基于维特比算法的语声转换

简志华[1,2],杨 震[2]
(1. 杭州电子科技大学通信工程学院,浙江杭州 310018; 2. 南京邮电大学通信与信息工程学院,江苏南京 210003)

摘 要: 本文提出了一种基于 Viterbi 搜索的语声转换算法,利用目标语音帧的转移概率矩阵来描述语音帧的时序信息,通过 Viterbi 搜索算法来寻找每帧语音的最佳 GMM 分量,它避免了传统的基于 GMM 的语声转换算法因丢失语音帧的时序信息所造成的频谱帧间不连续,同时也减少了因加权求平均所带来的语音频谱过于平滑,增强了共振峰. 客观评测和主观测试的实验结果都表明,本文算法改善了传统的基于 GMM 的语声转换算法的性能.

关键词: 语音处理; 语声转换; 韵律转换; 维特比算法
中图分类号: TN911.23 **文献标识码**: A **文章编号**: 0372-2112 (2009) 07-1470-06

A Method for Voice Conversion Based on Viterbi Algorithm

JIAN Zhi-hua[1,2], YANG Zhen[2]
(1. *School of Communication Engineering, HangZhou DianZi University, Hangzhou, Zhejiang* 310018, *China*;
2. *School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*)

Abstract: A novel method for voice conversion based on Viterbi algorithm is proposed in this paper. This method uses the matrix of transition probabilities of the target speaker's frames to represent the timing information of the speech sequence, and then determines the most appropriate component of the GMM by utilizing the Viterbi algorithm for converting each frame of the source speech. It avoids the spectral discontinuities caused by losing the relationship between the adjacent speech frames, and alleviates the spectral smoothing due to the weighted averaging in the traditional GMM-based algorithms and then enhances the formant. Both objective and subjective evaluation's results have demonstrated that the proposed method improves the performance of the conventional voice conversion system based on GMM.

Key words: speech processing; voice conversion; prosody transformation; Viterbi algorithm

1 引言

语音是人类交流和通信的重要手段. 语音信号包含了丰富的信息,既有语义信息,也包含了说话人个性特征、情感和态度等信息. 语声转换就是改变语音信号中源说话人(source speaker)的个性特征,使之具有目标说话人(target speaker)的身份特征,从而使得语音在经转换之后听起来就像是目标说话人的声音一样,而其中的语义信息并没有改变[1]. 语声转换技术的应用领域非常广泛[2],比如,具有说话人个性化特征的文语转换(TTS)和语音合成系统[3,4]、数字多媒体娱乐和电脑游戏[5]、帮助发声器官和听觉器官受损的人提高语音质量[6,7],以及作为语音识别系统中的说话人自适应模块用以降低因说话人差异而给识别系统带来的影响[8]等.

语声转换通常包含两个步骤,即训练阶段和转换阶段. 在训练阶段,通过提取源说话人语音和目标说话人语音的特征参数,形成两个特征向量空间,继而在某种匹配准则下,寻找将源说话人特征向量空间映射到目标说话人特征向量空间的最优匹配函数. 而转换阶段,就是用所求得的匹配函数将源说话人特征参数转换成目标说话人的特征参数,使得合成出来的语音具有目标说话人的身份特征. 语音特征参数的选取和建立恰当的匹配函数是语声转换系统中两个关键问题. 声道传输函数是影响说话人个性特征的主要因素,因此,反映声道特性的参数是语声转换常用的特征参数,如共振峰频率[9]、LPC 系数[10]、线谱对(LSP)[11-13]等. 在本文的研究中,拟将 LSP 作为语音的谱特征参数,这是因为 LSP 参数与共振峰频率和共振峰带宽密切相关,能够很好地反映声道特性;同时,与其他特征参数相比,LSP 具有非常良好的插值性,并且在 LSP 向量的某个分量估计有误的情况下,也仅影响到与其相关的部分频谱段[12]. 另外,语音的韵律信息也是影响说话人特性的因素[14],其主

收稿日期:2006-09-04;修回日期:2009-04-13
基金项目:国家 863 高技术研究发展计划重点项目(No. 2006AA010102)

要包括基音轮廓、音素时长和能量等. 韵律信息属于超音段特征,容易受到社会因素和心理状况的影响,目前的语声转换主要针对基音周期的改变,但基音只是考虑到了浊音信号,不对清音做变换. 本文对基音周期的转换是通过修改残差信号中强激励脉冲(IoSE, Instants of Significant Excitation)之间的时间间隔来实现的,它无需进行清/浊音的判决,对清音和浊音都同样处理[16]. 特征参数之间的匹配函数对语声转换系统的性能有重大影响. 匹配函数经历了一个从离散形式到连续形式的过程,早期的匹配函数是基于矢量量化(VQ)码本[8,13],是一种离散的形式. 它将源特征空间和目标特征空间分别量化成一个码本,通过统计直方图,形成两个码本之间各码字的匹配概率矩阵,从而完成源特征空间到目标特征空间的映射. 但这种基于 VQ 码本的匹配形式由于码字的有限性,使得特征参数限制在一个有限的集合中,引起了特征参数的不连续,极大地降低了语音的质量. 之后,Stylianou 等人提出了一种基于高斯混合模型(GMM)的连续的匹配函数,依据最小均方误差准则来估计匹配函数中的参量,保持了特征参数空间的连续性,提高了合成语音的质量,并证明了基于 VQ 的匹配函数是其一种特殊形式[10]. 在这基础上,Kain 应用联合密度估计(Joint Density Estimation,JDE)方法,改善了匹配函数的性能[12]. 但不管是基于 GMM 的匹配函数还是基于 VQ 的匹配函数,都没有考虑语音帧间的时序信息,对各帧都是独立进行转换,引起了语音频谱帧间的动态失真. 而文献[15]指出,频谱的帧间动态变化失真比一般的频谱失真更加影响语音质量. 为此,本文在 GMM 匹配算法的基础上,提出了一种基于维特比(Viterbi)搜索的语声转换算法. 通过建立目标语音各帧在 GMM 各个分量之间的转移概率矩阵来表征语音帧的时序信息,然后利用 Viterbi 算法寻找最佳的状态路径,再完成特征参数的转换.

2　基于 GMM 的语声转换算法

分别提取对称的训练语音库(对称语音库是指源说话人和目标说话人发相同的音,语音内容相同,并尽可能保持较好的时间对齐)中的源说话人语音和目标说话人语音的各帧特征参数,设分别为 L 维 X 和 Y,并用 DTW 算法对 X 和 Y 的时间序列进行对齐,形成两组具有相同数目并一一对应的特征参数时间序列,设分别为$\{X_n, n=1,\cdots,N\}$和$\{Y_n, n=1,\cdots,N\}$,然后再将相应的 X_n 和 Y_n 拼接成一个 $2L$ 维特征向量[12],即:

$$Z_n=\begin{bmatrix}X_n\\Y_n\end{bmatrix},\quad n=1,\cdots,N \tag{1}$$

因此,就形成了一个新的向量空间$\{Z_n, n=1,\cdots,N\}$. 利用高斯混合模型(GMM)对空间$\{Z_n, n=1,\cdots,N\}$进行概率密度建模,用 EM 算法训练 GMM 的参数,设 Z_n 的概率密度函数为:

$$p(Z_n)=\sum_{i=1}^{M}\pi_i\cdot N(Z_n;\mu_i,\Sigma_i),\quad n=1,\cdots,N \tag{2}$$

其中 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:

$$\mu_i=\begin{bmatrix}\mu_{iX}\\\mu_{iY}\end{bmatrix} \tag{3}$$

$$\Sigma_i=\begin{bmatrix}\Sigma_{iXX} & \Sigma_{iXY}\\\Sigma_{iYX} & \Sigma_{iYY}\end{bmatrix} \tag{4}$$

这样也同时求得了 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情况下,若 X 已知,则在最小均方误差(MMSE)估计准则下,对 Y 的估计为[10]:

$$E[Y|X=X_n]=\mu_Y+\Sigma_{YX}\Sigma_{XX}^{-1}(X_n-\mu_X) \tag{5}$$

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. 则在高斯混合模型下,根据式(2)～(4),对 Y 的 MMSE 估计为[12]:

$$\begin{aligned}Y_n&=E[Y|X=X_n]\\&=\sum_{i=1}^{M}\pi_i[\mu_{iY}+\Sigma_{iYX}\Sigma_{iXX}^{-1}(X_n-\mu_{iX})],\quad n=1,\cdots,N\end{aligned} \tag{6}$$

式(6)即为源说话人特征参数映射到目标说话人特征参数的匹配函数,在转换阶段,使用该式对源说话人的特征参数进行转换,形成转换后的频谱特征参数.

3　基于 Viterbi 搜索的语声转换算法

从式(6)可以看出,上面对源说话人每帧特征参数的转换都是独立进行的,没有考虑帧与帧之间的时序信息,这容易引起语音帧间的频谱动态失真,降低了语音的质量. 本文对语音帧间时序信息的描述,将采用目标语音帧的转移概率矩阵. 而同时,式(6)对目标特征参数的估计是一种加权求平均的思想,这会使得转换后的语音的频谱过于平滑,影响了语音的清晰度[6]. 本文算法采用 Viterbi 搜索,寻找最佳的 GMM 分量来进行特征参数的转换,这种单一分量的转换函数就避免了加权求平均所带来的频谱过于平滑的问题.

3.1　转移概率矩阵

由式(2)～(4)可以得到目标说话人特征参数空间$\{Y_n, n=1,\cdots,N\}$的概率密度分布函数,为:

$$p(Y_n)=\sum_{i=1}^{M}\pi_i\cdot N(Y_n;\mu_{iY},\Sigma_{iYY}),\quad n=1,\cdots,N \tag{7}$$

通过式(7),利用最大后验概率(MAP)对空间$\{Y_n, n=1,\cdots,N\}$中的向量进行分类,总共为 M 类,即每个

GMM 分量为一类,记为 Ω_i, $i=1,\cdots,M$,则有:

$$Y_n \in \Omega_j \mid j=\arg\max_i\left[\frac{\pi_i\cdot N(Y_n;\mu_{iY},\Sigma_{iYY})}{\sum_{i=1}^{M}\pi_i\cdot N(Y_n;\mu_{iY},\Sigma_{iYY})}\right],\quad j=1,\cdots,M \tag{8}$$

则称 Y_n 处于状态 j. 同理,再对 Y_{n+1} 进行状态判决. 因此,我们可以建立如下的统计直方图矩阵 H:

$$\mathrm{H}=\begin{bmatrix} h(1,1), & h(1,2), & \cdots, h(1,M)\\ h(2,1), & \cdots & \cdots\\ \cdots & h(i,j), & \cdots\\ h(M,1), & \cdots, & h(M,M)\end{bmatrix} \tag{9}$$

其中 $h(i,j)$ 是指 Y 从状态 i 转移到状态 j 的次数. 对矩阵 H 的每一行进行归一化,则可得到目标说话人特征参数的状态转移概率矩阵 P:

$$\mathrm{P}=\begin{bmatrix} p(1,1), & p(1,2), & \cdots, p(1,M)\\ p(2,1), & \cdots & \cdots\\ \cdots & p(i,j) & \cdots\\ p(M,1), & \cdots, & p(M,M)\end{bmatrix} \tag{10}$$

其中 $p(i,j)$ 是特征参数从状态 i 转移到状态 j 的概率,且满足 $\sum_{j=1}^{M}p(i,j)=1$, $\forall i=1,\cdots,M$.

3.2 Viterbi 搜索

利用 Viterbi 算法进行搜索的目的是为需要转换的源语句的特征向量序列寻找最佳的状态路径,也即是确定最佳的 GMM 分量序列. 设需要转换的一个源语句的特征向量序列为 $C_X=[X_1,\cdots,X_n,\cdots,X_T]$,则问题可以表述为:

$$\{i_n^*, n=1,\cdots,T\}=\arg\max_{\{i_n, n=1,\cdots,T\}}[q(i_1,X_1)\,p(i_1,i_2)\,q(i_2,X_2)\cdot p(i_2,i_3)\cdots q(i_{T-1},X_{T-1})\,p(i_{T-1},i_T)\,q(i_T,X_T)] \tag{11}$$

其中 $\{i_n^*, n=1,\cdots,T\}$ 表示最佳的状态路径, i_n 表示 X_n 所处的状态, $p(i_{n-1},i_n)$ 为式(10)中的转移概率,而 $q(i_n,X_n)$ 是后验概率,为:

$$q(i_n,X_n)=p(\Omega_{i_n}\mid X_n)=\frac{\pi_{i_n}\cdot N(X_n;\mu_{i_nX},\Sigma_{i_nXX})}{\sum_{i}^{M}\pi_i\cdot N(X_n;\mu_{iX},\Sigma_{iXX})} \tag{12}$$

式(11)是最佳路径的优化问题,可以通过 Viterbi 搜索算法来解决,其算法流程表示如下:

(1) 初始化:

$$\delta_1(i)=q(i,X_1),\ 1\le i\le M \tag{13}$$

$$\varphi_1(i)=0 \tag{14}$$

(2) 递归计算:

$$\delta_n(j)=\max_{1\le i\le M}[\delta_{n-1}(i)\,p(i,j)]\,q(j,X_n),\ 2\le n\le T,\ 1\le j\le M \tag{15}$$

$$\varphi_n(j)=\arg\max_{1\le i\le M}[\delta_{n-1}(i)\,p(i,j)],\ 2\le n\le T,\ 1\le j\le M \tag{16}$$

(3) 递归结束:

$$\mathrm{P}^*=\max_{1\le i\le M}[\delta_T(i)] \tag{17}$$

$$i_T^*=\arg\max_{1\le i\le M}[\delta_T(i)] \tag{18}$$

(4) 状态路径回溯:

$$i_n^*=\varphi_{n+1}(i_{n+1}^*),\ n=T-1,T-2,\cdots,1 \tag{19}$$

Viterbi 算法的图解可用图 1 所示. $[i_1^*, i_2^*, \cdots, i_T^*]$ 就是源特征向量序列 $C_X=[X_1,\cdots,X_n,\cdots,X_T]$ 所对应的 GMM 最佳分量序列,则相应的目标特征向量的估计值为:

$$\hat{Y}_n=E[Y\mid X=X_n]=\mu_{i_n^*Y}+\Sigma_{i_nYX}^{*}\Sigma_{i_nXX}^{*-1}(X_n-\mu_{i_n^*X}) \tag{20}$$

式(20)即为本文算法所用的谱特征参数的匹配函数.

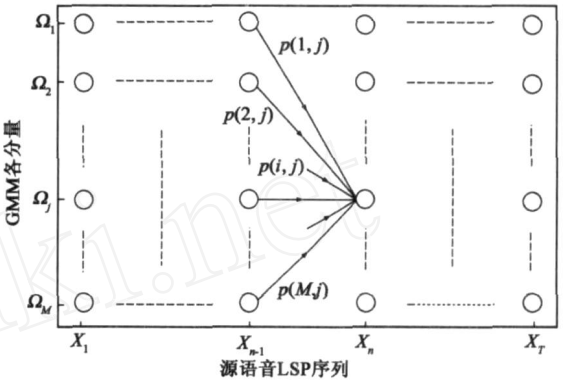

图1 Viterbi搜索示意图

4 韵律转换

在表征说话人个性特征的参数中,除了反映声道信息的谱包络特征参数外,语音信号的韵律特征也包含了丰富的说话人身份特征. 韵律特征参数主要包含基音轮廓、音素时长和能量等参数. 在本文的研究中,主要考虑源说话人和目标说话人之间的基音周期转换,而基音周期的转换是通过修改残差信号的 IoSE 之间的时间间隔来实现的,它无需进行清/浊音的判决,对清音和浊音都同样处理[16],其中,在浊音部分, IoSE 相当于基音信息,而在清音阶段, IoSE 反映了信号的突发时刻. 在以往的语声转换算法中,对韵律的转换往往只是针对浊音部分的基音周期,而忽略清音所包含的韵律信息. 采用 IoSE 作为韵律特征参数,则更充分考虑了语音信号的韵律信息,有利于语音韵律特性的转换. 残差信号是最小相位信号,利用最小相位信号的群时延平均斜率的特性,提取残差信号中的 IoSE,再计算出 IoSE 之间的时间间隔 Δ. 根据文献[17]的统计分析匹配思想,两者之间的匹配函数为:

$$\Delta^t=\mu_t+\frac{\sigma_t}{\sigma_s}(\Delta^s-\mu_s) \tag{21}$$

其中 Δ^s 和 Δ^t 是分别是源语音和目标语音的 IoSE 间隔, μ_s、μ_t 和 σ_s、σ_t 是其相应的均值和标准差. 利用式

(21)生成目标语音的 IoSE 序列,之后根据文献[16]的韵律转换算法,生成目标语音的残差信号.

5　实验与结果

5.1　语音库

本文实验所用的语音库是对称的语音库,由 500 个语句组成,长短不一,既有字和短语,也有长句,覆盖了大部分的汉语音节,由 4 个人发音,其中两个男声、两个女声,分别记为 M1、M2 和 F1、F2,每人每个语句都发音 3 次. 每个人都尽量以相同的朗读风格进行发音,以保持语音有比较好的时间对齐. 在信噪比不低于 30dB 的实验室环境下录制,信号抽样率为 16k,每个样点 16bit 量化. 语音库分成两部分,其中 380 个语句用于系统的训练,另外 120 个语句用于系统性能的测试. 实验分为客观评测和主观听觉测试,语声转换系统的整体框图如图 2 所示.

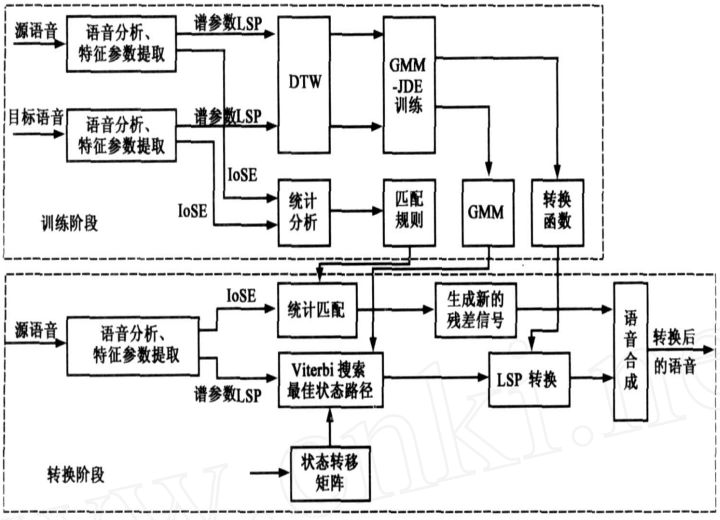

图2　语声转换系统框图

5.2　客观评测

实验采用 LSP 参数作为语音信号频谱信息的特征矢量,语音帧长为 20ms,帧移为 10ms,窗函数为 hamming 窗,GMM 的分量数 M 取 128. 另外,为了更好地和传统的基于 GMM 的语声转换算法相比较,两种语声转换系统对韵律的转换都采用上述基于 IoSE 韵律转换算法. 实验分为四个子任务,分别为男声到男声的转换(M1-to-M2)、男声到女声的转换(M2-to-F2)、女声到女声的转换(F2-to-F1)、女声到男声的转换(F1-to-M1). 图 3 是在 M1-to-M2 的情况下,一个源语句"社会主义"在转换前后和目标语音在语谱图方面的对照图. 从图上可以看出,基于 GMM 的算法和基于 Viterbi 搜索的算法对语声转换都有比较明显的效果,转换后的语谱图都更趋近于目标语音的语谱图. 但基于 Viterbi 搜索算法的语谱图在走势上要好于基于 GMM 算法的语谱图,也即降低了 GMM 算法的语音谱帧间动态失真,取得了较好的效果. 图 4 是在 M2-to-F2 的情况下,一帧语音(语音脚本为"可变电容器")的谱包络的对比图. 从图上可以看出,基于 GMM 算法的转换效果不佳,语音频谱被严重平滑,谱峰下降,而基于 Viterbi 算法的转换表现了较好的性能,语音的共振峰得到了强化. 这是因为基于 GMM 的算法对语音频谱的转换是采用的加权求平均的匹配函数,容易造成转换的语音频谱过于平滑,共振峰谱峰削弱,带宽拓展. 而基于 Viterbi 算法的语声转换对每一帧频谱特征参数都寻找最佳的 GMM 分量来进行转换,是单一分量的转换函数,避免了对转换函数进行加权处理. 另外,谱失真测度是一种常用的衡量频谱失真程度的方法,本文采用 Itakura 谱距离度量,定义[18]为:

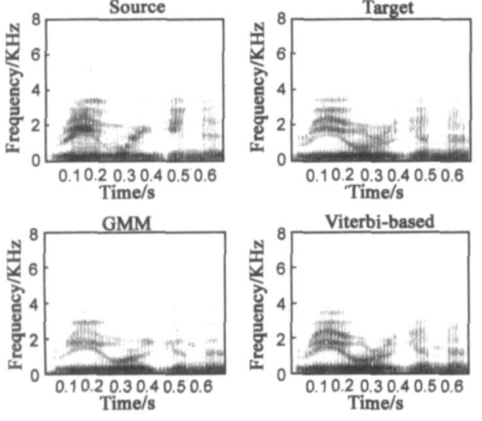

图3　语谱图对比(M1-to-M2)

$$d(s_1, s_2) = \log \frac{a_1 R_1 a_1^T}{a_2 R_2 a_2^T} \qquad (22)$$

因此,系统性能的优劣可以表示成如下的谱距离比值:

$$D = \frac{\sum_{n=1}^{N} d(con(n), tgt(n))}{\sum_{n=1}^{N} d(src(n), tgt(n))} \times 100\% \qquad (23)$$

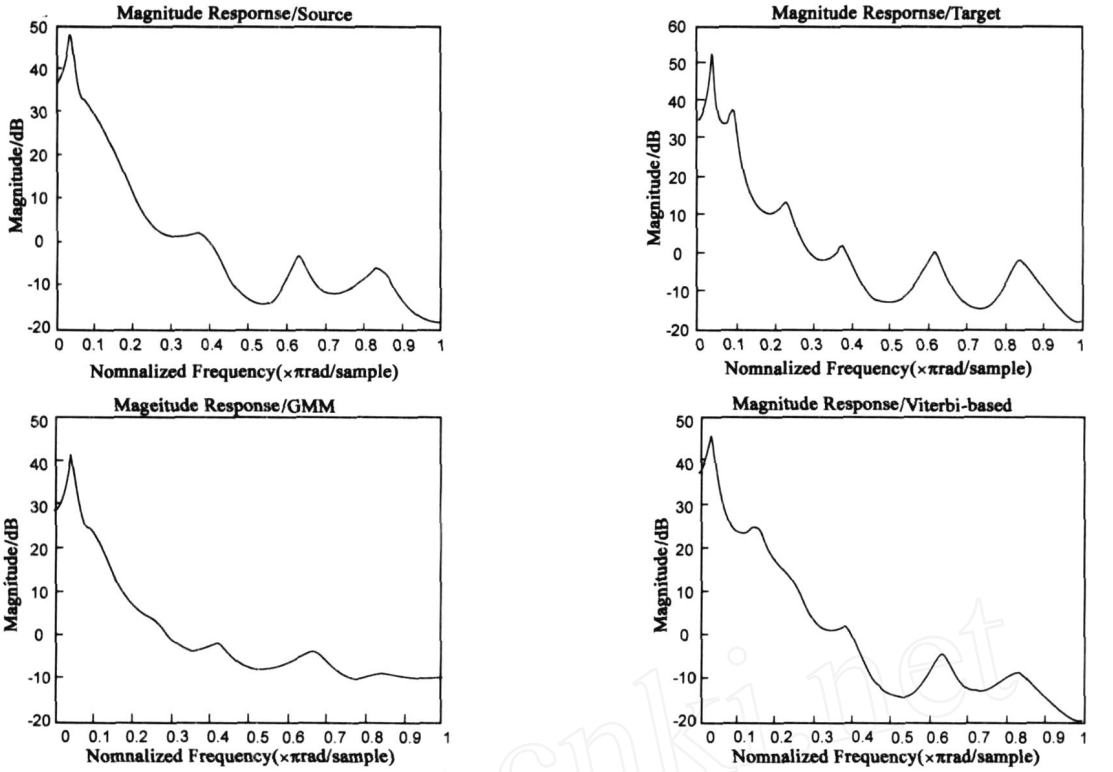

图4 频谱包络对比(M2-to-F2)

其中 N 是总共的语音帧数，con 表示转换后的语音. 当 D 越小，则转化后的语音谱越接近目标的语音谱，系统性能越好. 图 5 是一个语句在 M1-to-M2、M2-to-F2、F2-to-F1 和 F1-to-M1 四种情况下，在经过传统的基于 GMM 的算法和基于 Viterbi 搜索的算法转换后的 Itakura 谱距离比值的对照图. 从图上可以看出，不管是哪种情况，基于 Viterbi 搜索的转换算法都要优于基于 GMM 的转换算法. 同时还可以看出，异性之间的语声转换性能要优于同性之间的语声转换性能. 这是因为异性之间的特征参数空间距离更大，能转换过去的相对比值也会更大，效果也就更加明显一些. 图 6 是一个源语音在转换前后和目标语音的基音频率轨迹对比图，转换后的语音在基音频率轨迹上明显地趋近于目标语音.

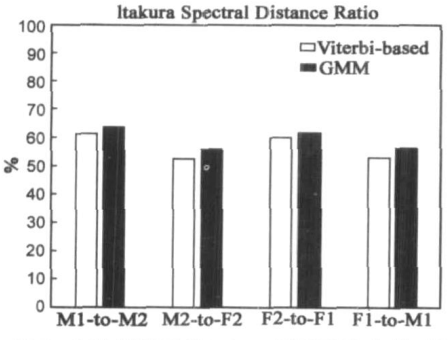

图5 几种情况下的Itakura谱距离比值的对照

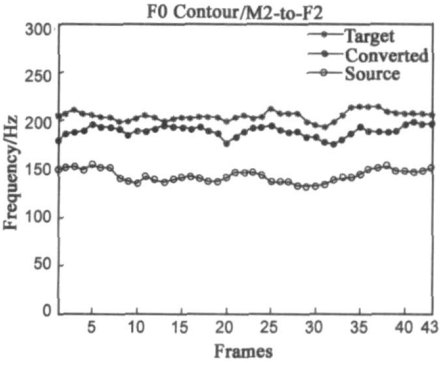

图6 基音频率轨迹对比

5.3 主观测试

主观听力感觉测试是对语音信号进行测试的一个重要组成部分. 在语声转换系统性能的测试中，ABX 测试法是一种常用的测试手段，它用来区分不同的说话人. A 和 B 分别表示源说话人语音和目标说话人语音，X 表示转换后的语音. 在实验测试中，要求受测者判断 X 更接近 A 还是更接近 B. 在本文实验中，分别要求 10 个受测者对转换后的语音做 ABX 测试，测试结果见表 1. 从测试结果来看，基于 Viterbi 搜索的语声转换算法要优于基于 GMM 的语声转换算法，同时，和客观评测实验结果相吻合的是，异性之间的转换效果要好于同性之间的转换.

表 1　几种情况下的 ABX 测试结果比较

	M1-TO-M2	M2-TO-F2	F2-TO-F1	F1-TO-M1
GMM	79.7 %	91.6 %	82.5 %	89.2 %
Viterbi-based	81.8 %	93.3 %	84.2 %	92.5 %

6　总结

本文提出了一种基于 Viterbi 搜索的语声转换算法,通过建立目标语音帧的转移概率矩阵来描述语音信号的帧间时序信息,继而利用 Viterbi 搜索算法,寻找各帧语音信号特征参数的最佳 GMM 分量,从而完成对语音信号的谱特征参数的转换,在韵律转换方面,是通过对残差信号中的强激励脉冲序列的统计分析匹配,从而达到对基音信息的改变. 实验结果表明,本文所提的语声转换算法,克服了基于 GMM 的传统语声转换算法所造成的频谱帧间动态失真的缺点,同时也改善了由于加权求平均所造成的语音频谱过于平滑的问题,使转换后的语音共振峰得到了加强.

参考文献:

[1] E Moulines, Y Sagisaka, Eds. Voice conversion: state of the art and perspectives[J]. Speech Communication, 1995, 16(2): 125-126.

[2] 左国玉, 刘文举, 阮晓刚. 声音转换技术的研究与进展[J]. 电子学报, 2004, 32(7): 1165 - 1172.
ZUO Guo-yu, LIU Wen-ju, RUAN Xiao-gang. Voice conversion technology and its development [J]. Acta Electronica Sinica, 2004, 32(7): 1165 - 1172(in Chinese).

[3] C H Wu, C C Hsia, T H Liu, J F Wang. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1109 - 1116.

[4] Chi-Chun Hsia, Chung-Hsien Wu, Jian-Qi Wu. Conversion function clustering and Selection using linguistic and spectral information for emotional voice conversion[J]. IEEE Transactions on Computers, 2007, 56(9): 1245 - 1254.

[5] Y Sato. Voice quality conversion using interactive evolution of prosodic control[J]. Applied Soft Computing, 2005, 5(2): 181 - 192.

[6] N Bi, Y Y Qi. Application of speech conversion to alarygeal speech enhancement[J]. IEEE Transactions on Speech and Audio Processing, 1997, 5(2): 97 - 105.

[7] C L Lee, W W Chang, Y C Chiang. Spectral and prosodic transformations of hearing-impaired mandarin speech [J]. Speech Communication, 2006, 48(2): 207 - 219.

[8] K Shikano, S Nakamura, M Abe. Speaker adaptation and voice conversion by codebook mapping[A]. IEEE Proceeding of ISCAS[C]. Singapore, 1991. 594 - 597.

[9] H Mizuno, M Abe. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt[J]. Speech Communication, 1995, 16(2): 153 - 164.

[10] Y Stylianou, O Cappe and E Moulines. Continuous probabilistic transform for voice conversion[J]. IEEE Transactions on Speech and Audio Processing, 1998, 6(2): 131 - 142.

[11] E Helander, J Nurminen, M Gabbouj. LSF mapping for voice conversion with very small training sets[A]. IEEE Proceeding of ICASSP2008[C]. Las Vegas, 2008. 4669 - 4672.

[12] A Kain. High resolution voice conversion [D]. Portland, Oregon: OGI School of Science and Engineering, Oregon Health and Science University, 2001.

[13] L M Arslan. Speaker transformation algorithm using segmental codebooks[J]. Speech Communication, 1999, 28(3): 211 - 226.

[14] D G Childers, B Yegnanarayana, K Wu. Voice conversion: factors responsible for quality[A]. IEEE Proceeding of ICASSP[C]. Florida, 1985. 748 - 751.

[15] H P Knagenhjelm, W B Kleijn. Spectral dynamics is more important than spectral distortion[A]. IEEE Proceeding of ICASSP[C]. Michigan, 1995. 732 - 735.

[16] K S Rao, B Yegnanarayana. Prosody modification using instants of significant excitation[J]. IEEE Transaction on Audio, Speech and Language, 2006, 14(3): 972 - 980.

[17] M M Hasan, A M Nasr and S Sultana. An approach to voice conversion using feature statistical mapping[J]. Applied Acoustics, 2005, 66(5): 513 - 532.

[18] L Rabiner, B H Juang. Fundaments of Speech Recognition [M]. Upper Saddle River, NJ, USA: Prentice Hall, Inc. 1993. 158 - 162.

作者简介:

简志华　男,1978 年 12 月生于江西新余. 分别于 2004 年和 2008 年获得南京邮电大学信号与信息处理专业硕士和博士学位,现为杭州电子科技大学通信工程学院讲师,主要研究方向有语音信号处理、语声转换和语音识别等.
E-mail: jianzh@hdu.edu.cn

杨　震　男,1961 年 11 月生于江苏苏州. 1999 年获得上海交通大学通信与信息系统专业工学博士学位,现为南京邮电大学教授、博士生导师,主要研究方向为语音信号处理、数字音频水印、认知无线电技术等.
E-mail: yangz@njupt.edu.cn