第 39 卷第 3 期　　声　学　学　报　　Vol. 39, No. 3
2014 年 5 月　　ACTA　ACUSTICA　　May, 2014

采用压缩感知的改进的语音转换算法*

简志华[1]　王向文[2]

(1　杭州电子科技大学通信工程学院　杭州　310018)

(2　上海电力学院电子与信息工程学院　上海　200090)

2012 年 12 月 19 日收到

2013 年 9 月 13 日定稿

摘要　提出了一种基于压缩感知的考虑语音帧间信息的语音转换算法。根据连续多帧语音的线谱对参数所构成的矢量在离散余弦变换域具有稀疏性, 利用压缩感知技术对该矢量压缩成短矢量, 并将该压缩后的短矢量作为特征参数训练语音转换函数。实验测试结果表明, 选择合适的语音帧数时, 该算法的性能要比传统的采用加权频率卷绕的转换算法提高 3.21% 。这说明, 充分有效地利用语音帧间的相关信息会使转换语音保持更稳定的帧间声学特性, 有利于提高语音转换系统的性能。

PACS 数: 43.72

A modified algorithm for voice conversion using compressed sensing

JIAN Zhihua[1]　WANG Xiangwen[2]

(1　*School of Communication Engineering, Hangzhou DianZi University*　Hangzhou　310018)

(2　*Institute of Electronic and Information Engineering, Shanghai University of Electric Power*　Shanghai　200090)

Received Dec. 19, 2012

Revised Sept. 13, 2013

Abstract　A voice conversion algorithm, which makes use of the information between continuous frames of speech by compressed sensing, is proposed in this paper. According to the sparsity property of the concatenated vector of several continuous Linear Spectrum Pairs (LSP) in the discrete cosine transformation domain, this paper utilizes compressed sensing to extract the compressed vector from the concatenated LSPs and uses it as the feature vector to train the conversion function. The results of evaluations demonstrate that the performance of this approach can averagely improve 3.21% comparing with the conventional algorithm based on weighted frequency warping when choosing the appropriate numbers of speech frame. The experimental results also illustrate that the performance of voice conversion system can be improved by taking full advantage of the inter-frame information, because those information can make the converted speech remain the more stable acoustic properties which is inherent in inter-frames.

引言

在语音信号中, 说话人的个性特征是一种非常重要的信息。语音转换就是要改变源说话人语音中的个性特征信息, 使之具有目标说话人的个性特征, 也即转换后的语音听起来就像是目标说话人的声音一样, 但语音的语义内容保持不变[1]。语音转换在多个领域都具有潜在的应用价值, 比如: 个性化的语音合成和文语转换系统[2]、语音情感的转换[3−4]、用于语音翻译系统[5]、在多媒体娱乐中的应用[6]、用于听觉障碍人的语音辅助系统[7]、说话人伪装身份通信和用于语音识别系统的前端预处理模块[8]。

语音转换是通过提取并改变语音信号中与说话人身份相关的声学特征参数, 之后再利用改变后的特征参数来合成具有目标说话人身份特征的语音。影响说话人特性的因素有很多, 比如语言表达方式、口音、心理状态等。但语音转换研究主要考虑的是声

* 国家自然科学基金 (61201301) 和浙江省教育厅项目 (Y201016542) 资助

学层面的因素，比如声道特性和基音频率。语音转换研究的核心问题是寻找能够精确反映源说话人特征参数和目标说话人特征参数之间映射关系，即转换函数。早期的语音转换函数是基于矢量量化 (Vector Quantization, VQ) 模型 [9]。但这种基于 VQ 的转换算法由于将特征参数矢量离散化，特征参数被限制在一个有限的矢量码本中，导致频谱的不连续性，转换性能和语音质量都不理想。1998 年，Y.Stylianou 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 [10]。不久，A. Kain 对基于 GMM 的算法进行了改进，提出了联合矢量 GMM 模型 [11]，简化了运算，也使得基于 GMM 的转换算法逐渐地成为语音转换的主流算法 [12−15]。但由于基于 GMM 的转换函数是基于统计平均，使频谱过于平滑，导致转换后的语音质量和自然度下降。为了提高语音质量，A.Pribilova 等人提出了一种基于频率卷绕的转换算法，具有较好的语音质量，但转换效果不佳 [16]。D.Erro 综合了基于 GMM 的转换算法和频率卷绕算法的优势，提出了一种在 GMM 模型的基础上进行加权的频率卷绕算法 (WFW, Weighted Frequency Warping)，较好地平衡了语音质量和转换性能之间的矛盾 [17]。

包括 WFW 在内的以上算法都有一个共同的特点，那就是在训练时都是单独地对每帧语音进行处理，没有考虑语音帧间的相关信息。事实上，语音帧间具有很强的相关性，这些相关信息不仅有利于改善转换效果，也有利于提高转换后的语音质量。本文正是基于上述考虑，提出了采用压缩感知理论来考虑语音特征参数的帧间相关信息。

1 传统的 WFW 转换算法

在平行语料库的情况下，对源语音和目标语音相对应的语句利用动态时间规整 (Dynamic Time Warping, DTW) 算法进行时间上的对齐，假定源说话人的语音特征参数矢量序列为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n,\cdots,\boldsymbol{x}_N\}$，目标说话人语音特征参数序列为 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n,\cdots,\boldsymbol{y}_N\}$。将 $\boldsymbol{x}_n$ 与对应的 $\boldsymbol{y}_n$ 拼接成一个新的联合矢量 $\boldsymbol{z}_n$，即 $\boldsymbol{z}_n=[\boldsymbol{x}_n^{\mathrm{T}},\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}$，其中符号 “T” 表示矩阵转置。因此，就得到了联合矢量空间 $\boldsymbol{Z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_n,\cdots,\boldsymbol{z}_N\}$，对该空间用 GMM 进行建模，并用期望最大 (Expectation Maximization, EM) 算法训练 GMM 模型参数，即：

$$p(\boldsymbol{z})=\sum_{i=1}^{M}\alpha_i N(\boldsymbol{z};\boldsymbol{\mu}_i^z,\boldsymbol{\Sigma}_i^z), \tag{1}$$

其中，α_i 是权重，$\boldsymbol{\mu}_i^z$ 和 $\boldsymbol{\Sigma}_i^z$ 分别表示第 i 个分量的均值向量和协方差矩阵，M 表示高斯分量的总个数，并且有：

$$\boldsymbol{\mu}_i^z=\begin{bmatrix}\boldsymbol{\mu}_i^x\\ \boldsymbol{\mu}_i^y\end{bmatrix},\quad \boldsymbol{\Sigma}_i^z=\begin{bmatrix}\boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy}\\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy}\end{bmatrix}. \tag{2}$$

这样，由 $\boldsymbol{Z}$ 的 GMM 模型就可以得到 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的 GMM 模型。$\boldsymbol{\mu}_i^x$ 和 $\boldsymbol{\mu}_i^y$ 分别对应着 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的第 i 个子空间的均值矢量，由于系统在进行转换时所采用的特征参数是线谱对 (Linear Spectrum Pair, LSP) 参数，因此 $\boldsymbol{\mu}_i^x$ 和 $\boldsymbol{\mu}_i^y$ 就具有 LSP 参数的特性。在语音信号的全极点模型中，LSP 和语音谱包络的谱峰分布有对应关系，因此根据相对应的谱峰频率，利用分段线性频率卷绕函数就可以得到第 i 个子空间的频率卷绕曲线函数 $W_i(f)$[17]。因此，对于语音帧 $\boldsymbol{x}_n$ 来讲，完整的频率卷绕曲线函数为：

$$W^{(n)}(f)=\sum_{i=1}^{M}\beta_i(\boldsymbol{x}_n)\cdot W_i(f), \tag{3}$$

其中 $\beta_i(\boldsymbol{x}_n)$ 是后验概率，为：

$$\beta_i(\boldsymbol{x}_n)=\frac{\alpha_i N(\boldsymbol{x}_n;\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{M}\alpha_j N(\boldsymbol{x}_n;\boldsymbol{\mu}_j^x,\boldsymbol{\Sigma}_j^{xx})}. \tag{4}$$

2 改进的 WFW 转换算法及系统框架

2.1 MWFW 算法

从上节可以看出，WFW 算法是单独对每帧语音的特征参数进行训练和转换，没有考虑到语音帧间的相关性。而事实上，语音帧间的相关信息具有重要的作用，为了利用语音帧间的相关性，同时也考虑到在转换函数训练时，语音段比语音帧更具有稳定性，有利于提高语音质量，本文采用压缩感知 (Compressed Sensing, CS) 技术来提取语音帧间的相关信息。

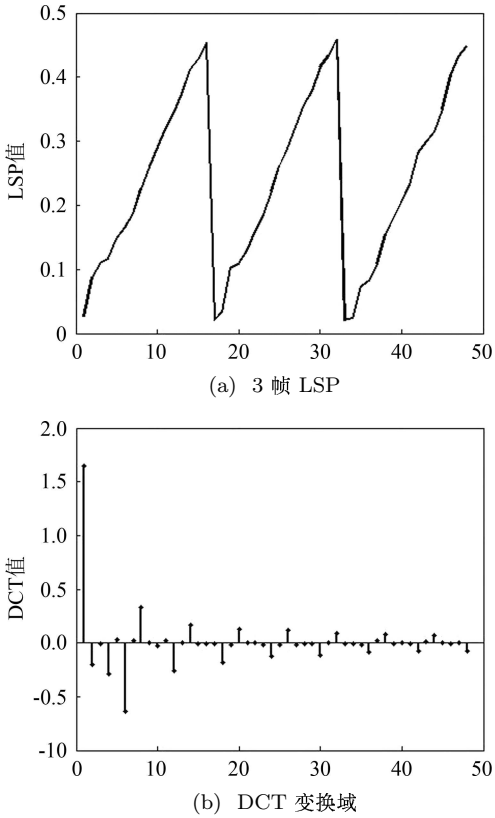
CS 理论指出 [18]，只要信号是可压缩的或在某个变换域是稀疏的，那么就可以用一个与变换基不相关的观测矩阵将高维信号投影到一个低维空间上，然后通过求解一个优化问题就可以从这些少量投影中以高概率重构出原信号。

由于 LSP 参数具有良好的量化和插值特性，使它成为目前语音转换中使用最为广泛的特征参数。假定 $\boldsymbol{x}_n$ 是当前时刻语音帧的 LSP 参数，即 L 维的列矢量，则 $\boldsymbol{x}_{n-(\tau-1)/2},\cdots,\boldsymbol{x}_{n-1},\boldsymbol{x}_n,\boldsymbol{x}_{n+1},\cdots,\boldsymbol{x}_{n+(\tau-1)/2}$ 是一以 $\boldsymbol{x}_n$ 为中心的由 τ (τ 为奇数) 帧语音 LSP 参数构成的矢量序列。将这一矢量序列按时间先后顺序拼接起来形成一个长的矢量，即有：

$$\boldsymbol{X}_n = \Big[\boldsymbol{x}_{n-(\tau-1)/2}^{\mathrm{T}}, \cdots, \boldsymbol{x}_{n-1}^{\mathrm{T}}, \boldsymbol{x}_n^{\mathrm{T}}, \boldsymbol{x}_{n+1}^{\mathrm{T}}, \cdots, \boldsymbol{x}_{n+(\tau-1)/2}^{\mathrm{T}}\Big]^{\mathrm{T}}, \tag{5}$$

则 $\boldsymbol{X}_n$ 是一个 $\tau L \times 1$ 维的列矢量。将 $\boldsymbol{X}_n$ 进行离散余弦变换 (DCT), 其结果如图 1 所示 ($\tau = 3$) 。图 1(a) 是相邻 3 帧 LSP 参数的轨迹图，每帧 LSP 为 16 维的矢量; 图 1(b) 是三帧联合矢量 $\boldsymbol{X}_n$ 进行 DCT 变换后的值，为 48 维的矢量。从图 1 中可以看出，多帧 LSP 参数具有准周期的性质; 联合矢量 $\boldsymbol{X}_n$ 在 DCT 域具有很好的稀疏性，其大部分的系数都为零或者接近于零，即 $\boldsymbol{\Psi}\boldsymbol{X}_n$ 是稀疏的，其中 $\boldsymbol{\Psi}$ 是 DCT 变换矩阵。这说明在 DCT 域，联合矢量 $\boldsymbol{X}_n$ 采用压缩感知理论是完全可行的。令观测矩阵 $\boldsymbol{\Phi}$ 为一个 $I \times \tau L$ 维的高斯随机矩阵，并且可以保证 $\boldsymbol{\Phi}$ 和 $\boldsymbol{\Psi}$ 是不相关的 [19], 于是观测矢量为：

$$\boldsymbol{g}_n^x = \boldsymbol{\Phi}\Psi\boldsymbol{X}_n. \tag{6}$$



(a) 3 帧 LSP

(b) DCT 变换域

图 1 三帧 LSP 参数的轨迹及联合进行 DCT 变换后的值

根据 CS 理论，I 是一个比 τL 小得多的值。因此，第 n 帧语音的 LSP 特征参数 $\boldsymbol{x}_n$ 就变换成了 I 维的矢量 $\boldsymbol{g}_n^x$ 。这样，采用 $\boldsymbol{g}_n^x$ 做特征参数，不仅包含了当前语音帧的信息，也包含了前后几帧语音的信息。在本文中，$I = \lceil \xi \log(\tau L/\xi) \rceil$, 其中符号 $\lceil\cdot\rceil$ 表示不小于某数的最小整数，ξ 是稀疏度。而稀疏度是根据 LSP 联合矢量在 DCT 域具有较大系数的能量和占总能量的比值来确定，本文采用比值为 90% 。由于 I 是压缩以后的矢量维数，因此考虑帧间相关信息并不会增加 GMM 模型训练算法的计算量。

同理，提取目标说话人相应的 $\boldsymbol{g}_n^y$, 再将 $\boldsymbol{g}_n^x$ 和相对应的 $\boldsymbol{g}_n^y$ 拼接起来就构成了矢量 $\boldsymbol{g}_n$，即 $\boldsymbol{g}_n = [\boldsymbol{g}_n^{xT}, \boldsymbol{g}_n^{yT}]^{\mathrm{T}}$ 。用式 (1) 的 GMM 模型对矢量空间 $\{\boldsymbol{g}_n\}$ 进行建模，得到参数集 $\{\alpha_i, \boldsymbol{\mu}_i^g, \boldsymbol{\Sigma}_i^g\}$, 其中 $\boldsymbol{\mu}_i^g$ 同样可以展成：

$$\boldsymbol{\mu}_i^g = \begin{bmatrix} \boldsymbol{\mu}_i^{gx} \\ \boldsymbol{\mu}_i^{gy} \end{bmatrix}. \tag{7}$$

在物理意义上，$\boldsymbol{\mu}_i^{gx}$ 和 $\boldsymbol{\mu}_i^{gy}$ 相当于 τ 帧语音的 LSP 参数经过 CS 压缩后所得到的值。为了能得到 GMM 各子空间的频率卷绕函数，将 $\boldsymbol{\mu}_i^{gx}$ 和 $\boldsymbol{\mu}_i^{gy}$ 分别用正交匹配追踪法 [19] 重构出各自连续的 τ 帧 LSP 参数, 并取其对应的位于中间的 LSP 参数 $\overline{\boldsymbol{\mu}}_i^x$ 和 $\overline{\boldsymbol{\mu}}_i^y$ 。此时，和 WFW 算法一样，利用 $\overline{\boldsymbol{\mu}}_i^x$ 和 $\overline{\boldsymbol{\mu}}_i^y$ 获得第 i 个子空间的频率曲线函数 $W_i(f)$, 再用后验概率进行加权就可以得到整体的频率卷绕函数 $W^{(n)}(f)$, 但后验概率为：

$$\beta_i(\boldsymbol{g}_n^x) = \frac{\alpha_i N(\boldsymbol{g}_n^x; \boldsymbol{\mu}_i^{gx}, \boldsymbol{\Sigma}_i^{gxx})}{\sum_{j=1}^{M} \alpha_j N(\boldsymbol{g}_n^x; \boldsymbol{\mu}_j^{gx}, \boldsymbol{\Sigma}_j^{gxx})}. \tag{8}$$

这样，在训练阶段就获得了用于转换的改进的频率卷绕函数式 (3) 。在转换阶段，分别提取每帧源语音信号的 $\boldsymbol{g}_n^x$ 参数和频谱函数 $X(f)$, 然后用训练阶段得到的 $W^{(n)}(f)$ 对 $X(f)$ 进行转换，并用 $\boldsymbol{g}_n^x$ 计算后验概率。

2.2 基音频率转换函数

由于人耳的听觉特性和频率近似满足对数关系，因此对浊音帧中基音频率的转换就在对数域中进行，转换方法是基于统计的方法，为：

$$\log f_0' = \mu_{\log f_0}^y + \frac{\sigma_{\log f_0}^y}{\sigma_{\log f_0}^x}\left(\log f_0 - \mu_{\log f_0}^x\right), \tag{9}$$

其中 f_0 和 f_0' 分别是源语音和转换后的基音频率，而参数 $\mu_{\log f_0}^x$ 、$\sigma_{\log f_0}^x$ 和 $\mu_{\log f_0}^y$ 、$\sigma_{\log f_0}^y$ 分别是源语音和目标语音的基音频率在对数域中的均值和均方差。而对于清音帧来讲，就保持不变，不进行处理。

2.3 转换系统框图

整个转换系统原理框图可见图 2 。从图上可以看出，语音转换可以分为两个阶段，即训练阶段和转换阶段。在训练阶段，语音信号首先利用 STRAIGHT

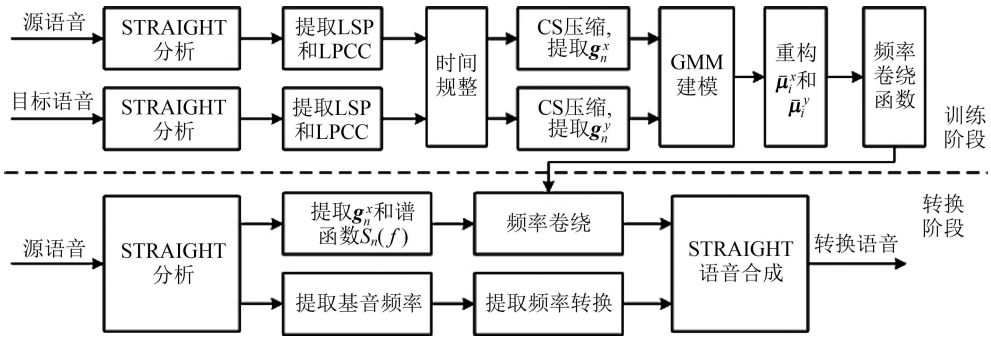


图 2　语音转换系统框图

模型进行建模. 该模型可以作为语音信号的分析 / 合成模型, 它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, 语音信号经过 STRAIGHT 模型之后会被分解成平滑的短时频谱和基音频率 [20]。然后再通过短时频谱分别提取出语音特征参数 LSP 和 LPCC, 其中 LSP 参数用于训练转换函数，而 LPCC 则用于时间规整。

在 STRAIGHT 模型下, 这两种参数的提取是通过计算相应的 LPC 系数而得到 [21], 具体如下：

$$S(k) = |X(k)|^2, \quad 0 \leqslant k \leqslant \frac{K}{2}, \tag{10}$$

$$R(i) = \frac{1}{K}\sum_{k=0}^{K-1} S(k)\exp\left(\mathrm{j}\frac{2\pi ki}{K}\right), \tag{11}$$

其中，$X(k)$, $0 \leqslant k \leqslant K-1$, 是由 STRAIGHT 模型获得的短时频谱，$S(k)$ 是功率谱，并且有 $S(k) = S(K-k)$, 而 $R(i)$ 是计算的自相关系数。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。在有了自相关系数之后，就可以根据 LPC 分析的自相关求解法获得 LPC 系数，然后再分别转换成 LSP 和 LPCC 参数。本文在进行 LPC 分析时，取阶次为 16 。之后，转换系统就利用 LPCC 参数进行时间规整，形成两个一一对应的时间序列，再利用本文前述的算法训练得到频率卷绕函数，即转换函数。

在转换阶段, 源语音信号经过 STRAIGHT 模型后，得到 $\boldsymbol{g}_n^x$ 和短时频谱 $S_n(f)$, 其中 $\boldsymbol{g}_n^x$ 用来计算式 (8) 中的后验概率, 以得到频率卷绕函数中的加权系数。然后再对短时频谱 $S_n(f)$ 进行转换，转换后的短时频谱和转换后的基音频率通过 STRAIGHT 语音模型就合成出转换后的语音信号。

3　实验与结果

3.1　语音库

本实验所采用的语音库 CASIA 是购自中国科学院自动化所模式识别实验室，是在高信噪比 (> 30 dB) 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，发音是采用中性的朗读风格。我们抽取其中 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 200 个语句，每个语句大致是 2~3s 时长的短语和短句，其中 150 个用于训练， 50 个用于测试。而且每个人的发音内容相同, 也即是对称的语音库。

3.2　客观评测

实验的语音帧长为 20 ms, 帧移为 10 ms, 采用 Hamming 窗。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) 。由于语音信号的听觉感觉特性和对数域的频谱密切相关，本文采用如下公式来反映两帧语音信号的频谱差异 [22]。

$$d(S_1, S_2) = \frac{1}{K}\sum_{k=1}^{K}\left(10\log_{10}a_k^1 - 10\log_{10}a_k^2\right)^2, \tag{12}$$

其中 S_1 和 S_2 表示两语音信号，a_k^1 和 a_k^2 表示两语音信号对应帧的 STRAIGHT 频谱的第 k 个频点的幅度，K 是总的频点数。$d(S_1, S_2)$ 是绝对距离，而根据语音转换的特点，我们采用如下的相对距离来作为反映语音转换性能的客观准则。

$$D = \frac{\sum_{n=1}^{N} d\left(S_{\mathrm{tgt}}(n), S_{\mathrm{conv}}(n)\right)}{\sum_{n=1}^{N} d\left(S_{\mathrm{tgt}}(n), S_{\mathrm{src}}(n)\right)} \times 100\%, \tag{13}$$

其中 S_{src}, S_{tgt} 和 S_{conv} 分别表示源、目标和转换语音，N 表示测试中总共的语音帧数。从上式可以看出，D 越小，转换性能就越好。本文试验中，GMM 的分量个数为 8, 与文献 17 一致。

图 3 表示的是在 τ 分别等于 3, 5, 7, 9 几种情况下的 MWFW 算法和 WFW 算法的性能对比图。从图上可以看出，有些情况下的 MWFW 频谱相对距离比 WFW 小，有些情况下要大，但从整体上来讲 MWFW 的性能要好，特别是 MWFW5 在 4 个转换方向上都要好于 WFW 。这是由于当 $\tau = 5$ 时， 5 帧语音所构成的语音段能够较好地反映出语音的帧间相关性和稳定性，当 τ 越大时，相关性则越来越弱，就不利于语音转换性能的提高。

图 4 是连续 5 帧语音的 LSP 参数的轨迹图。其中 src 表示源语音， conv 表示用本文算法 MWFW 对 src 进行转换后得到的 LSP, 而 tgt 表示相对应的目标语音。从图上可以看出，转换后的 LSP 轨迹与目标语音的 LSP 轨迹更吻合，这说明 MWFW 算法非常有效。

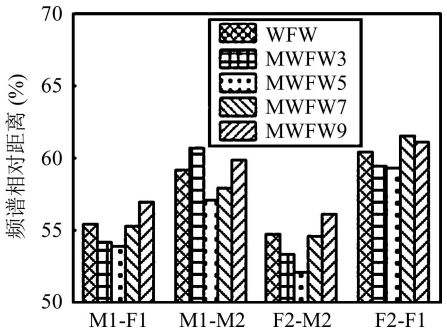


图 3 几种转换情况下的频谱相对距离的对比

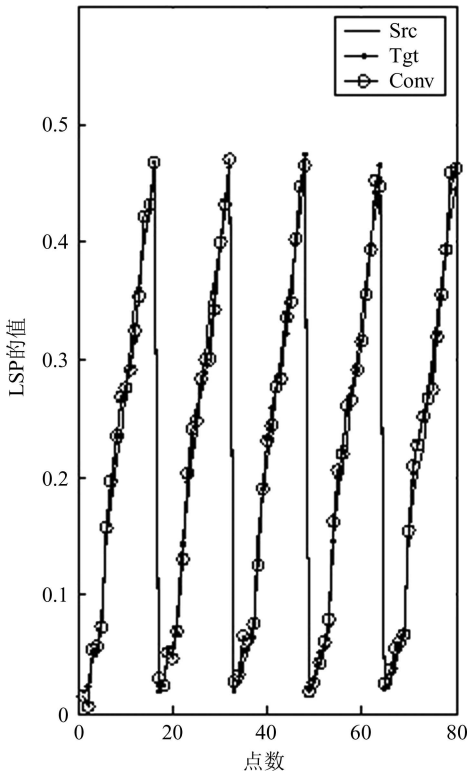


图 4 连续 5 帧 LSP 参数的对照图

图 5 是一段语音分别经过 WFW 算法和 MWFW 算法转换后得到的语谱图。从图上可以看出，两种算法都有较好的转换效果，但不管是从语音信号的高频分量还是低频分量来讲， MWFW 都比 WFW 保留了更多的目标语音频率分量。

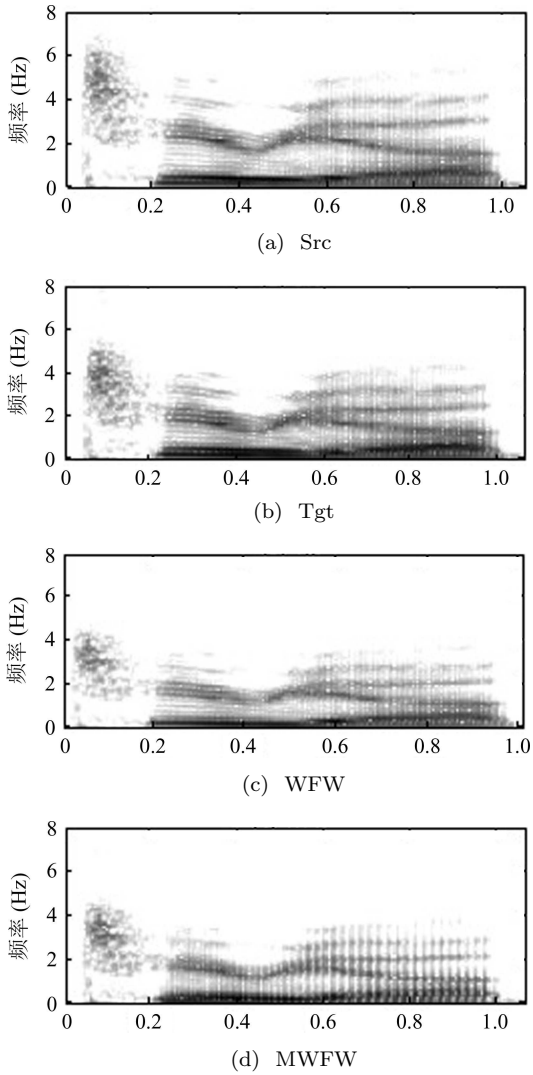


图 5 同一语音段在几种情况下的语谱图对照

3.3 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： 一是相似度测试，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。而相似度的测试又包括 ABX 测试和倾向度测试。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。倾向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实验人员总共为 20 人，既有专业研究人员，也有普通学生，在每个转换方向上评测 50 个语句，也就是说每个人要对 200 个语句进行评价。表 1 是 ABX 的测试结果。从表 1 可以看出，在异性之间的转换，ABX 的结果要好于同性之间。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像目标说话人，而不像源说话人。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 D 的结果相吻合。表 2 是倾向度的测试结果。从表 2 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，倾向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。因此，同性之间转换的倾向度测试分要好于异性之间。表 3 是转换后语音的 MOS 分。从表 3 来看，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 测试结果

	M1-F1	M1-M2	F2-M2	F2-F1
WFW	92.1%	88.7%	93.4%	89.1%
MWFW5	94.3%	90.1%	95%	89.3%

表 2 倾向度测试结果

	M1-F1	M1-M2	F2-M2	F2-F1
WFW	3.13	3.37	3.18	3.34
MWFW5	3.17	3.41	3.19	3.38

表 3 MOS 分测试结果

	M1-F1	M1-M2	F2-M2	F2-F1
WFW	2.89	3.14	2.83	3.12
MWFW5	2.94	3.22	2.91	3.20

4 总结

语音信号在时序上具有很强的相关性, 语音帧间相关信息已在许多语音处理系统中得到广泛应用。本文首先梳理了目前许多语音转换算法都没有考虑到语音帧间的相关信息，接着分析了语音信号的相邻多帧线谱对特征参数在离散余弦变换域具有很好的稀疏性, 就提出了采用压缩感知技术对多帧语音的线谱对参数压缩成一个短矢量, 并将该矢量作为特征参数用来进行训练语音转换函数的算法。另外，由于对数域的频率能更好地拟合人耳的听觉特性，对基音频率的转换就在对数域中利用统计均值和均方差所构成的函数进行转换。实验结果表明，不管是语音相似度还是语音质量，本文提出的转换算法都要好于传统的算法。这是因为在将多帧语音特征参数压缩成一个短矢量之后再同时进行转换会更加具有稳定性，特别是对于语音发音非平稳段，这是语音帧间的固有特性，因此有利于转换的准确性，提高转换系统性能。

与说话人身份特性相关的特征有许多种，目前的转换算法主要采用的是反映频谱特性的一些声学特征参数，而许多高层次的特征参数，比如语速、口音、情感以及一些韵律特征等，由于当前对这些特征参数还没有很好的提取和建模方法, 都未在语音转换中得到充分利用。另外，由于汉语是一种有调语言，声调对汉语语音的自然度和清晰度都有很重要的影响，汉语语音转换应该要考虑到声调的作用，以便提高转换语音的质量和准确性。这些都是今后语音转换研究工作的重点。

参 考 文 献

1 左国玉，刘文举，阮晓钢. 语音转换技术的研究与进展. 电子学报， 2004; **32**(7): 1165—1172

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

3 Zeynep Inanoglu, Steve Young. Data-driven emotion conversion in spoken English. *Speech Communication*, 2009; **51**(3): 268—283

4 Daniel Erro, Eva Navas, Inma Hernaez, Ibon Saratxaga. Emotion conversion based on prosodic unit selection. *IEEE Transactions on Audio, Speech and Language Processing*, 2010; **18**(5): 974—983

5 Khadivi S, Ney H. Integration of speech recognition and machine translation in computer-assisted translation. *IEEE Transaction on Audio, speech and Language Processing*, 2008; **16**(8): 1551—1564

6 Yuji Sato. Voice quality conversion using interactive evolution of prosodic control. *Applied Soft Computing*, 2005; **5**(2): 181—192

7 Doi H, Nakamura K, Toda T, Saruwatari H, Shikano K. An evaluation of alaryngeal speech enhancement methods based on voice conversion. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011: 5136—5139

8 Choi H C, King R W. On the use of spectral transformation for speaker adaptation in HMM based isolated-word speech recognition. *Speech Communication*, 1995; **17**(3): 131—143

9 Abe M, Nakamura S, Shikano K, Kuwabara H. Voice conversion through vector quantization. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), New York, USA, 1988: 655—658

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

11 Kain A, Macon M W. Design and evaluation of a voice conversion algorithm based on spectral envelop mapping and residual prediction. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Salt Lake City, USA, 2001: 813—816

12 Oytun Turk, Marc Schroder. Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. *IEEE Transactions on Audio, Speech and Language Processing*, 2010; **18**(5): 965—973

13 Kunikoshi A, Qian Y, Soong F, Minematsu N. Improved F0 modeling and generation in voice conversion. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011: 4568—4571

14 俞一彪, 曾道建, 姜莹. 采用独立说话人模型的语音转换. 声学学报, 2012; **37**(3): 346—352

15 Chen Linghui, Ling Zhenhua, Dai Lirong. Non-parallel training for voice conversion based on FT-GMM. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011: 5116—5119

16 Pribilova A, Pribil J. Non-linear frequency scale mapping voice conversion in text-to-speech system with cepstral description. *Speech Communication*, 2006; **48**(12): 1691—1703

17 Daniel Erro, Asuncion Moreno, Antonio Bonafonte. Voice conversion based on weighted frequency warping. *IEEE Transactions on Audio, Speech and Language Processing*, 2010; **18**(5): 922—931

18 Duarte M F, Eldar Y C. Structured Compressed sensing: from theory to applications. *IEEE Transactions on Signal Processing*, 2011; **9**(9): 4053—4085

19 Tropp J A, Gilbert A C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 2007; **53**(12): 4655—4666

20 Kawahara H, Masuda-Katsuse I, de Cheveigne A. 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. *Speech Communication*, 1999; **27**(3): 187—207

21 康永国, 双志伟, 陶建华, 张维. 基于混合映射模型的语音转换算法研究, 声学学报, 2006; **31**(6): 555—562

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