基于混合线性变换的语声转换算法

简志华 杨 震
(南京邮电大学信号与信息处理研究所 南京 210003)

摘 要：针对在没有对称语音库的情况下，该文提出了一种基于混合线性变换的语声转换算法，在最大似然估计准则下，使用 EM 迭代算法计算变换函数的参量。为了减小线性加权对语音谱包络的平滑作用，使用线性调频 Z 变换来调节语音信号的 LPC 系数。客观评测和主观感受的实验结果都表明，基于混合线性变换的语声转换算法也可以取得与传统语声转换技术相当的转换效果，解除了传统语声转换技术需要对称语音库的要求。

关键词：语声转换；混合线性变换；最大期望算法；线性调频 Z 变换
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2007)07-1700-03

An Algorithm for Voice Conversion Based on Mixtures of Linear Transformation

Jian Zhi-hua Yang Zhen
(*Institute of Signal and Information Processing, Nanjing Univ. of Post and Telecom.,Nanjing* 210003, *China*)

Abstract: This paper proposes an algorithm for voice conversion based on mixtures of linear transformation which avoids the need for parallel training corpus inherent in conventional approaches. In maximum likelihood framework, the EM algorithm is used to compute the parameters of the transfer function. And the chirp Z-transform is utilized to enhance the smoothed spectral envelop due to the linear weighted averaging. The proposed voice conversion system is evaluated using both objective and subjective measures. The experiment results demonstrate that the proposed approach is capable of effectively transforming speaker identity and can achieve comparable results of the conventional methods where a parallel corpus is needed.

Key words: Voice conversion; Ms-LT; EM algorithm; Chirp Z-transform

1 引言

语声转换就是改变语音信号中源说话人的个性特征，使之具有目标说话人的特性，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]。在语声转换技术的研究过程中，主要形成了以矢量量化码本[2,3]、人工神经网络[4,5]和高斯混合模型[6,7]为核心匹配算法的语声转换技术。在具有对称语音库的基础上，这些传统的语声转换算法通过某种匹配误差最小的优化准则，比如最小均方误差准则，来求取转换函数的参数，利用转换函数实现源说话人特征空间到目标说话人特征空间的映射。但有时候，很难或者根本不可能录制对称的语音库。这时，传统的语声转换技术就很难有比较好的转换效果，这也是现有的语声转换技术存在的一个重要的不足之处。针对没有对称语音库的情况，本文提出了一种基于混合线性变换(Mixtures of Linear Transformation, Ms-LT)的语声转换算法，它是以最大似然估计为优化准则，无需对称的训练语音库。同时，语音信号的韵律特征也包含了说话人许多个性化的特征信息，本文拟在提取残差信号中的强激励脉冲(Instants of Significant Excitation, IoSE)[8]的基础上，通过对IoSE序列的各相邻时间间隔进行统计分析匹配，从而实现韵律特征的转换。

本文结构安排如下：第 2 节先着重论述基于 Ms-LT 的谱包络转换算法的原理，接着，提出了用线性调频 Z 变换改善因加权求平均所导致的过于平滑的谱包络；第 3 节将讨论韵律的变换算法；实验的描述和结果安排在第 4 节；第 5 节对全文进行总结。

2 基于 Ms-LT 的谱包络转换

2.1 Ms-LT 基本原理

假定 $\boldsymbol{D}_{\boldsymbol{x}}=\{\boldsymbol{x}_n,n=1,\cdots,N_{\boldsymbol{x}}\}$ 表示源说话人特征向量空间，$\boldsymbol{x}$ 是定义在 $\boldsymbol{D}_{\boldsymbol{x}}$ 上的 h 维随机向量，利用 GMM 对空间 $\boldsymbol{D}_{\boldsymbol{x}}$ 进行概率密度建模，则有：

$$f_{\boldsymbol{x}}(\boldsymbol{x})=\sum_{i=1}^{M}p(\pi_i)N(\boldsymbol{x};\boldsymbol{\mu}_{\boldsymbol{i}},\boldsymbol{\Sigma}_{\boldsymbol{i}}) \tag{1}$$

其中 $\sum_{i=1}^{M}p(\pi_i)=1$，$\forall i,0<p(\pi_i)<1$，而 $\boldsymbol{\mu}_{\boldsymbol{i}}$ 和 $\boldsymbol{\Sigma}_{\boldsymbol{i}}$ 分别是第 i 个分量的均值向量和协方差矩阵。设 $\boldsymbol{y}$ 是定义在目标说话人特征向量空间 $\boldsymbol{D}_{\boldsymbol{y}}=\{\boldsymbol{y}_n,n=1,\cdots,N_{\boldsymbol{y}}\}$ 上的 h 维随机向量，假设 $\boldsymbol{y}$ 与 $\boldsymbol{x}$ 之间的映射关系用如下的混合线性变换来拟合，即：

2006-06-06 收到，2006-10-30 改回
江苏省青蓝工程项目(QL003YZ)资助课题

$$\boldsymbol{y}=\sum_{i=1}^{M}p(\pi_i)\sum_{k=1}^{K}p(\lambda_k\mid\pi_i)[\boldsymbol{A}_k\boldsymbol{x}+\boldsymbol{b}_k] \tag{2}$$

$\boldsymbol{A}_k$ 是 $h\times h$ 维变换矩阵，$\boldsymbol{b}_k$ 是 h 维偏移向量，$p(\lambda_k\mid\pi_i)$ 是第 k 个变换 λ_k 的先验概率，且满足 $\sum_{k=1}^{K}p(\lambda_k\mid\pi_i)=1,\ \forall i=1,\cdots,M$ 。参数集 $\theta=\{\boldsymbol{A}_k,\boldsymbol{b}_k,p(\lambda_k\mid\pi_i),\forall k=1,\cdots,K,i=1,\cdots,M\}$ 是本文所需要求解的未知参量。θ 的求解采用似然最大估计准则，即 $\theta^*=\arg\max_{\theta}\left(\log\left\{\prod_{n=1}^{N_y}f_y(\boldsymbol{y}_n\mid\theta)\right\}\right)$。则求取 θ^* 的迭代算法为[9]：

(1)分别对 $p(\lambda_k\mid\pi_i)$，$\boldsymbol{A}_k$ 和 $\boldsymbol{b}_k$ 进行初始化：$p(\lambda_k\mid\pi_i)=1/K$，$\boldsymbol{A}_k^{(0)}=\boldsymbol{I}$，$\boldsymbol{b}_k^{(0)}=0$，$\forall k=1,\cdots,K$，$\forall i=1,\cdots,M$，$\boldsymbol{I}$ 表示单位阵；

(2)求取中间变量 R_{ik}，$\bar{\boldsymbol{\mu}}_{ik}$ 和 $\bar{\boldsymbol{\Sigma}}_{ik}$；

(3)对 $p^{(l)}(\lambda_k\mid\pi_i)$，$\boldsymbol{A}_k^{(l)}$ 和 $\boldsymbol{b}_k^{(l)}$ 的值进行更新，求取下一步的值 $p^{(l+1)}(\lambda_k\mid\pi_i)$，$\boldsymbol{A}_k^{(l+1)}$ 和 $\boldsymbol{b}_k^{(l+1)}$；

(4)判断收敛条件 $\dfrac{L(\boldsymbol{y}\mid\theta^{(l+1)})-L(\boldsymbol{y}\mid\theta^{(l)})}{L(\boldsymbol{y}\mid\theta^{(l)})}\le\text{threshold}$ 是否满足，如果满足，则停止迭代，结束程序；否则返回第(2)步，重新迭代求解。

迭代程序结束后，将最后所求得的 $p^*(\lambda_k\mid\pi_i)$，$\boldsymbol{A}_k^*$ 和 $\boldsymbol{b}_k^*$ 代入式(2)中，就可以得到由源特征矢量变换成目标特征矢量的转换函数。

2.2 CZT 谱包络增强

从式(2)可以看出，$\boldsymbol{y}$ 是对 $\boldsymbol{x}$ 在经各个局部子空间变换后取加权平均的结果，而这样的加权平均会引起变换后语音的谱包络过于平滑，共振峰带宽拓展，谱峰下降，影响了语音的可懂度和清晰度。因此，为了改善语音质量，本文拟采用具有局部放大和细化作用的CZT增强谱包络[10]。在本文的语音谱包络增强实现中，对语音的第 i 阶LPC参数 a_i 乘以因子 c^{-i}，即 $\hat{a}_i=c^{-i}a_i$，在实验中，取 $c=0.98$。

3 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外，语音信号的韵律特征也包含了丰富的反映说话人身份特征的参数。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[11]。提取语音残差信号中的IoSE，再计算出IoSE之间的时间间隔 Δ。根据文献[12]的统计分析匹配思想，两者之间的匹配函数为

$$\Delta^t=\mu_t+\frac{\sigma_t}{\sigma_s}(\Delta^s-\mu_s) \tag{3}$$

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(3)生成目标语音的 IoSE 序列，之后根据文献[11]的韵律转换算法，生成目标语音的残差信号。

4 实验与结果

4.1 语音库

本文实验所用的语音库是参照了文献[7]的部分设计思想，由 250 个语句组成，由 3 个人发音，其中两个男声、一个女声，在信噪比不低于 30dB 的实验室环境下录制，信号抽样率为 16kHz，每个样点 16bit 量化。语音库分成 3 个部分，前 100 个句子组成第 1 部分，中间 100 个句子构成第 2 部分，用错开的第 1 部分和第 2 部分组成一个非对称的训练用语音库，最后 50 个句子用于系统性能的测试。

4.2 客观评测

本文采用线谱对(LSP)参数作为语音信号频谱信息的特征矢量，实验分为两个部分，即男声转换成男声(M-to-M)和女声转换成男声(F-to-M)。谱失真测度采用 Itakura 谱距离度量为

$$d(\text{src},\text{tgt})=\log\frac{\boldsymbol{a}_{\text{src}}\boldsymbol{R}_{\text{tgt}}\boldsymbol{a}_{\text{src}}^{\text{T}}}{\boldsymbol{a}_{\text{tgt}}\boldsymbol{R}_{\text{tgt}}\boldsymbol{a}_{\text{tgt}}^{\text{T}}} \tag{4}$$

其中 $\boldsymbol{a}_{\text{src}}$ 和 $\boldsymbol{a}_{\text{tgt}}$ 分别是源语音和目标语音的 LPC 系数构成的向量，$\boldsymbol{R}_{\text{tgt}}$ 是目标语音的相关系数矩阵。因此，系统性能的优劣可以表示成如下的谱距离比值：

$$D=\frac{\sum_{n=1}^{N}d(\text{con}(n),\text{tgt}(n))}{\sum_{n=1}^{N}d(\text{src}(n),\text{tgt}(n))}\times100\% \tag{5}$$

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 1 是基于GMM的联合密度估计(Joint Density Estimation, JDE)算法[7]和本文的Ms-LT算法之间的Itakura谱距离比值的对比图，分别对在M-to-M和F-to-M两种语声变换情况下进行了比较。从图上可以看出，在GMM的混合数 M 较小的情况下，Ms-LT的性能要优于GMM-JDE算法。这是因为用较少的 M 个子空间不能精确地描述整个特征矢量空间，而Ms-LT算法在每个子空间内使用了 K (在本实验中，取 K =6)个线性变换，这相当于在每个子空间内又进一步细

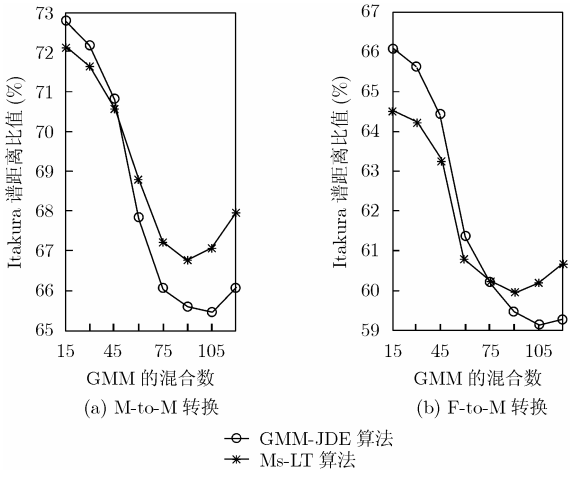

图 1 GMM 与 Ms-LT 之间的 Itakura 谱距离比值的对比

划分了 K 个小空间，使得 Ms-LT 比 GMM-JDE 更能够准确地刻画整个矢量空间。而随着 M 的增大，GMM-JDE 对空间的细化能力增强，其性能逐渐超越 Ms-LT，这是由 Ms-LT 的似然函数估计本身的缺陷所造成的。当 M 增大到一定的值时，Ms-LT 和 GMM-JDE 的性能都有一定程度的下降，这是因为 M 越大，要估计的参数越来越多，而训练数据有限，因此，训练数据的不足导致了所估计的参数准确性不足，系统性能下降。图 2 是 M-to-M 使用 Ms-LT 算法转换后，源说话人、目标说话人、转换后以及经 CZT 增强后的语音谱包络形状对比图。从图上可以看出，虽然转换后的语音共振峰和目标语音的共振峰存在一定的偏移，但与源语音的频谱相比，转换后的语音在谱包络形状上更接近于目标语音的谱包络，这说明本文的转换算法是非常有效的。而同时，转换后的谱包络被平滑了很多，在经 CZT 增强后，共振峰得到了增强，提高了语音的清晰度。图 3 是源说话人、目标说话人以及转换后语音的基音周期轨迹图，从图上来看，转换后的基音周期轨迹的走势趋近于目标说话人的基音周期轨迹，这在语音韵律的层面上，改善了语声转换的效果。

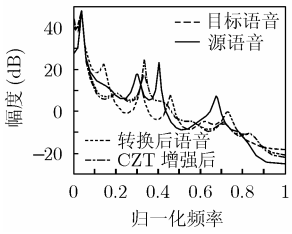

图 2 谱包络形状对比图(M-to-M)

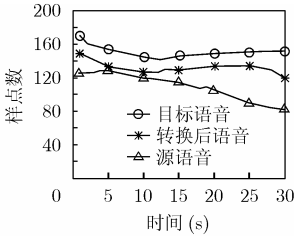

图 3 源、目标和转换后语音的基音周期轨迹(F-to-M)

4.3 主观评测

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 5 个受测者对 50 个转换后的语音做 ABX 测试，测试结果见表 1。从测试结果来看，F-to-M 的语音转换效果要明显高于 M-to-M 的语音转换效果，同时 Ms-LT 的转换效果要弱于 GMM-JDE，但差距不大，亦即在没有对称语音库的情况下，使用 Ms-LT 转换算法也能取得较为满意的效果。

表 1 ABX 测试结果对照表(M=90)

	GMM-JDE	Ms-LT
F-to-M	92.8%	89.6%
M-to-M	79.2%	74.4%

5 结束语

本文提出了一种基于混合线性变换的语声转换算法，它无需对称的语音库，在最大似然估计准则下，使用 EM 迭代算法计算变换函数的参量。为了减小线性加权对语音谱包络的平滑作用，使用线性调频 Z 变换来调节语音信号的 LPC 系数，增强共振峰。在韵律转换上，通过对语音残差信号中的强激励脉冲序列的统计分析匹配，从而达到对基音的修改。实验结果表明，Ms-LT 语声转换算法也可以取得与传统语声转换技术相当的转换效果，解除了传统语声转换技术对需要对称语音库的要求，极大地方便了用户的使用。

参 考 文 献

[1] Childers D G, Wu K, and Hicks D M, *et al*.. Voice conversion. *Speech Communication*, 1989, 8(2): 147–158.

[2] Abe M, Nakamura S, Shikano K, and Kuwabara H. Voice conversion through vector quantization. IEEE Proceedings of ICASSP, New York, USA, Apr. 11-14, 1988: 565–568.

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

[4] Narendranath M, Murthy H A, and Rajendran S, *et al*.. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 1995, 16(2): 207–216.

[5] Iwahashi N and 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.

[6] Stylianou Y, Cappe O, and Moulines E. Continuous Probabilistic Transform for Voice Conversion. *IEEE Trans on Speech and Audio Processing*, 1998, 6(2): 131–142.

[7] Kain A and Macon M W. Spectral voice conversion for text-to-speech synthesis. IEEE Proceedings of ICASSP, Seattle, USA, May 12-15, 1998: 285–288.

[8] Smits R and Yegnanarayana B. Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. on Speech and Audio Processing*, 1995, 3(5): 325–333.

[9] Diakoloukas V D and Digalakis V V. Maximum likelihood stochastic transformation adaptation of hidden Markov models. *IEEE Trans. on Speech and Audio Processing*, 1999, 7(2): 177–187.

[10] Wang T T. The segmented chirp z-transform and its application in spectrum analysis. *IEEE Trans. on Instrumentation and Measurement*, 1990, 39(2): 318–324.

[11] Rao K S and Yegnanarayana B. Prosody modification using instants of significant excitation. *IEEE Trans. on Audio, Speech and Language*, 2006, 14(3): 972–980.

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

简志华： 男，1978 年生，博士生，研究方向为语音信号处理、语声转换及语音识别.

杨 震： 男，1961 年生，教授，博士生导师，研究方向为语音信号处理、数字音频水印和无线通信技术.