关联、情感对音色的约束等问题。根据研究角度的不同，声音变换的方法大致可以分为两类：声学修改法和映射法。早期的方法大多数属于第一类。Atal 等人研究了声码器的 LPC 系数与语音特性之间的关系[3]。Childers 等人利用基频、声管长度、声门脉冲包络和能量等参数实现了有限词条的声音转换，提出了“音段不同，分析过程就应该不同”的思想 [4][5]。Valbret 等人采用多变量线性回归的方法进行频谱特征的变换，而利用基音同步叠加方法对激励信号进行韵律特征的修改[6]。但也只是简单地对基频和时长作因子变换，因此重建语音的韵律特征与目标语音仍有较大的差距。随着计算机技术和统计学习理论的不断进步，映射法成为声音变换研究的主流方法。Abe 等提出了基于矢量量化的码本映射方法[7]以后，一些研究人员又提出了改进方案[8][9]。Stylianou 提出了基于高斯混合模型（GMM，Gaussian Mixture Model）建立映射函数的方法，采用谐波噪声模型（HNM，Harmonic + Noise Model）中的谐波系数作为声学特征矢量[10]。HNM 虽然有利于时长与基频的修改，但其系数难于估计。Stylianou 没有考虑清浊音的差异，而且韵律方面的处理限于基频和时长的线性变换。映射方法必须利用源说话人和目标说话人在文本相同的情况下发出的语音作为训练数据。为了解决某些条件下训练数据无法得到的矛盾，Mouchtaris 等人又研究了非平行语音库的声音变换问题[11]。国内的学者也展开了声音变换的相关研究。陈一宁等发现，基于 GMM 的线性变换法会破坏相邻帧的特征矢量的连续性。他们提出了基于平滑 GMM 和最大后验概率自适应的变换方法，力图减少重建语音的频谱跳变，从而改善重建语音的质量[12]。左国玉等采用遗传径向基神经网络捕捉语音频谱的映射关系，以实现不同说话人之间的声音转换[13]。还提出了一种声调码本映射技术，缩小重建语音和目标语音之间个性特征的差异[14]。康永国等分析了重建语音频谱过平滑的问题，提出了使用码本映射和高斯混合模型共同转换声学特征细节的混合变换算法[15]。

上述方法虽然能够取得较好的变换效果，但仍然存在一些问题。众所周知，清音的时长一般比浊音短很多，而且也没有明显的共振峰结构，各个频带的能量趋于均匀分布[16]，因此高斯混合模型的训练算法易于发散，不能得到稳定的映射函数。这提示我们应该对清浊音分别进行处理。另外，通常的声音变换算法将频谱特征的映射作为研究的重点，对韵律特征却只做简单的因子变换，因此重建语音与目标语音在韵律上仍有感知差距。这又提示我们要考虑韵律的细部和动态特征。根据以上分析，我们提出了一种基于统计学习的混合参数化特征模型（HPCM，Hybrid Parametric Characteristic Model），用来描述语音中的个性化特征。该模型包括浊音声学模型、清音声学模型与韵律补偿模型三部分。与 Stylianou 方法不同的是，浊音声学模型采用 mel 倒谱矢量[17]和 LSF（Line Spectral Frequency）矢量作为参数。Mel 倒谱矢量易于估计，但不能很准确地刻画共振峰的分布情况，LSF 矢量的引入有效地弥补了这一不足。清音声学模型采用 LSF 矢量和噪声源参数矢量作为参数，反映了说话人发清音的特点。通过统计学习得到的韵律补偿模型能够准确地刻画说话人的音高、能量与时长等分布情况，使重建语音具有目标语音韵律的动态特征。

在汉语音节的声音变换问题上，利用混合参数化特征模型进行了实验。结果表明，该模型能够缩小重建语音与目标语音之间的感知距离，提高重建语音的清晰度与可懂度。而且该模型还能够直接应用于语音个性化特征的分类和说话人的辨识。

1 混合参数化特征模型

人之所以能够根据语音中的个性化特征来区分发音人，是由于在大脑中存在一套可训练的个性化特征的分析、学习和记忆系统。从音色的角度看，在声学上它直接体现为频谱能量分布的差异。该差异可看成是由声学特征矢量的类中心和分布的不同导致的。

HPCM 采用的特征参数集如图 1 所示。通过线性预测分析，获得清音 LSF 矢量和噪声源参数。通过 mel 倒谱分析和线性预测分析，获得浊音的倒谱矢量和 LSF 矢量。上述参数作为 HPCM 的声学参数。清浊音时长、浊音能量及音高构成韵律补偿模型的参数空间。一个说话人的混合参数化特征模型表示为上述参数空间的多个类中心和方差矩阵。

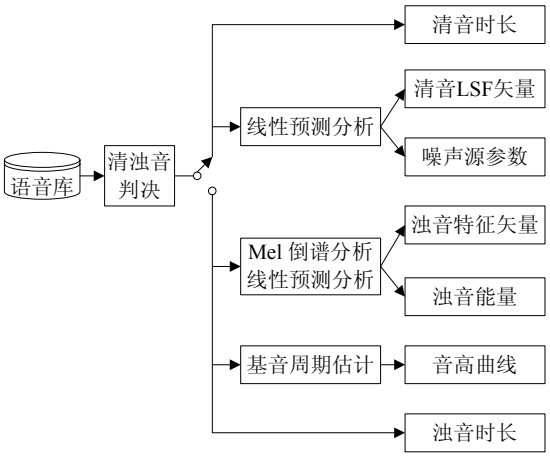

图 1 混合参数化特征模型的参数集

鉴于浊音和清音在生成机理和声学表现上的差异，浊音和清音的声学模型采用了不同的结构和训练方法。

1.1 浊音声学模型

浊音是一种准周期信号，具有明显的共振峰结构。由于声带振动周期、声腔谐振特性、发音用的气力和持续时间的差异，不同的浊音具有不同的音高、时长、幅度和共振峰结构。

本文采用了 mel 倒谱系数 $\{c_\alpha[k]\}$ 作为短时频谱的等价表示[17]。由 mel 倒谱系数构成的估计谱 $\hat{g}_\alpha(\boldsymbol{\Omega}_\alpha)$ 定义为式（1）。

$$\hat{g}_\alpha(\boldsymbol{\Omega}_\alpha)=\sum_{k=-N}^{N}c_\alpha[k]e^{-jk\boldsymbol{\Omega}_\alpha}\quad c_\alpha[k]=c_\alpha[-k]\quad(1)$$

其中 $\boldsymbol{\Omega}_\alpha$ 是频域变换，α 为变换系数，N 为 mel 倒谱的阶数。$\boldsymbol{\Omega}_\alpha$ 将线性频率轴变换为人耳感知的对数频率轴，定义如式（2）。

$$\boldsymbol{\Omega}_\alpha=\boldsymbol{\Omega}+2\tan^{-1}\frac{1-\alpha\cos\boldsymbol{\Omega}}{\alpha\sin\boldsymbol{\Omega}}\qquad(2)$$

最优的估计谱 $\hat{g}_\alpha(\boldsymbol{\Omega}_\alpha)$ 的 $N+1$ 个系数构成的 mel 倒谱矢量代表了原信号的短时频域特征。其中 $c_\alpha[0]$ 代表当前语音帧的短时能量，不参与浊音声学模型的训练。鉴于 mel 倒谱矢量不能很准确地刻画共振峰的分布情况，还要对每一帧短时语音信号进行 K 阶线性预测分析，得到 LSF 矢量。该矢量与 mel 倒谱矢量构成了描述说话人的浊音声学空间的 $P=(N+K)$ 维特征矢量。

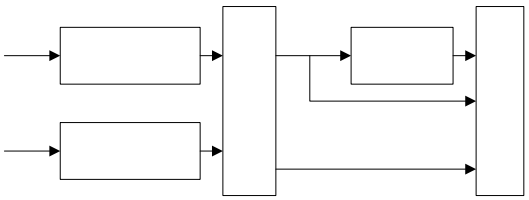

图 2 浊音声学模型和映射函数的训练框图

图 2 所示的是浊音模型与映射函数的训练框图。首先对源说话人和目标说话人的语音中的浊音段分帧和加窗，然后进行 mel 倒谱分析与线性预测分析。再利用 DTW（Dynamic Time Warping）得到时间对齐后的 L 个矢量对 $\{\boldsymbol{x}_i\}$ 与 $\{\boldsymbol{y}_i\}$。源说话人的特征矢量 $\{\boldsymbol{x}_i\}$ 输入到高斯混合模型的训练过程中，得到 M 个均值 $\{\boldsymbol{\mu}_i^{\nu}\}$ 和方差矩阵 $\{\boldsymbol{\Sigma}_i^{\nu}\}$。这些均值看成为说话人空间的中心点，它们近似地对应于不同的音位。方差表示的是说话人空间的分布情况，代表了说话人的音位的变化规律。

为了实现浊音特征矢量的变换，采用了 Stylianou 定义的如式（3）所示的浊音特征映射函数[10]。

$$\begin{aligned}&F^{\nu}(\boldsymbol{x},\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_M,\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_2,\cdots,\boldsymbol{\Gamma}_M)\\&=\sum_{i=1}^{M}P(C_i|\boldsymbol{x})\left[\boldsymbol{v}_i+\boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{\nu^{-1}}(\boldsymbol{x}-\mu_i^{\nu})\right]\end{aligned}\qquad(3)$$

其中 M 是高斯混合数，$P(C_i|\boldsymbol{x})$ 为后验概率，$[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_M]$（记为 $\boldsymbol{V}$）与 $[\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_2,\cdots,\boldsymbol{\Gamma}_M]$（记为 $\boldsymbol{\Gamma}$）是待定参数。最优的映射函数参数通过求解式（4）得到。

$$(\tilde{\boldsymbol{V}},\tilde{\boldsymbol{\Gamma}})=\arg\min_{\boldsymbol{V},\boldsymbol{\Gamma}}\sum_{i=1}^{L}\left\|F^{\nu}(\boldsymbol{x}_i,\boldsymbol{V},\boldsymbol{\Gamma})-\boldsymbol{y}_i\right\|^2\qquad(4)$$

1.2 清音声学模型

清音与浊音不同，它是气流在声腔中受到阻碍时所发出的噪声。清音没有明显的共振峰结构。但是不同清音的频谱分布不同，主要表现在谱质心（强频集中区）的频率不同。清音的特征取决于发音部位和方法以及送气的方式和力度。

给定一段时长为 t 的清音，首先进行 K 阶线性预测分析，得到 K 个 LSF 参数 $\{B_i\}$ 及残差信号 $e(n)$。对 $e(n)$ 进行高斯混合模型估计，得到 Q 个均值 $\{\mu_i^u\}$ 和方差 $\{\sigma_i^u\}$，作为噪声源参数。因此这段清音表示为 $\{B_1\sim B_K,\mu_1^u\sim\mu_Q^u,\sigma_1^u\sim\sigma_Q^u,t\}$。

清音模型的训练包括 LSF 矢量码本生成和噪声源模型训练两部分，如图 3 所示。

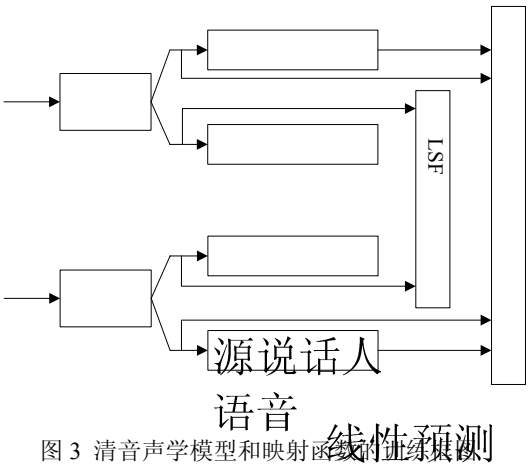

图 3 清音声学模型和映射函数的训练框图

假设源说话人的训练集内存在 U 段清音，则 K 阶线性预测分析得到 LSF 矢量集 $\{\boldsymbol{B}_i^s, i=1,2,\cdots,U\}$。然后利用 LBG 算法构建出一个含有 K 个矢量的清音 LSF 码本矢量，记为 $\{\boldsymbol{\varepsilon}_m^s, m=1,2,\cdots,K\}$。类似地得到目标说话人的清音 LSF 码本矢量，记为 $\{\boldsymbol{\varepsilon}_m^t, m=1,2,\cdots,K\}$。清音 LSF 矢量的映射函数定义为式（5）。

$$F_l^u(\boldsymbol{B}^s,\boldsymbol{H},\boldsymbol{O})=(\boldsymbol{\varepsilon}_1^t\boldsymbol{\varepsilon}_2^t\cdots\boldsymbol{\varepsilon}_K^t)\cdot[\boldsymbol{H}\cdot(\boldsymbol{\varepsilon}_1^t\boldsymbol{\varepsilon}_2^t\cdots\boldsymbol{\varepsilon}_K^t)\boldsymbol{B}^s+\boldsymbol{O}] \tag{5}$$

其中 $\boldsymbol{H}$ 是转换矩阵，$\boldsymbol{O}$ 是补偿矢量。求解式（6）所示的误差函数的最小值，得到最优的变换参数 $\{\hat{\boldsymbol{H}},\hat{\boldsymbol{O}}\}$。

$$\Delta^u=\sum_{i=1}^{U}\left\|F_l^u(\boldsymbol{B}_i^s,\boldsymbol{H},\boldsymbol{O})-\boldsymbol{B}_i^t\right\|^2 \tag{6}$$

对于 U 个噪声源参数矢量 $\{\mu_1^u\sim\mu_Q^u,\sigma_1^u\sim\sigma_Q^u\}$，采用浊音模型的训练方法，得到噪声源参数的映射函数 F_n^u。

1.3 韵律补偿模型

韵律信息在语音个性化特征的感知过程中扮演了重要的角色，声音变换必须考虑韵律特征的转换。因此我们在 HPCM 中增加了一个韵律补偿模型。其参数包括音高及其差分、清音时长、浊音时长和浊音能量。为了更好地反映人耳对语音的感知特性，音高采用音阶标示，参考值为 440Hz。经过清浊音判决就可获得清音时长和浊音时长。Mel 倒谱矢量的 $c_\alpha[0]$ 作为浊音短时能量。

图 4 所示的是韵律补偿模型和映射函数的训练框图。说话人音高、时长和浊音能量模型独立地进行训练。它们的训练方法和浊音声学模型的训练方法相同。对源说话人的清浊音时长、能量和音高，进行统计学习得到相应的 GMM 模型。

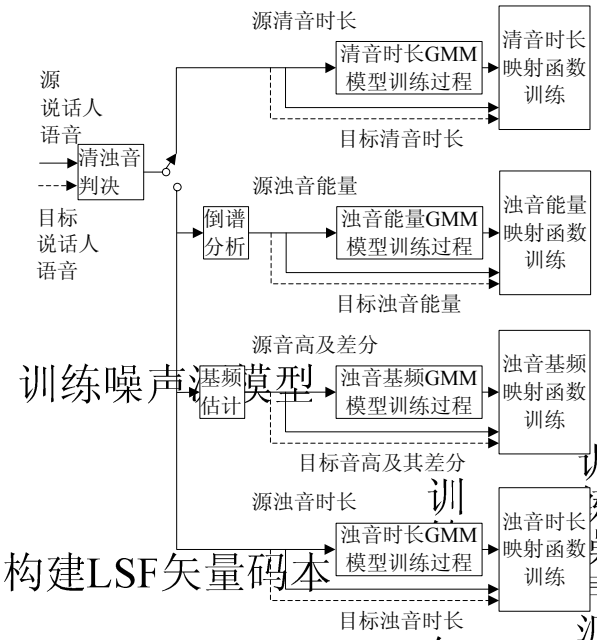

图 4 韵律补偿模型和映射函数的训练框图

通过最小化映射函数的平均误差，估计出映射函数的最优参数。通过时长映射函数将源说话人的清浊音时长变换为目标说话人的清浊音的持续时间。浊音能量的映射函数将源说话人的能量参数转换为目标的能量参数，并利用该参数与浊音特征映射函数得到的矢量，估计出重建语音的能量谱。根据音高映射函数计算的音高曲线，构造出重建语音的脉冲激励信号。经过韵律补偿，重建语音具有了目标说话人的韵律特征。

上述的浊音声学模型、清音声学模型和韵律补偿模型，构成了混合参数化特征模型（HPCM）。前两个模型采用了频谱相关的声学参数，刻画了语音频谱的分布情况，因此可看成是音色的间接模型。它们的训练集是高维矢量空间的点集。基于清浊声学模型的映射函数，将实现语音个性化音色特征的变换。韵律补偿模型刻画了说话人的音高、清浊音时长和浊音能量的特性。补偿模型的训练集是低维空间的点集。补偿模型的映射函数将完成语音个性化韵律特性的变换。

2 基于 HPCM 的声音变换算法

图 5 所示的是基于 HPCM 的声音变换算法的框图。源说话人的语音首先经过清浊音判决，清音段和浊音段分别进行变换，然后将变换结果综合起来作为最终结果。

对于清音段，首先进行线性预测分析，得到 K 阶预测系数，从而计算出 LSF 矢量 $\boldsymbol{B}^s$ 和残差信号

$e(n)$。再通过清音 LSF 映射函数，将 $\boldsymbol{B}^s$ 变换为 $\boldsymbol{B}^t$。从 $e(n)$ 估计出噪声源参数，利用训练得到的 F_n^u 生成新的噪声源参数。再结合转换后的清音时长、LSF 矢量 $\boldsymbol{B}^t$，重建出清音段。

对于源说话人的浊音段，首先分帧和加窗，然后通过 mel 倒谱和线性预测分析得到若干个特征矢量 $\boldsymbol{x}_i$。再利用浊音特征映射函数 F^v，逐帧将 $\{\boldsymbol{x}_i\}$ 变换为 $\{\boldsymbol{y}_i\}$。另一方面，利用音高映射函数得到的新的音高曲线，得到脉冲激励信号 $s(n)$。$\{\boldsymbol{y}_i\}$ 与 $s(n)$ 输入到 mel 对数谱逼近数字滤波器（MLSADF，Mel Log Spectrum Approximation Digital Filter）[17]中，重建出浊音段。重建的清音段和浊音段构成音节，进而可合成出语句。

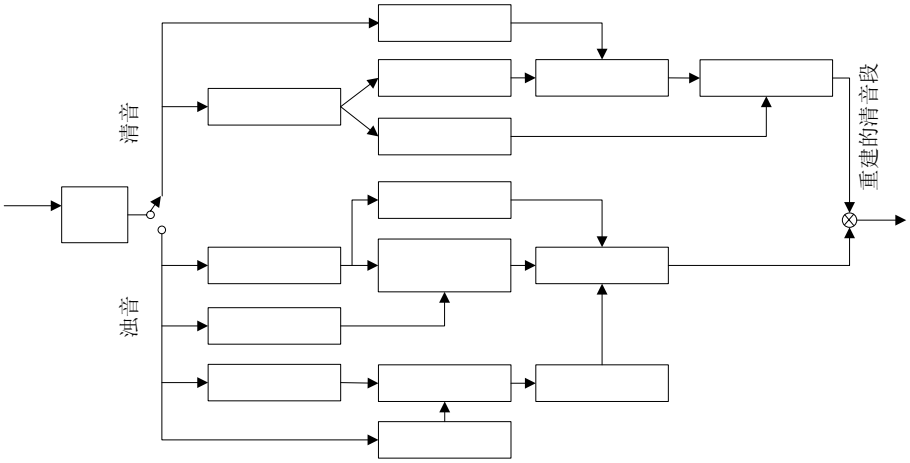

图 5 基于 HPCM 的声音变换算法的框图

3 实验

两个男声的音库（记为 M1 和 M2）和女声的音库（记为 F1 和 F2）取自清华大学计算机科学与技术系搜集的语音语料库。这些语料库是在极低环境噪声下，利用高质量麦克风录制的。四个发音人的录音文本完全相同，均包括汉语的 1268 个不同的有调音节。从这些音节中，随机抽取 1218 个音节作为训练集，余下的 50 个音节作为测试集。在 HPCM 训练过程中，清浊音边界的标定采用自动和人工相结合的方式，而音高曲线采用自动方式提取。对于清音，设定线性预测阶数为 16，噪声源参数估计时的高斯混合数为 2，噪声源模型训练时的高斯混合数为 32。对于浊音，加权 mel 倒谱的分析阶数为 26，高斯模型混合数为 64。时长、能量和音高模型的训练中，高斯混合数设为 32。

实验采用了 3 种不同的声音变换方法，分别是 Stylianou 提出的变换方法（记为 GMM）、基于本文提出的浊音模型和清音模型的变换方法（记为 AM）和基于 HPCM 的变换方法（记为 HPCM，即 AM 加上韵律补偿模型的方法）。应用以上 3 种方法完成模型训练后，分别对测试集内 50 个汉语音节进行变换，得到 150 段声音数据。实验进行了 M1 到 M2、M1 到 F1、F2 到 M2 和 F1 到 F2 等四对说话人的声音变换，总共得到 600 段重建语音。

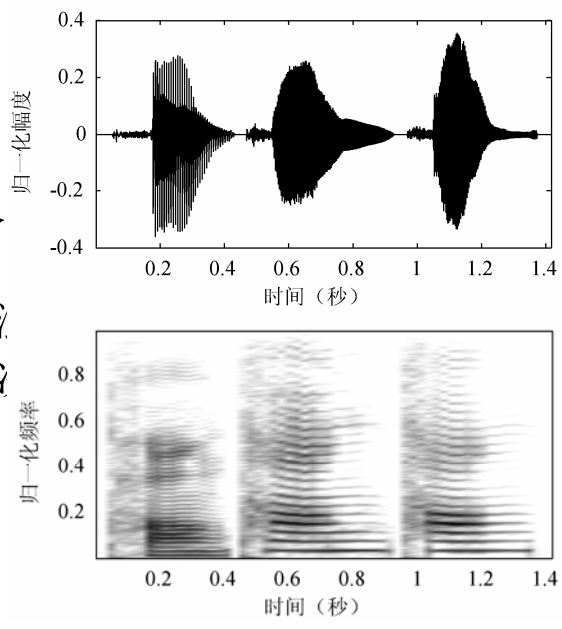

图 6 基于 HPCM 的声音变换的实例

图 6 所示的是基于 HPCM 的方法对音节“汤 tang1”的变换结果。其中波形所示的第一段是 M1 的语音，第二段是 F1 的语音，第三段则是重建语音。可以看出，重建语音的共振峰结构和 F1 的共振峰结构几乎一致。

3.1 客观评价

对于一段源说话人和目标说话人的语音以及一段重建语音，采用 mel 倒谱距离缩小比（CDRR，Cepstrum Distance Reduction Ratio）[18]作为客观度量标准，定义如式（7）。

$$CDRR = \left[1 - \frac{D\left(\hat{C}^t, C^t\right)}{D\left(C^s, C^t\right)}\right] \times 100(\%) \qquad (7)$$

其中 $\hat{C}^t$、C^s 和 C^t 分别是重建语音、源说话人语音和目标说话人语音的 mel 倒谱矢量集。$D(\cdot,\cdot)$ 是平均 mel 倒谱距离。*CDRR* 越大，重建语音听感上越接近目标说话人的语音。在客观评价过程中，不需要进行清浊音判决，但利用了 DTW 算法对齐 mel 倒谱矢量。

浊音声学模型采用了 mel 倒谱系数和线谱对系数作为描述短时语音的频谱参数。二者均表征了声道谱的特性，mel 倒谱系数反映了语音感知的非线性频域特性，LSF 则更好地刻画了短时语音的共振峰特征。但我们知道，mel 倒谱系数是通过求解优化问题得到的结果。它和共振峰结构不存在直接的关系，但可以用来重建出较高质量的语音。为了验证二组系数合用能够提高浊音声学模型的性能，我们进行了一个对比实验。表 1 为该实验的客观评价结果。可以看出，MCC+LSF 作为组合系数建立的 HPCM，可以取得更好的变换效果。

表 1　HPCM 对浊音段进行变换的客观评价

发音人	浊音段 *CDRR* (%)	
	MCC	MCC+LSF
M1→M2	31.87	**32.34**
M1→F1	43.48	**45.37**
F2→M2	37.65	**39.23**
F1→F2	27.16	**27.99**

表 2 所示的是三种声音变换方法对清浊音段变换后，产生的客观评价结果。它说明，和 GMM 方法相比，基于清浊声学模型的 AM 方法更能减小重建语音和目标语音的频谱差异。加入韵律补偿模型后，HPCM 在浊音段进一步减小了重建语音和目标语音的频谱距离。

另外，从表 2 中还可以看出，同性语音变换的 *CDRR* 比异性语音变换的小。这是由于异性语音之间的 比同性语音之间的大。由于异性语音之间的韵律特性差异较大，和 M1→M2、F1→F2 相比，在 M1→F1、F2→M2 中，HPCM 的韵律补偿发挥了更大的作用。因此，清浊音分别建模的 AM 方法优于 GMM 方法。在 AM 方法基础上增加韵律补偿模型的 HPCM 方法，表现出更好的性能。

表 2　三种变换方法的客观评价结果

发音人	*CDRR* (%)					
	GMM		AM		HPCM	
	清音	浊音	清音	浊音	清音	浊音
M1→M2	10.13	29.12	**13.05**	32.17	13.03	**32.58**
M1→F1	10.18	36.84	**14.19**	43.75	14.20	**45.32**
F2→M2	11.78	34.19	**13.92**	38.22	13.95	**39.15**
F1→F2	11.06	25.22	**13.73**	27.86	13.79	**28.06**

3.2 主观评价

为了从心理感知的角度评价重建语音的质量，实验采用 ABX 和平均意见得分（Mean Option Score，MOS）两种测试方式。10 位听音人参加了主观评价。

ABX 测试的目的是定量地评价声音变换算法改变说话人个性化特征的性能。首先将源说话人的语音、目标说话人的语音以及变换后得到的重建语音打乱次序，随机地播放给听音人，让其选出语音个性化特征最接近的两段。测试结束后，统计出平均正确响应率。响应率越高，说明重建语音的个性化特征越接近目标说话人的个性化特征。表 3 是四种方法的 ABX 测试结果。

表 3　ABX 测试结果

发音人	平均正确响应率(%)		
	GMM	AM	HPCM
M1→M2	72.6	77.3	**82.5**
M1→F1	100.0	100.0	**100.0**
F2→M2	100.0	100.0	**100.0**
F1→F2	70.6	74.9	**79.1**

表 3 说明了听音人易于判断重建语音的性别特征，表明了 AM 方法比 GMM 方法取得了更高的正确响应率，证实了清浊音分别建模更能准确地刻画说话人的个性化特征。另外，HPCM 方法进行了韵律特征的转换，使得重建语音和目标语音听感上更加接近。这也说明，语音的个性化特征差异不仅表现在频谱特征空间上，同样也表现在韵律特征空间上。

实验采用了主观意见得分作为评价重建语音的质量标准。让听音人听完重建语音后，给出意见分（5：优秀，4：良好，3：一般，2：较差，1：很差）。测试结束后，统计出平均意见得分。得分

越高，说明重建语音的清晰度与可懂度越好。图 7 所示的是四种方法获得的平均意见得分。

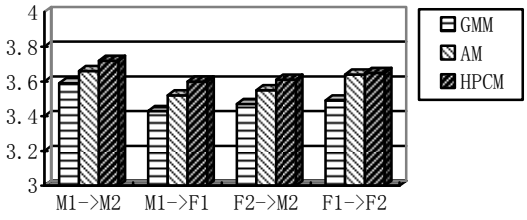

图 7 平均意见得分测试结果

从图 7 可以看出，男声之间的声音变换优于其他声音之间的变换。这是由于男声之间的声学特征与韵律特征的差异较小。图 7 也说明，清浊音和韵律特征的分别建模对于改善重建语音的质量有着比较明显的作用。

4 结论

在源滤波器模型的基础上，利用统计学习方法，建立了一个混合参数化特征模型。该模型刻画了说话人声学参数空间与韵律个性化特征空间的分布。鉴于清浊音的声学差异，采用了不同的方法分别对清浊音建立了声学模型。为了进一步提高重建语音的清晰度与可懂度，还对音高、能量与时长分别建立了基于统计学习的韵律补偿模型。清浊音的声学模型和韵律补偿模型综合为混合参数化特征模型。该模型的训练过程实际上模拟了人对语音个性化特征的学习机制。因此混合参数化特征模型不仅可以用来建立声音变换算法，还可以直接应用到语音个性化特征的分类和说话人的辨识上。在汉语音节的声音变换中该模型已经取得了较好的效果，而连续语流和非对齐文本情形下的变换问题尚待进一步研究。

参 考 文 献

1 Trask R L. 语音学和音系学词典. 北京：语文出版社，2000：259

2 张家騄. 超音段特征间的相互作用. 声学学报，1993；**18**(4)：263－271

3 Atal B S，Hanauer S L. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*，1971；**50** (2)：637－655

4 Childers D G，Yegnanarayana B，Wu K. Voice conversion: factors responsible for quality. In：Proc. IEEE ICASSP，1985：748－751

5 Childers D G，Wu K. Voice conversion. *Speech Communication*，1989；**8** (2)：147－158

6 Valbret H，Moulines E，Tubach J P. Voice transformation using PSOLA technique. *Speech Communication*，1992；**11**(2-3)：175－187

7 Abe M，Sagayama S. A voice conversion based on phoneme segment mapping. *J. Acoust. Soc. Japan (E)*， 1992；13 (3)：131－139

8 Arslan L M. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*，1999；**28**(3)：211－226

9 Lee K S，Won D，Youn D H. Voice conversion using low dimensional vector mapping. *IEICE Transactions on Information and Systems*，2002；**E85-D** (8)：1297－1305

10 Stylianou Y，Cappé O，Moulines E. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*，1998；**6** (2)：131－142

11 Mouchtaris A，der Spiegel J V，Mueller P. Non-parallel training for voice conversion by maximum likelihood constrained adaptation. In：Proc. IEEE ICASSP，2004：I1－I4

12 Chen Y，Chu M，et al. Voice conversion with smoothed GMM and MAP adaptation. In： Proc. Eurospeech，2003：2413－2416

13 左国玉，刘文举，阮晓钢. 基于遗传径向基神经网络的声音转换. 中文信息学报，2004；**18**(1)：78－84

14 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. 数据采集与处理，2005；**20**(2)：144－149

15 康永国，双志伟，陶建华等. 高斯混合模型和码本映射相结合的语音转换算法. 第八届全国人机语音通讯学术会议论文集，2005

16 吴宗济. 普通话元音和辅音的频谱分析及共振峰的测算. 声学学报，1964；**1**(1)：33－39

17 Tokuda K，Kobayashi T，Imai S. Adaptive cepstral analysis of speech. *IEEE Trans. on Speech and Audio Processing*，1995；**3**(6)：481－489

18 Iwahashi N， Sagisaka Y. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. *Speech Communication*，1995；**16**(2)：139－151