考虑帧间动态特征的音色变换算法

张晓洲[1]，黄德智[2]，蔡莲红[1]
(1. 清华大学 计算机科学与技术系 普适计算教育部重点实验室，北京 100084;
2. 北京法国电信研发中心有限公司，北京 100080)

摘 要：基于 Gauss 混合模型的音色变换算法在预测目标说话人频谱时会出现过平滑问题，导致声音转换结果的音质下降。该文分析了造成过平滑问题的原因，并提出一种考虑帧间动态特征的音色变换改进算法，在估计参数的目标函数中加入了连续性和方差的影响，从而改善了映射结果的帧间连续性，并使方差最大化，克服了过平滑现象。实验表明该算法在保证变换结果的目标倾向性的同时，能够使变换语音的音质主观意见得分由 3.11 提高到 3.89，证明动态特征对提高音色变换的音质有重要意义。

关键词：音色变换；高斯混合模型；动态特征；连续性；方差
中图分类号：TN912.3

Voice Conversion Considering Inter-frame Dynamic Features

ZHANG Xiaozhou[1]， HUANG Dezhi[2]，CAI Lianhong[1]
(1. Key Laboratory of Pervasive Computing, Ministry of Education; Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. France Telecom R&D Beijing Co., Ltd, Beijing 100080, China)

Abstract: In conventional Gaussian Mixture Model (GMM) based voice conversion system, speech quality of converted utterances is degraded due to the over-smoothing of predicted spectrum. A novel conversion method considering inter-frame dynamic features is proposed to alleviate the over-smoothing by taking account of continuity and variance in object function. As results, the predicted features are continuous and variance maximized in one syllable. Experimental results show that the method improves the opinion score of converted speech quality from 3.11 to 3.89, while effectively convert speaker's individuality, and prove that dynamic features are important to the quality of voice conversion.

Key words: voice conversion; GMM; dynamic feature; continuity; variance

音色变换是一种改变说话人声音特征的技术，它可以转换源说话人的声音，得到听感上接近目标说话人的重建语音，而保持源说话人语音信息内容不变[1]。音色变换可应用在文语转换系统上，实现合成语音的个性化。此外，在多语种合成系统中的音色归一化以及和谐人机语音交互方面都有很大的应用价值。

音色变换系统通常包括分析、变换和合成 3 个部分。分析部分从源语音中提取代表说话人个性的声学特征；变换部分通过事先建立的规则将源特征映射到目标特征；合成部分用目标特征合成目标语音。其中最为关键的是变换部分中声学特征映射规则的建立。现有的映射规则包括码本映射、Gauss 混合模型（GMM）[2]、人工神经网络、线性多变量回归、隐马尔可夫模型（HMM）等。基于 GMM 的音色变换用概率统计模型刻画说话人的声学特征空间，是目前最为成功的方法。但变换后的频谱存在过平滑问题，转换声音模糊发闷，导致语音音质下降。

在传统 GMM 变换规则的参数估计过程中，只考虑单帧源特征矢量变换后尽量逼近目标，而没有考虑帧间关联，估计出的协方差参数不能反映特征矢量的时序变化，导致过平滑现象。

本文提出一种考虑帧间连续性和方差等动态特征的参数估计算法，使得变换后一个音节内部的特征矢量集在保证连续性的同时具有足够的方差。

1 声学特征

说话人的声学特征与语音产生器官的生理情况密切相关，对说话人个性感知有重要贡献。音色变换在分析端从语音信号中提取声学特征，然后通过映射规则加以变换，最后在合成端重建出接近目标的语音。常见的分析合成方法如线性预测（LPC）滤波器，它将语音信号分解为残差和 LPC 谱。由于残差与原语音波形等长，必须通过建模来降低参数维数；其次还

收稿日期：2005-10-14
基金项目：国家自然科学基金资助项目（60275014）
作者简介：张晓洲（1983-），男（汉），浙江，硕士研究生。
通讯联系人：蔡莲红，教授，E-mail：clh-dcs@tsinghua.edu.cn

要考虑残差和 LPC 谱分别变换后的匹配问题。

STRAIGHT[3]是一种高质量的语音分析合成方法。它将语音信号分解为脉冲激励和光滑声道谱，声源部分仅包含基频信息，避免了声源特征和声道特征分别变换后的不匹配问题，并且在时长、基频等参数修改幅度较大时仍保持良好的重建音质，适合应用于音色变换中。STRAIGHT 分析得到的光滑声道谱是 FFT 功率谱，无法直接参与后续的训练过程。我们利用 Mel 倒谱分析[4]对其降维，把 40 阶 Mel 倒谱系数作为声学特征，刻画说话人的声道谱。非正式听测表明，40 阶 Mel 倒谱系数重建的语音与直接从光滑声道谱恢复的语音在音质上非常接近。

2 考虑帧间动态特征的音色变换算法

2.1 基于 GMM 的映射规则

Gauss 混合模型（GMM）用 m 个 Gauss 函数的线性组合来描述数据集的概率分布。概率密度函数为

$$P(\boldsymbol{x})=\sum_{i=1}^{m}\alpha_i N(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i). \qquad (1)$$

其中

$$N(\boldsymbol{x};\mu,\boldsymbol{\Sigma})=\frac{\left|\boldsymbol{\Sigma}\right|^{-1/2}}{(2\pi)^{p/2}}\mathrm{e}^{[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]}, \qquad (2)$$

$$\sum_{i=1}^{m}\alpha_i=1,\quad a_i\geq 0. \qquad (3)$$

$N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ 代表 p 阶矢量 $\boldsymbol{x}$ 在均值为 $\boldsymbol{\mu}$、协方差为 $\boldsymbol{\Sigma}$ 下的正态分布，α_i 是各个混合成分的权重。GMM 参数可以通过期望最大化（EM）算法估计。

基于 GMM 的音色变换方法认为声学特征在声学空间内的不同分布造成了说话人之间的差异，变换规则就是用 GMM 将空间分布参数化并构造线性映射函数。估计映射函数参数的方法有两种：最小二乘法（Least Squares Optimization）和联合概率法（Joint Density）。这两种方法性能相近，但联合概率法由于特征矢量维数增加了一倍，需要更多的训练数据保证 EM 算法的估计精度。最小二乘法的形式化描述如下：

给定 N 对对齐的源说话人和目标说话人语音特征矢量 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$，最小二乘法首先利用 EM 算法估计出源说话人连续概率空间的 m 组参数 $(\alpha_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$，每一组刻画了一类声学子空间的概率分布。根据 Bayes 准则，特征矢量 $\boldsymbol{x}$ 属于第 i 类声学子空间 C_i 的条件概率为

$$P(C_i\mid\boldsymbol{x})=\frac{\alpha_i\mid\boldsymbol{\Sigma}_i\mid^{-1}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}}{\sum_{j=1}^{m}\alpha_j\mid\boldsymbol{\Sigma}_j\mid^{-1}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)}}. \qquad (4)$$

定义映射函数为

$$\begin{aligned}F(\boldsymbol{x}_t)&=\sum_{i=1}^{m}P(C_i\mid\boldsymbol{x}_t)[\boldsymbol{\upsilon}_i+\boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)]\\&=F(\boldsymbol{x}_t,\boldsymbol{\upsilon}_1,\boldsymbol{\upsilon}_2,\cdots,\boldsymbol{\upsilon}_m,\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_2,\cdots,\boldsymbol{\Gamma}_m).\end{aligned} \qquad (5)$$

变换的目标函数为

$$\varepsilon=\sum_{t=1}^{N}\|F(\boldsymbol{x}_t)-\boldsymbol{y}_t\|^2. \qquad (6)$$

其中：$\boldsymbol{x}_t$、$\boldsymbol{y}_t$ 分别表示源矢量和目标矢量。最后，通过最小二乘法估计映射函数的参数 $\boldsymbol{\upsilon}$ 和 $\boldsymbol{\Gamma}$，使目标函数最小化。

基于 GMM 的音色变换方法在转换语音的目标倾向性上较为成功，但由于变换后的频谱会出现过平滑现象，导致音质下降。

2.2 过平滑问题的分析

频谱的过平滑现象包括两个方面：一是单帧的短时谱细节丢失；二是邻近帧的短时谱相似性过大，缺乏变化。这些都不符合真实语音的短时谱特性，造成听感上的音质下降。

Toda[5]认为统计模型的平均效应造成过平滑；Chen[6]认为过平滑的原因在于 GMM 估计的协方差矩阵大部分都是很小的值，导致变换后的特征矢量都靠近加权类中心，缺乏变化；Toda[7]还指出 GMM 变换后的特征向量在整个句子内的全局方差要远小于源和目标句子的全局方差。本文认为在传统的 GMM 变换规则的参数估计过程中，只考虑了单帧源特征矢量变换后尽量逼近目标，而没有考虑帧间关联，使得估计出的协方差参数不能反映特征矢量的时序变化，导致过平滑现象。

为克服过平滑现象，我们改进映射函数参数的估计算法，考虑语音的帧间连续性和方差等动态特征，使得变换后的声学特征在一个音节内部既保证连续性，同时又具有足够的方差，符合真实语音的特性。对动态特征的考虑体现在代价函数里而不是与静态特征一起组成高维特征，可以避免训练数据维数过高而导致 GMM 训练结果不稳定。

2.3 考虑连续性和方差的 GMM 参数估计

对一个音节内部的特征矢量集$(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k)$，反映矢量连续变化的程度如下：

$$D=\sum_{t=2}^{k}\|\boldsymbol{x}_t-\boldsymbol{x}_{t-1}\|^2. \tag{7}$$

矢量集的方差为：

$$V=\sum_{t=1}^{k}\|\boldsymbol{x}_t-\boldsymbol{x}_{\text{mean}}\|^2=\sum_{t=1}^{k}\|\boldsymbol{x}_t-\frac{1}{k}\sum_{\tau=1}^{k}\boldsymbol{x}_\tau\|^2. \tag{8}$$

反映了矢量集整体变化的程度。

如果变换后的矢量在连续变化的同时具有足够的方差，就可以改善过平滑现象。假设训练数据有 N 对特征矢量，M 对音节，定义考虑连续性和方差的目标函数为：

$$\begin{aligned}\Delta&=w_1\sum_{t=1}^{N}\|F(\boldsymbol{x}_t,\boldsymbol{v},\boldsymbol{\Gamma})-\boldsymbol{y}_t\|^2\\&+w_2\sum_{k=1}^{M}\sum_{t=2}^{M_k}\|F(\boldsymbol{x}_t,\boldsymbol{v},\boldsymbol{\Gamma})-F(\boldsymbol{x}_{t-1},\boldsymbol{v},\boldsymbol{\Gamma})\|^2\\&-w_3\sum_{k=1}^{M}\sum_{t=1}^{M_k}\|F(\boldsymbol{x}_t,\boldsymbol{v},\boldsymbol{\Gamma})-\frac{1}{M_k}\sum_{\tau=1}^{M_k}F(\boldsymbol{x}_\tau,\boldsymbol{v},\boldsymbol{\Gamma})\|^2,\\&w_1+w_2+w_3=1,\ w_1\ge 0,\ w_2\ge 0,\ w_3\ge 0.\end{aligned} \tag{9}$$

式（9）由三个子目标组成：第一部分是单帧矢量逼近，第二部分是音节内部矢量集的连续性，第三部分是整体方差，w_1、w_2 和 w_3 表示各部分权重。GMM 参数估计问题就转换为目标函数的最小化问题。

2.4 目标函数最小化

将映射函数式（5）改写为线性变换形式如下：

$$\begin{aligned}F(\boldsymbol{x}_t)&=F(\boldsymbol{x}_t,\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m,\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_2,\cdots,\boldsymbol{\Gamma}_m)\\&=\sum_{i=1}^{m}P(C_i\mid\boldsymbol{x}_t)[\boldsymbol{v}_i+\boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)]\\&=\sum_{i=1}^{m}P(C_i\mid\boldsymbol{x}_t)[\boldsymbol{A}_i\boldsymbol{x}_t+\boldsymbol{B}_i].\end{aligned} \tag{10}$$

其中：$\boldsymbol{A}_i=\boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{-1}$，$\boldsymbol{B}_i=\boldsymbol{v}_i-\boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$，成为映射函数的待定参数。进一步，可以将式（10）改写为

$$\begin{aligned}&F(\boldsymbol{x}_t,\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_m,\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_m)^{\mathrm{T}}\\&=\sum_{i=1}^{m}[P(C_i\mid\boldsymbol{x}_t)\ \ P(Ci\mid\boldsymbol{x}_t)\boldsymbol{x}_t^{\mathrm{T}}]\begin{bmatrix}\boldsymbol{B}_i^{\mathrm{T}}\\\boldsymbol{A}_i^{\mathrm{T}}\end{bmatrix}\\&=\boldsymbol{e}_t^{\mathrm{T}}\bullet\boldsymbol{J}.\end{aligned} \tag{11}$$

其中：

$$\boldsymbol{e}_t=\begin{bmatrix}P(C_1\mid\boldsymbol{x}_t)\\\vdots\\P(C_m\mid\boldsymbol{x}_t)\\P(C_1\mid\boldsymbol{x}_t)\boldsymbol{x}_t^{\mathrm{T}}\\\vdots\\P(C_m\mid\boldsymbol{x}_t)\boldsymbol{x}_t^{\mathrm{T}}\end{bmatrix},\quad\boldsymbol{J}=\begin{bmatrix}\boldsymbol{B}_1^{\mathrm{T}}\\\vdots\\\boldsymbol{B}_m^{\mathrm{T}}\\\boldsymbol{A}_1^{\mathrm{T}}\\\vdots\\\boldsymbol{A}_m^{\mathrm{T}}\end{bmatrix}. \tag{12}$$

$\boldsymbol{e}_t$ 是由 $\boldsymbol{x}_t$ 确定的行向量，$\boldsymbol{J}$ 是由待定参数组成的待定矩阵。

因此，考虑连续性和方差的目标函数式（9）就变成：

$$\begin{aligned}\Delta&=w_1\sum_{t=1}^{N}\|\boldsymbol{e}_t^{\mathrm{T}}\cdot\boldsymbol{J}-\boldsymbol{y}_t^{\mathrm{T}}\|^2\\&+w_2\sum_{k=1}^{M}\sum_{t=2}^{M_k}\|(\boldsymbol{e}_t^{\mathrm{T}}-\boldsymbol{e}_{t-1}^{\mathrm{T}})\cdot\boldsymbol{J}\|^2\\&-w_3\sum_{k=1}^{M}\sum_{t=1}^{M_k}\|\boldsymbol{e}_t^{\mathrm{T}}\cdot\boldsymbol{J}-\frac{1}{M_k}\sum_{\tau=1}^{M_k}\boldsymbol{e}_\tau^{\mathrm{T}}\cdot\boldsymbol{J}\|^2.\end{aligned} \tag{13}$$

目标函数的最小值出现在导数零点。对式（13）求导并求零点 $\boldsymbol{J}$ 即得到 GMM 映射函数的参数。

3 实验及分析

3.1 实验条件

选取两个男声音库（M1 和 M2）和两个女声音库（F1 和 F2）进行 M1 到 M2（MM）、M1 到 F1（MF）、F2 到 M2（FM）和 F1 到 F2（FF）等 4 组实验。音库均取自清华大学计算机系搜集的语音语料库。4 个发音人的录音文本完全相同，覆盖了汉语 450 个不同有调音节。每组实验选取 180 句约 2 万个特征矢量作为训练数据，5 句作为测试数据。实验对传统的 GMM 参数训练方法（GMM）和考虑动态特征的参数训练方法（GMM-DF）从客观和主观两个方面进行了比较。

3.2 客观评价

采用 Mel 倒谱距离缩小比（CDRR）作为客观度量标准，定义为

$$\text{CDRR}=\left[1-\frac{D(\hat{\boldsymbol{C}}^t,\boldsymbol{C}^t)}{D(\boldsymbol{C}^s,\boldsymbol{C}^t)}\right]\times 100\%. \tag{14}$$

其中：$\hat{\boldsymbol{C}}^t$、$\boldsymbol{C}^s$ 和 $\boldsymbol{C}^t$ 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。表 1 是平均客观评价结果。

表 1　CDRR 客观评价结果

权重			CDRR / %			
w_1	w_2	w_3	MM	MF	FM	FF
1.0	0.0	0.0	42.0	50.0	55.3	34.1
0.8	0.1	0.1	40.2	49.1	53.7	31.2
0.5	0.3	0.2	42.8	51.2	56.1	35.9
0.3	0.5	0.2	43.1	52.7	56.6	36.0
0.2	0.7	0.1	43.7	55.4	58.3	37.5

表 1 中 w_1 取 1，w_2 和 w_3 均取 0 时相当于传统 GMM 方法，其余 4 种情况均采用了 GMM-DF 方法。从表 1 可以看出 GMM-DF 的客观评价结果要优于 GMM，重建语音更接近目标说话人。当 w_1 取 0.2，w_2 取 0.7，w_3 取 0.1 时 CDRR 最大，重建语音最接近目标说话人。

图 1 给出了 GMM 和 GMM-DF 声学特征转换结果在音节内部的方差比较。GMM-DF 能有效改善过平滑问题，使转换后的声学特征各维具有足够的方差。

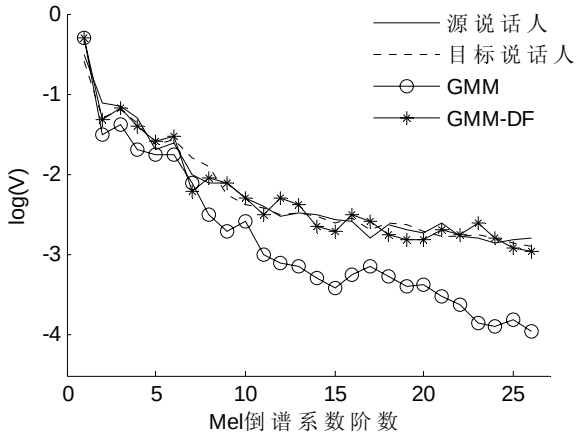

图 1　声学特征方差比较

图 2 显示了一帧短时谱的比较。可见 GMM-DF 转换的频谱细节更丰富，更接近目标谱。

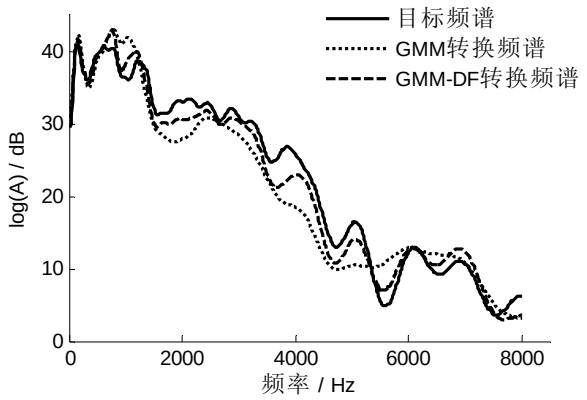

图 2　短时谱结果比较

3.3 主观评价

主观评价采用 ABX 和音质主观意见得分两种方式。参与测试的都是长期从事语音处理的研究人员。

ABX 测试定量地评价音色变换算法改变说话人个性特征的性能。将源语音、目标语音和变换语音随机播放给听音人，让其选出语音个性特征最接近的两段。测试结束后，统计出平均正确率。正确率越高，说明重建语音的个性特征越接近目标说话人的个性特征。表 2 是 4 组实验的 ABX 测试结果。

表 2　ABX 测试结果

实验组别	平均正确率 / %	
	GMM	GMM-DF
MM	70.3	75.1
MF	100.0	100.0
FM	100.0	100.0
FF	73.7	78.8

表 2 说明 GMM-DF 比 GMM 具有更高的变换准确率，证明动态特征是说话人个性中的重要成分。

实验采用了平均主观意见得分（MOS）作为评价重建语音的音质标准。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。MOS 越高，说明重建语音的清晰度与可懂度越好。GMM 和 GMM-DF 的 MOS 结果如图 3 所示。

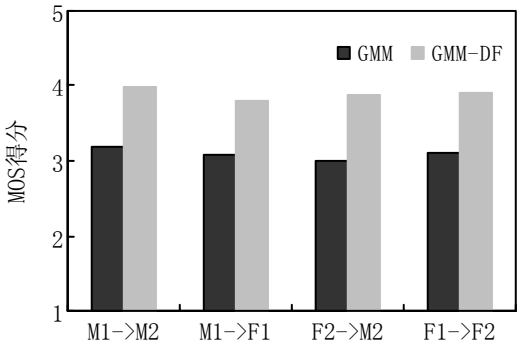

图 3　音质平均意见得分

从图 3 可以看出 GMM-DF 的主观意见平均得分由 3.11 提高到 3.89，音质优于传统的 GMM。由于 GMM-DF 克服了过平滑现象，改善了音质，因此对 ABX 测试也有一定正面影响，使其在 ABX 测试中得分也较高。

4 总结

传统的基于 GMM 的音色变换算法在预测目标说

话人声学特征时会出现频谱过平滑问题。本文分析了过平滑问题的原因，并提出了一种考虑帧间连续性和方差等动态特征的映射函数参数估计方法，使得转换结果不仅逼近目标特征，而且在音节内部具有帧间连续性和方差最大化。实验结果表明算法在保证变换语音目标倾向性的同时，能够有效提高变换语音的音质。

参 考 文 献 (References)

[1] 左国玉, 刘文举, 阮晓钢. 声音转换技术的研究与进展[J]. 电子学报, 2004, 32(7): 1165-1172.
ZUO Guoyu, LIU Wenju, RUAN Xiaogang. Voice conversion technology and its development [J]. *Acta Electronica sinica*. 2004, 32(7): 1165-1172. (in Chinese)

[2] Stylianou Y, Cappe O, Moulines E. Continuous probabilistic transform for voice conversion [J]. *IEEE Trans Speech and Audio Proc*, 1998. 6: 131-142.

[3] Kawahara H, Masuda-katsuse I, De Cheveign A. Restructuring speech representations using a pitchadaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction: possible role of a repetitive structure in sounds [J]. *Speech Communication*, 1999. 27: 187-207.

[4] 牟晓隆, 胡起秀, 吴文虎. 与文本无关的复合策略说话人辨识系统[J]. 清华大学学报, 1997, 37(3): 16-19.
MOU Xiaolong, HU Qixiu, WU Wenhu. Text-independent speaker identification system based on multiple strategies [J]. *J Tsinghua Univ*, 1997, 37(3): 16-19. (in Chinese)

[5] Toda T, Saruwatari H, Shikano K. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum [C]. *Proc ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001. 2: 841-844.

[6] Chen Yining, Chu Min, Chang Eric, et al, Voice conversion with smoothed GMM and MAP adaptation [C]. *Proc Eurospeech*, Geneva, Switzerland, 2003. 2413-2416.

[7] Toda T, Black A W, Tokuda K. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter [C]. *Proc ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, USA, 2005. 9-12.