一种用于语音转换的区域最近邻迭代训练算法

简志华* 王向文
(杭州电子科技大学通信工程学院 杭州 310018)

摘 要：针对非对称语音库情况下的语音转换，该文提出一种新的改进的语音转换训练算法 ILNCA。与原有的训练算法 INCA 不一样的是，ILNCA 首先利用高斯混合模型(GMM)分别对源、目标语音特征参数空间进行分类。然后根据 Kullback–Leibler(KL)距离最小原则对源、目标 GMM 模型的子空间进行匹配，最后利用最近邻准则在相对应的子空间中进行源、目标语音特征参数矢量的对齐。客观测试和主观听觉实验都表明由于该文算法采用了更加精确的矢量对齐方法，能取得比 INCA 算法更优异的转换性能。

关键词：语音转换；与文本无关；最近邻准则；迭代训练
中图分类号：TN911.23 **文献标识码**： A **文章编号**：1009-5896(2012)09-2091-06
DOI: 10.3724/SP.J.1146.2012.00398

An Iterative Training Algorithm Based on Local Nearest Neighbor for Voice Conversion

Jian Zhi-hua Wang Xiang-wen
(*School of Communication Engineering*, *Hangzhou DianZi University*, *Hangzhou* 310018, *China*)

Abstract: A novel algorithm named Iterative combination of a Local nearest Neighbor search step and a Conversion step Alignment (ILNCA), a modified version of the Iterative combination of a nearest Neighor search step and a Conversion step Alignment (INCA), is proposed for training voice conversion system under the situation of nonparallel corpus. Unlike INCA, ILNCA uses firstly Gaussian Mixture Model (GMM) to represent the spectral feature spaces of both source speaker and target speaker respectively, and then matches each individual Gaussian components of the GMM from source speaker to target speaker and vice versa according to Kullback-Leibler (KL) distance. Finally, ILNCA performs the frame alignment of phonetically equivalent acoustic vectors from source and target speaker in their mapped sub-spaces, not in the whole space like INCA. Both object and subject evaluations are conducted. The experimental results demonstrate that the approach can achieve better performance than INCA because of the accurate vector alignment.

Key words: Voice conversion; Text independent; Nearest neighbor; Iterative training

1 引言

语音转换的目的是要改变源说话人语音中的个性特征，使得转换后的语音听起来就像是目标说话人的声音一样，而其中的语义信息不变[1,2]。较早期的语音转换算法是基于矢量量化(Vector Quantization, VQ)模型[3]。但这种基于 VQ 的算法将特征参数矢量离散化，导致频谱的不连续性，转换性能和语音质量都不理想。之后，文献[4,5]等学者针对基于 VQ 转换算法的不足，提出了一种基于高斯混合模型(Gaussian Mixture Model, GMM)的具有连续形式的转换函数，具有较好的转换性能。为了进一步提高转换性能和语音质量，文献[6]提出了基于隐马尔科夫模型(Hidden Markov Model, HMM)的转换算法，用 HMM 中的状态变量来描述语音的帧间时序信息。

以上算法在训练阶段都需要有一个对称的语音库，即源说话人和目标说话人录制相同文本内容的语音。但在实际应用当中，有许多场合根本就不可能获得对称的语音库，为了满足这种应用需求，学术界做了许多有益的探索。文献[7]在原有的转换函数的基础上，采用自适应方法得到非对称语音库情况下的转换函数，但原有的转换函数是经过对称语音库训练得到的，在实际使用上也并不便利。文献[8]根据说话人识别的原理预先建立目标说话人统计模型，然后根据最大似然估计法循环迭代计算转换矩阵，从而使得转换后语音在目标说话人统计模型下具有最大的输出概率。该算法过于依赖目标说话人统计模型，一旦所建立的模型不准确，将对转换

2012-04-09 收到，2012-06-12 改回
浙江省教育厅科研项目(Y201016542)和浙江省自然科学基金项目(Y1101040)资助课题
*通信作者：简志华 jianzh@hdu.edu.cn

效果产生很大的影响。文献[9]根据最近邻的声学特征参数对应着相同的音素这一原理，提出了一种循环迭代的转换函数训练算法 INCA(Iterative combination of a nearest Neighbor search step and a Conversion step Alignment)。但对整个语音特征参数空间来讲，声学距离最近的特征参数未必对应着相同的音素，因此这就会影响转换函数的有效性。

针对 INCA 算法的不足，本文提出了一种基于 GMM 空间分类和 Kullback–Leibler(KL)距离的转换函数训练算法 ILNCA。

2 KL 距离矩阵

KL 散度(Kullback–Leibler divergence)可用来度量两个概率分布函数之间的相似度，在统计学、信息论等领域有广泛的应用。假设 $f(x)$ 和 $g(x)$ 分别表示两个连续分布的概率密度函数，则 KL 散度 $D_{\mathrm{KL}}\left(f\,||\,g\right)$ 定义为[10]

$$D_{\mathrm{KL}}\left(f\,||\,g\right)=\int_{-\infty}^{+\infty}f(x)\lg\frac{f(x)}{g(x)}\mathrm{d}x \tag{1}$$

对于任意的概率分布 $f(x)$ 和 $g(x)$ 来说，$D_{\mathrm{KL}}(f\,||\,g)\geq 0$；当且仅当 $f(x)=g(x)$ 时，$D_{\mathrm{KL}}\left(f\,||\,g\right)=0$。由于 KL 散度具有不对称性，即 $D_{\mathrm{KL}}(f\,||\,g)\neq D_{\mathrm{KL}}(g\,||\,f)$，所以它不具有距离的属性。为此，可将 KL 距离定义为

$$D(f,g)=\frac{1}{2}\left[D_{\mathrm{KL}}\left(f\,||\,g\right)+D_{\mathrm{KL}}\left(g\,||\,f\right)\right] \tag{2}$$

对于具有高斯分布的概率密度函数来讲，$D_{\mathrm{KL}}(f\,||\,g)$ 具有闭式的解，即当 $f(x)$ 和 $g(x)$ 都服从正态分布时，$D_{\mathrm{KL}}\left(f\,||\,g\right)$ 为[11]

$$\begin{aligned}D_{\mathrm{KL}}\left(f\,||\,g\right)=\frac{1}{2}\Bigg[&\lg\frac{\left|\boldsymbol{\Sigma}_g\right|}{\left|\boldsymbol{\Sigma}_f\right|}+\mathrm{Tr}\left[\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\Sigma}_f\right]-d\\&+\left(\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\right)^{\mathrm{T}}\boldsymbol{\Sigma}_g^{-1}\left(\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\right)\Bigg]\end{aligned} \tag{3}$$

其中 $\boldsymbol{\mu}_f$，$\boldsymbol{\mu}_g$ 和 $\boldsymbol{\Sigma}_f$，$\boldsymbol{\Sigma}_g$ 分别表示 $f(x)$ 和 $g(x)$ 中的均值向量和协方差矩阵，$\mathrm{Tr}[\cdot]$ 表示矩阵的迹，d 表示矢量的维数。

假定 $\boldsymbol{X}=\{\boldsymbol{x}_n\}$ 和 $\boldsymbol{Y}=\{\boldsymbol{y}_k\}$ 分别表示源说话人和目标说话人的语音特征参数空间，分别用 GMM 对这两个特征参数空间进行建模，得到的概率密度函数分别为

$$f(\boldsymbol{x})=\sum_{i=1}^{N_x}\alpha_i\cdot N\left(\boldsymbol{x};\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^x\right) \tag{4}$$

$$g(\boldsymbol{y})=\sum_{j=1}^{N_y}\beta_j\cdot N\left(\boldsymbol{y};\boldsymbol{\mu}_j^y,\boldsymbol{\Sigma}_j^y\right) \tag{5}$$

参量 $\{\alpha_i,\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^x\}$ 和 $\{\beta_i,\boldsymbol{\mu}_i^y,\boldsymbol{\Sigma}_i^y\}$ 可分别由期望最大(EM)算法迭代训练得到。从分类的角度来看，GMM 的每个高斯分量对应着一个子空间，假定 Ω_i^x 是源说话人特征参数空间中的第 i 个子空间，Ω_j^y 是目标说话人特征参数空间中的的第 j 个子空间，则 Ω_i^x 与 Ω_j^y 之间的 KL 距离为

$$\begin{aligned}D_{ij}&=\frac{1}{2}\left[D_{\mathrm{KL}}\left(\Omega_i^x\,||\,\Omega_j^y\right)+D_{\mathrm{KL}}\left(\Omega_j^y\,||\,\Omega_i^x\right)\right]\\&=\frac{1}{4}\Bigg[\lg\frac{\left|\boldsymbol{\Sigma}_j^y\right|}{\left|\boldsymbol{\Sigma}_i^x\right|}+\mathrm{Tr}\left[\left(\boldsymbol{\Sigma}_j^y\right)^{-1}\boldsymbol{\Sigma}_i^x\right]-d\\&\quad+\left(\boldsymbol{\mu}_i^x-\boldsymbol{\mu}_j^y\right)^{\mathrm{T}}\left(\boldsymbol{\Sigma}_j^y\right)^{-1}\left(\boldsymbol{\mu}_i^x-\boldsymbol{\mu}_j^y\right)\Bigg]\\&\quad+\frac{1}{4}\Bigg[\lg\frac{\left|\boldsymbol{\Sigma}_i^x\right|}{\left|\boldsymbol{\Sigma}_j^y\right|}+\mathrm{Tr}\left[\left(\boldsymbol{\Sigma}_i^x\right)^{-1}\boldsymbol{\Sigma}_j^y\right]-d\\&\quad+\left(\boldsymbol{\mu}_j^y-\boldsymbol{\mu}_i^x\right)^{\mathrm{T}}\left(\boldsymbol{\Sigma}_i^x\right)^{-1}\left(\boldsymbol{\mu}_j^y-\boldsymbol{\mu}_i^x\right)\Bigg]\end{aligned} \tag{6}$$

因此，可以计算源、目标语音特征参数空间中任意两个子空间的 KL 距离，则形成了如下的 KL 距离矩阵：

$$\boldsymbol{D}=\begin{bmatrix}D_{11}&D_{12}&\cdots&D_{1N_y}\\D_{21}&D_{22}&\cdots&D_{2N_y}\\\vdots&\vdots&\ddots&\vdots\\D_{N_x1}&D_{N_x2}&\cdots&D_{N_xN_y}\end{bmatrix} \tag{7}$$

3 ILNCA 算法

在非对称语音库情况下，假定 $\boldsymbol{X}=\{\boldsymbol{x}_n,n=1,\cdots,N\}$ 和 $\boldsymbol{Y}=\{\boldsymbol{y}_k,k=1,\cdots,K\}$ 分别表示源、目标说话人的语音声学特征参数空间，ILNCA 算法是在空间分类和 KL 距离的基础上，利用 KL 距离最小原理对源语音、目标语音的子空间进行匹配，然后在相对应的子空间内，根据最近邻准则将源、目标语音的特征参数进行对齐，最后利用循环迭代思想对转换函数进行训练。该算法的具体步骤如下：

(1)初始化：定义中间矢量集 $\boldsymbol{X}'=\{\boldsymbol{x}_n'\}$，并初始化为 $\boldsymbol{x}_n'=\boldsymbol{x}_n$。

(2)计算 KL 距离矩阵：分别用式(4)和式(5)对 $\boldsymbol{X}'=\{\boldsymbol{x}_n'\}$ 和 $\boldsymbol{Y}=\{\boldsymbol{y}_k\}$ 进行 GMM 建模，用期望最大(EM, Expectation Maximization)训练算法得到 GMM 参量 $\{\alpha_i,\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^x\}$ 和 $\{\beta_i,\boldsymbol{\mu}_i^y,\boldsymbol{\Sigma}_i^y\}$，然后根据式(6)计算出 KL 距离矩阵 $\boldsymbol{D}$。

(3)区域最近邻距离匹配：对任意的矢量 $\boldsymbol{x}_n'$ 根据最大后验概率准则进行分类，即

$$I\left(\boldsymbol{x}_n'\right)=\underset{k=1,2,\cdots,N_x}{\arg\max}\left(\frac{\alpha_k\cdot N\left(\boldsymbol{x}_n';\boldsymbol{\mu}_k^x,\boldsymbol{\Sigma}_k^x\right)}{\sum_{i=1}^{N_x}\alpha_i\cdot N\left(\boldsymbol{x}_n';\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^x\right)}\right) \tag{8}$$

在将 $\boldsymbol{x}_n^{'}$ 分类到第 $I(\boldsymbol{x}_n^{'})$ 子空间后，再根据 KL 距离最小准则将子空间 $I(\boldsymbol{x}_n^{'})$ 匹配到 $\boldsymbol{Y}$ 中的某子空间 J，即

$$J = \arg\min_j D_{I(\boldsymbol{x}_n^{'})j} \tag{9}$$

也就是说，$\boldsymbol{x}_n^{'}$ 应该在 $\boldsymbol{Y}$ 中的第 J 个子空间中根据最近邻准则找到其匹配的矢量，并将该矢量的下标存储为 $p(n)$，有

$$p(n) = \arg\min_k d\left(\boldsymbol{x}_n^{'}, \boldsymbol{y}_k\right), \quad \boldsymbol{y}_k \in \Omega_J^y \tag{10}$$

其中 $d(\boldsymbol{x}_n^{'}, \boldsymbol{y}_k)$ 表示矢量间的距离测度，距离测度的选择依赖于特征参数的类型，由于本文以线性预测倒谱系数(LPCC)作为特征参数，故距离测度采用欧式距离。

同理，对于任意的目标语音矢量 $\boldsymbol{y}_k$，也根据前述方法在 $\boldsymbol{X}^{'}$ 的某个子空间 I 中根据最近邻准则找到其所匹配的矢量 $\boldsymbol{q}(k)$。

这样，任意的 $\boldsymbol{x}_n^{'}$ 和 $\boldsymbol{y}_k$ 都有相匹配的矢量，不会造成遗漏，能充分地利用训练集中的数据，另外删掉那些重复匹配的矢量对。因此，就形成了一对相匹配的矢量序列。

(4)转换函数训练：将前面匹配好的矢量序列对 $\{\boldsymbol{x}_n^{'}, \boldsymbol{y}_{p(n)}\}$ 和 $\{\boldsymbol{x}_{\boldsymbol{q}(k)}^{'}, \boldsymbol{y}_k\}$ 中相对应的矢量拼接成一个更长的矢量 $\boldsymbol{z}$，即

$$\boldsymbol{z} = \left[\left(\boldsymbol{x}_n^{'}\right)^{\mathrm{T}}, \boldsymbol{y}_{p(n)}^{\mathrm{T}}\right]^{\mathrm{T}} \text{ 和 } \boldsymbol{z} = \left[\left(\boldsymbol{x}_{\boldsymbol{q}(k)}^{'}\right)^{\mathrm{T}}, \boldsymbol{y}_k^{\mathrm{T}}\right]^{\mathrm{T}} \tag{11}$$

然后用 GMM 对新的矢量空间 $\boldsymbol{Z} = \{\boldsymbol{z}\}$ 进行建模，有

$$p(\boldsymbol{z}) = \sum_{i=1}^{M} \omega_i \cdot N\left(\boldsymbol{z}; \boldsymbol{\mu}_i^z, \boldsymbol{\Sigma}_i^z\right) \tag{12}$$

参数集 $\{\omega_i, \boldsymbol{\mu}_i^z, \boldsymbol{\Sigma}_i^z\}$ 用 EM 算法训练得到。由于矢量 $\boldsymbol{z}$ 是由 $\{\boldsymbol{x}_n^{'}\}$ 和 $\{\boldsymbol{y}_k\}$ 中相对应的矢量拼接而成，故均值向量 $\boldsymbol{\mu}_i^z$ 和协方差矩阵 $\boldsymbol{\Sigma}_i^z$ 分别可以分解写成如下的结构形式：

$$\boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^{x'} \\ \boldsymbol{\mu}_i^{y} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i^z = \begin{bmatrix} \boldsymbol{\Sigma}_i^{x'x'} & \boldsymbol{\Sigma}_i^{x'y} \\ \boldsymbol{\Sigma}_i^{yx'} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix} \tag{13}$$

因此，根据式(13)可得转换函数为

$$F(\boldsymbol{x}) = \sum_{i=1}^{M} \lambda_i(\boldsymbol{x}) \cdot \left[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx'}\left(\boldsymbol{\Sigma}_i^{x'x'}\right)^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_i^{x'}\right)\right] \tag{14}$$

其中 $\lambda_i(\boldsymbol{x})$ 是后验概率，为

$$\lambda_i(\boldsymbol{x}) = \frac{\omega_i \cdot N\left(\boldsymbol{x}; \boldsymbol{\mu}_i^{x'}, \boldsymbol{\Sigma}_i^{x'x'}\right)}{\sum_{i=1}^{M} \omega_i \cdot N\left(\boldsymbol{x}; \boldsymbol{\mu}_i^{x'}, \boldsymbol{\Sigma}_i^{x'x'}\right)} \tag{15}$$

(5)更新中间矢量 $\boldsymbol{x}_n^{'}$：利用得到的转换函数式(14)对 $\boldsymbol{x}_n$ 进行转换，转换后的值作为中间矢量 $\boldsymbol{x}_n^{'}$ 的更新值，即为

$$\boldsymbol{x}_n^{'} = F\left(\boldsymbol{x}_n\right), \ \forall n \tag{16}$$

(6)收敛判决：判断收敛条件是否满足，如果满足，则停止迭代，得到转换函数式(14)；如果不满足，则回到第(2)步，进入下一个循环。收敛条件采用 $\dfrac{\overline{d}^{(m-1)} - \overline{d}^{(m)}}{\overline{d}^{(m-1)}} < \delta\%$，其中 $\overline{d}$ 是匹配后的序列中相对应矢量的平均距离，为

$$\overline{d} = \frac{1}{N+K}\left[\sum_{n=1}^{N} d\left(\boldsymbol{x}_n^{'}, \boldsymbol{y}_{p(n)}\right) + \sum_{k=1}^{K} d\left(\boldsymbol{x}_{q(k)}^{'}, \boldsymbol{y}_k\right)\right] \tag{17}$$

而 $\overline{d}^{(m-1)}$，$\overline{d}^{(m)}$ 分别表示第 $m-1$ 步和第 m 步迭代循环时的 $\overline{d}$。

4　实验与结果

4.1 语音信号模型及特征提取

本文采用 STRAIGHT 模型作为语音信号的分析/合成模型，它可以允许语音参数在进行较大幅度的修改之后还具有很高的语音合成质量，非常适合用于语音转换研究[12]。STRAIGHT 模型将语音信号分解成平滑的频谱和基音频率，因此，可以通过如下计算得到相应的线性预测编码(Linear Prediction Coding, LPC)模型的参数：

$$s(l) = \left|x(l)\right|^2, \quad 0 \le l \le L/2 \tag{18}$$

$$R(i) = \frac{1}{L}\sum_{l=0}^{L-1} s(l)\exp\left(j\frac{2\pi li}{L}\right), \quad 0 \le i \le L-1 \tag{19}$$

其中 $x(l)$，$0 \le l \le L-1$，是由 STRAIGHT 模型获得的频谱，$s(l)$ 是功率谱，且有 $s(l) = s(L-l)$，而 $R(i)$ 是自相关系数。根据功率谱和自相关系数互为傅里叶变换对可以得到自相关系数，因此，就可以根据 LPC 分析的自相关求解法获得 LPC 模型参数。之后，根据 LPC 参数获得 LPCC 参数，本文采用 LPCC 参数作为语音信号的特征参数。

4.2 客观评测

本实验所用的语音是在高信噪比的实验室环境下录制的汉语语音，信号的采样率为 16 kHz，每个样点 16 bit 量化。本文抽取其中 4 个人的语音，即 2 个男声和 2 个女声，分别命名为 M1, M2 和 F1, F2。每个人都取 300 个发音内容相同的语句，为了形成实验对照组，我们从每个人中选取 200 个语句作为对称的语音库 Ξ。同时，从每个人的 300 个语句中随机地选择不同的 200 个语句，从而形成非对称的语音库 $\widehat{\Xi}$。

实验的语音帧长为 20 ms，帧移为 10 ms，采用 Hamming 窗。另外，由于源、目标训练语音的样本数大致相同，因此，式(4)，式(5)和式(12)中 GMM 的分量个数取相同的数值，即 $N_x = N_y = M$。整个实验根据转换方向的不同分为 4 部分，分别是男声

转换成女声(M1-F1)、男声转换成男声(M1-M2)、女声转换男声(F2-M2)和女声转换成女声(F2-F1)。由于语音信号的频谱失真和听觉感觉特性密切相关，并且人耳的听觉特性又具有对数规律，因此本文采用文献[8]中的对数频谱相对距离来反映语音信号的频谱差异。

图 1 是在 4 个不同转换方向下 ILNCA, INCA[9] 和 GMM(具有对称语料基于 GMM 模型的算法)3 种转换算法的相对距离的对比图。从图上可以看出，在 GMM 模型的分量个数 M 较小的情况下，ILNCA 算法和 INCA 算法的转换性能差不多，但是随着 M 的逐渐增大，ILNCA 算法的性能要越来越明显地好于 INCA，甚至 M 增加到一定程度，ILNCA 的性能快要接近对称语音库转换算法 GMM 的性能。这是因为随着 M 增大，也即随着子空间数目的逐步增加，ILNCA 能够越来越精准地将源、目标语音的特征参数矢量进行对齐，这样就会使得转换函数更加准确，转换系统的性能也就越好。

图 2 表示的是 ILNCA 算法中收敛门限 δ 对算法的循环迭代次数和总体的相对距离的影响。从图上可以看到，随着 δ 的增加，即收敛的条件越松，算法的迭代次数会迅速减少，所要求的计算量也就越来越少；但相对距离会逐渐增大，即算法的转换性能是下降的。在 $\delta = 5$ 左右，如果 δ 再继续增大，迭代次数的减少会变得非常慢，而转换性能是急剧下降的。因此，在本文的实验中，我们取 $\delta = 5$ 。

图 3 表示的是一段语音信号在转换方向 F2-M2 下的经过 ILNCA, INCA 和 GMM 这 3 种算法转换后的语谱图对比，其中图 3(a)和图 3(b)分别表示源语音和目标语音的语谱图。从图上可以看出，ILNCA 算法比 INCA 算法能够更好地保留高频分量和一些能量较弱的频率分量，而这些频率分量有利于提高转换后的语音质量。

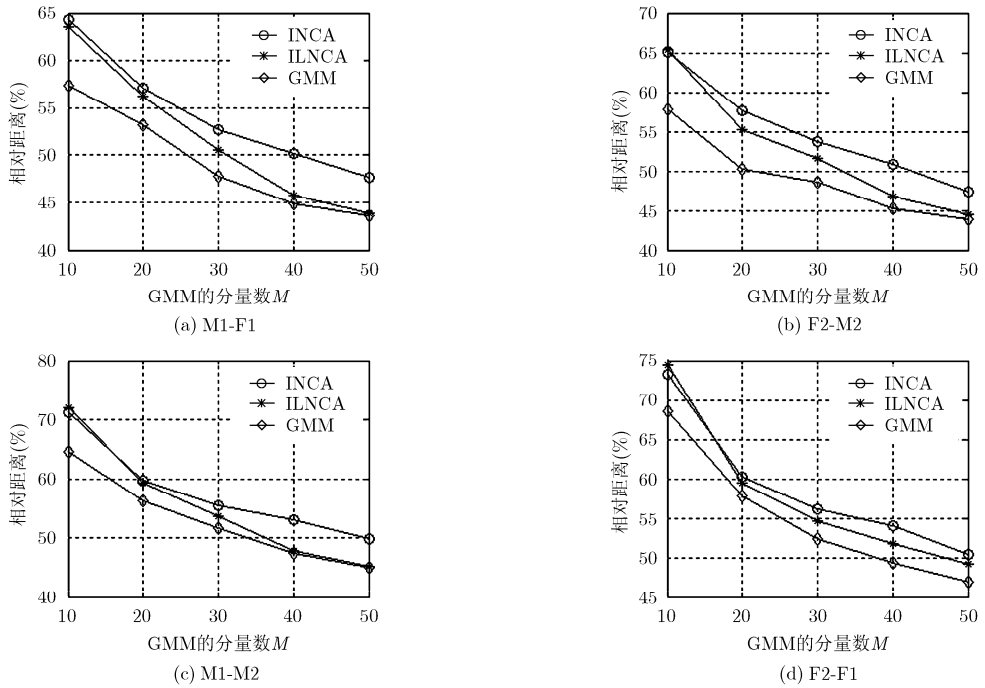

图 1 不同转换方向下 3 种转换算法的相对距离的性能对比

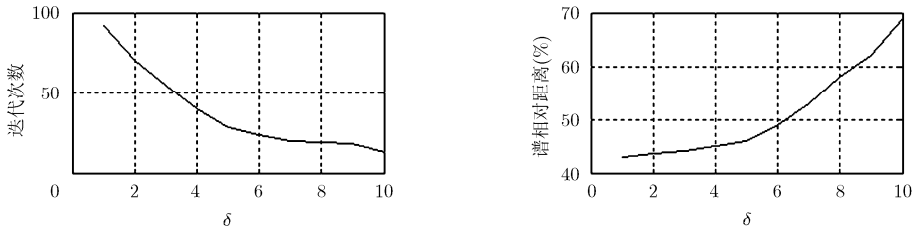

图 2 ILNCA 算法收敛门限 δ 对循环迭代次数和总的相对距离的影响

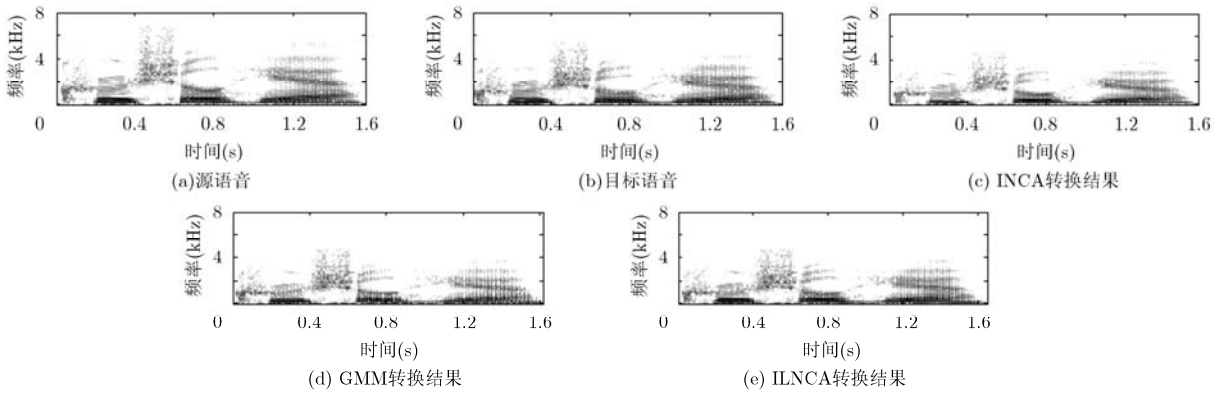

图 3 一段语音经过 3 种算法转换后的语谱图对照($M = 50$)

4.3 主观测试

人耳的听觉感受是评价语音转换的重要方法，毕竟转换后的语音最终还是要给人听的。主观听觉测试主要包括两方面，一是转换性能测试，主要是为了反映转换的程度，二是语音质量评价，语音质量的好坏对语音转换技术的应用具有非常大的影响。而转换性能测试又包括 ABX 测试和相似度测试。ABX 测试中的 A 和 B 分别表示源说话人和目标说话人，X 指的转换后的语音，该测试的目的主要是为了反映转换后的语音听起来是像源说话人还是更像目标说话人，如果听起来像源说话人则得分为 0，如果像目标说话人则得分为 1，然后将总分加起来再去除以总共测试的语音个数。相似度测试是为了反映转换后的语音跟目标语音的相似程度，用 5 分制来打分，其中 1 表示“完全不同”，5 表示“完全相同”，其余几个分值介于它们之间。而对语音质量的评价采用的是常用的 MOS 打分。

在本文实验中，参与主观听觉测试的人数为 20 人，每人在每个转换方向上评测 50 个语句，这样在每个转换方向上就要进行 1000 次评测。表 1 是 ABX 的测试结果。从表中可以看出，异性之间转换的 ABX 分值要高于同性之间。这是因为，异性之间的频谱距离虽然比同性之间的要大，但它的转换程度要大于同性，这样就导致转换后的语音听起来很明显地不再像源说话人，而是像目标说话人。这是一种相对的结果，这一结果也和客观测试中的频谱相对距离的结果相吻合。图 4 是相似度的测试结果。从图中的结果来看，同性之间转换的打分要高，也即同性之间转换时转换语音和目标语音更像。这一结果和 ABX 结果其实并不矛盾，相似度反映的转换语音和目标语音之间的绝对距离，对于同性之间转换来讲，源语音和目标语音之间频谱距离要小于异性转换，转换后语音和目标语音之间频谱距离也自然要小，所以，同性之间转换的相似度测试分要好于异性之间。从表 1 和图 4 综合来看，本文算法 ILNCA 比 INCA 在转换性能上具有明显的改善，只是略低于对称语音库情况下的转换算法。图 5 是转换后语音的 MOS 分。从图中可以看出，ILNCA 算法在语音质量上比 INCA 改善得并不大，只是略有提高。另外也可以看出，同性转换的语音质量要好于异性之间。这是因为，异性语音频谱之间的距离一般要大于同性之间，转换的程度也要大些，而对语音参数修改的程度要大，对语音质量的影响也越大，这就导致了异性转换之间的语音质量有所下降。

表 1 ABX 测试值(%)

算法	M1-F1	F2-M2	M1-M2	F2-F1
GMM	95.6	94.9	91.9	92.1
INCA	91.4	91.2	87.6	88.2
ILNCA	94.2	94.1	88.3	88.7

5 结束语

本文针对 INCA 转换算法的不足，提出了一种

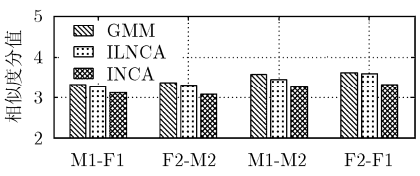

图 4 3 种算法的相似度得分对照

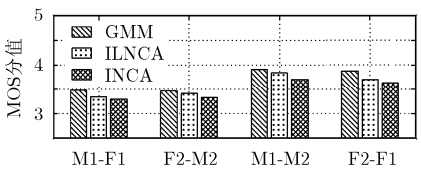

图 5 3 种算法的 MOS 打分对照

改进的用于非对称语料的转换函数迭代训练算法 ILNCA。ILNCA 算法首先利用 GMM 模型分别对源、目标语音信号的特征参数空间进行建模，然后利用 KL 距离最小原则将源、目标语音特征参数空间的各个子空间进行匹配，最后在相匹配的两个子空间中利用最近邻准则将源、目标语音的特征参数矢量进行对齐，从而对转换函数进行迭代训练。客观测试和主观听觉评测结果表明，本文算法在转换性能上比 INCA 有较大的改善，同时在语音质量上也略有提高。

参 考 文 献

[1] 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-jü, and Ruan Xiao-gang. Voice conversion technology and its development[J]. *Acta Electronica Sinica*, 2004, 32(7): 1165–1172.

[2] Saito D, Watanabe S, Nakamura A N, *et al*.. Statistical voice conversion based on noisy channel model[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, 20(6): 1784–1794.

[3] Abe M, Nakamura S, Shikano K, *et al*.. Voice conversion through vector quantization[C]. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), New York, USA, April 11–14, 1988: 655–658.

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

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

[6] Wu C H, Hsia C C, Liu T H, *et al*.. 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.

[7] Mouchtaris A, Van der Spiegel J, and Mueller P. Nonparallel training for voice conversion based on a parameter adaptation approach[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, 14(3): 952–963.

[8] Ye Hui and Young S. Quality-enhanced voice morphing using maximum likelihood transformations[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2006, 14(4): 1301–1312.

[9] Erro D, Moreno A, and Bonafonte A. INCA algorithm for training voice conversion systems from nonparallel corpora [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, 18(5): 944–953.

[10] Cover T M and Thomas J A. 信息论基础[M]. 北京: 清华大学出版社, 2003: 231–232.

[11] Hershey J R and Olsen P A. Approximating the Kullback Leibler divergence between Gaussian mixture models [C]. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, USA, April 15–20, 2007, Vol. IV: 317–320.

[12] Kawahara H, Masuda-Katsuse I, and Cheveigne A D. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds[J]. *Speech Communication*, 1999, 27(3): 187–207.

简志华： 男，1978 年生，讲师，研究方向为语音转换、语音识别以及无线网络中的语音通信技术.

王向文： 男，1978 年生，讲师，研究方向为视频信号处理、多媒体信号处理.