语音转换及相关技术综述

李波[1,2]，王成友[1]，蔡宣平[1]，唐朝京[1]，张尔扬[1]

（1. 国防科技大学 电子科学与工程学院智能感知系统联合研究中心，湖南 长沙 410073；

2. 空军工程大学 电讯工程学院，陕西 西安 710077）

摘 要：给出了语音转换的定义，介绍了语音转换的用途，分析了表征说话人个性特征的语音参数，研究了语音转换的系统结构，对语音转换的实现主要从频谱包络和韵律两个方面的转换进行了研究讨论，分析并介绍了语音转换现在的发展水平及存在的问题。

关键词：语音处理；语音转换；频谱包络；韵律特征

中图分类号：TP391 **文献标识码：**A **文章编号：**1000-436X(2004)05-0109-10

A survey of voice conversion and its relevant technology

LI Bo[1,2], WANG Cheng-you[1], CAI Xuan-ping[1], TANG Chao-jing[1], ZHANG Er-yang[1]

(1. Joint Center for Intelligent and Sensing System, College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China;

2. The Telecommunication Engineering Institute, Air Force Engineering University, Xi' an 710077, China)

Abstract: The definition of voice conversion and its use are presented; the coefficients of speech's individual characteristics are analyzed; the structure of voice conversion is studied; the implementation of voice conversion is studied through spectral envelope conversion and prosodic conversion two parts; the status quo and problems of voice conversion are also introduced.

Key words: speech processing; voice conversion; spectral envelope; prosodic features

1 语音转换的概念及意义

语音转换（VC，voice conversion **或** VT，voice transformation**）是指改变一个说话人（源说话人，**source speaker**）的语音个性特征，使之具有另外一个说话人（目标说话人，**target speaker**）的语音个性特征**[1~7]**。语音包含很多信息，其中最主要的是语义信息，另外一个很重要的信息为语音的个性化信息。语音转换就是要保留原有语义信息不变，而改变语音的个性化信息，使一个人的语音经语音转换后听起来象是另外一个人说的语音。**

语音转换的用途是很广泛的，下面列举几个应用例子。

(1) **在文语转换（**TTS，text-to-speech**）系统中的应用。现有的** TTS **系统主要有共振峰合成（**formant synthesis**）、波形叠接相加合成（**PSOLA**）和基于数据库的合成等方法。不论是哪种方法，它们合成的语音的个性特征一般都是单一的，缺乏个性化，这也就限制了它的**

收稿日期：2003-06-25；**修订日期：**2003-11-01

应用，但如果将合成的语音再通过一个 VC 系统，或者将合成单元先通过 VC 转换，再进行 TTS 合成，将其转换为特定人的声音特征，使单调的合成语音具有更多的个性特征，也就使之应用更加广泛有效。例如，对于采用了 TTS 的有声 E-mail 系统，如果再采用 VC 技术，使有声 E-mail 的声音特征具有发送 E-mail 者的语音特征，这样 TTS 的应用就更加具有吸引力。这也正是 TTS 系统正在发展的一个方向。TTS 与 VC 的结合也是实现极低速率语音编码的有效方案。

(2) 在电影配音中的应用。在电影配音中，尤其是用另外一种语言进行配音时，往往配音演员不是演员本人，常常使配音与原演员的个性特征相差很大，配音效果不理想，但如果将配音再进行 VC 转换，使之重新具有演员本人的个性特征，那么配音效果就会理想的多。

(3) 语音转换思想可以用于恢复受损语音，帮助声道受损的说话人的语音提高可懂度。

(4) 语音转换可用于单个说话人的语音质量的控制，可以纠正在 TTS 中录音人长时间的录音而导致录音质量发生的变化。

(5) 可用于保密通信中进行语音个性化的伪装。

(6) 可以用于语音识别的前端预处理，以减少说话人差异的影响。

2 语音的个性化特征描述

表征语音个性化的语音特征可以分为以下三类：

· 音段特征：描述的是语音的音色特征。特征参数主要包括共振峰的位置、共振峰的带宽、频谱倾斜（spectral tilt）、基音频率、能量等。音段特征主要与发音器官的生理学和物理学特征有关，也与说话人的情绪状态有关。

· 超音段特征：描述的是语音的韵律特征。特征参数主要包括音素的时长、基音频率的变化（音调）、能量等。

· 语言特征（linguistic cues）：包括习惯用语、方言、口音等。

超音段特征和语言特征都是语音的很重要的个性特征，但对于说话人来说，超音段特征主要受社会和心理状况的影响[3]，容易随意的改变，例如，放慢说话速度、降低音量、说的更加柔软一些等；语言特征则与人的生活环境、成长过程和个人习惯有很大关系，随意性很大，不易对其建模。而音段特征与语音发音器官的生理学和物理学特征紧密相连，也与说话人的情绪状态有关，可以认为是不可改变的。

现在报道的语音转换系统，主要是对音段特征进行控制和转换；对于超音段特征如基音频率轮廓、能量轮廓、和说话人速率等特征一般都是进行平均值转换以与目标语音的平均特征值相匹配，之所以没有对超音段特征进行详细的建模、控制和转换，主要是由于在现在语音技术水平下，很难对高层的语音特征进行提取和操作。对于语言特征，在语音转换中几乎没有对其研究的报道。

对于各声学参数对语音的个性特征的贡献大小，Matsumoto[8]研究得出的结论是基音频率贡献最大，其次是共振峰频率，再次是基音频率的波动和声源频谱倾斜（voice source spectral tilt）；Furui[9]研究报道说由倒谱系数得到的长时平均谱包络对语音的个性特征贡献最大，特别是 2.5~3.5 kHz 频率范围的谱包络，平均基音频率为其次；Nakatsui 认为基音频率比声道的共振特性对语音的个性特征贡献大，而 Itoh[10]等则认为相反，他们认为频谱包络对语音的个

性特征影响最大，接着是基音频率和它的时间轮廓结构。各个语音参数对语音的个性特征的贡献大小的次序，虽然研究者们对此的结论不是完全相同，但可以肯定，无论哪个声学参数都无法包含所有语音的个性化信息，语音的个性特征是许多声学参数共同作用的结果，Kuwabara[11]认为声学参数的重要性因人而异，并且与实验的语音材料也有很大关系。

现在报道的语音转换系统中，用于转换的语音特征可以分为包含共振峰频率、共振峰带宽、频谱倾斜的表征声道滤波特性的频谱包络特征和包含基音频率、时长、能量的韵律特征两大类。

3　语音转换系统结构

语音转换的实现在总体上分为训练和转换两个阶段。

在训练阶段，系统基于某个语音模型对源语音（source speech）和目标语音（target speech）进行分析并提取语音特征，将这些语音特征进行对齐，再进行训练得到转换规则。图 1 所示，为语音转换系统结构图。

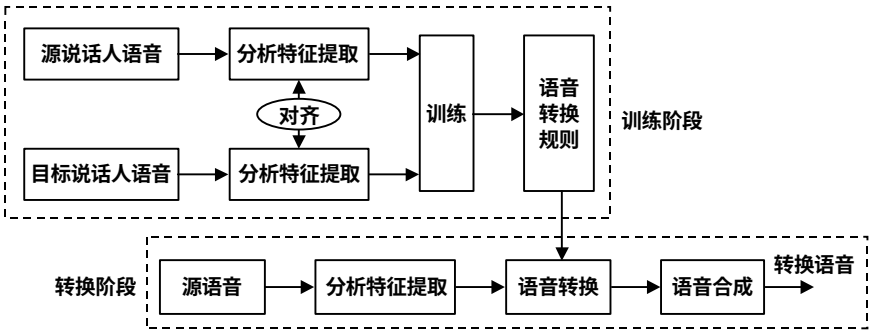

图 1　语音转换系统结构图

语音模型：选择的语音模型要能够准确提取语音频谱包络特征和韵律特征，能够准确有效的实现频谱包络和韵律特征的控制和转换，现在语音转换的研究大都采用声源-滤波（source-filter）的语音模型，将语音分解为声源激励部分和声道滤波部分，具体的说，所采用的语音模型主要有 LPC 语音模型和基于倒谱包络的语音模型。LPC 语音模型是应用较多的语音模型[1~3,5,7]，LPC 模型符合语音产生原理，它可以将语音有效的分解为谱包络部分（由 LPC 系数表示）和激励部分（由 LPC 的残差表示）。对于谱包络部分，由 LPC 系数得到的推演参数 LSF 可以与频谱包络的共振峰很好的对应，且控制和转换准确、有效、容易，在文献[1~3,5,7]中就是通过转换 LSF 的分布来实现频谱包络的转换；由 LPC 系数得到的伪对数面积比(PLAR, pseudo log area ratio)[12]和 PARCOR 系数[13]可以与声道的生理结构对应，通过对 PLAR 和 PARCOR 的转换也可以实现对 LPC 频谱包络的转换；Iwahashi[6]通过对 LPC 倒谱和对数面积比的转换来实现对频谱包络的控制转换；Lee[14]通过对 LPC 倒谱的处理来控制和转换频谱包络；Mizuno[15]则直接由 LPC 系数提取共振峰频率和频谱倾斜参数通过向量量化码书法来实现语音转换，Narendranath[16]也是先由 LPC 系数提取前三个共振峰频率再用神经网络法来实现转换。基于倒谱包络的语音模型也是一种有效的语音转换模型，Stylianou[4]和 Türk[17]采用基于倒谱包络的语音模型来实现对频谱包络的控制转换。对于韵律的转换，基于

LPC 的语音模型，可以对 LPC 残差进行韵律转换，这样还可以实现对声门波的转换，以达到高质量的语音转换，也可以通过在频域将分离掉谱包络部分所得的激励部分来实现韵律转换[17]。

在训练阶段都要先进行源语音和目标语音的分析和特征提取，提取语音的模型参数，源语音和目标语音对应于相同语音内容的语音特征要进行对齐，采用的方法有动态时间规整[1~4,6,18]、非监督 HMM 法[5,7]等。通过这些对齐的参数来估计转换规则，转换规则就是要捕捉源语音和目标语音特征之间的对应关系。

在转换阶段，首先对源语音进行分析并提取语音特征，再根据在训练阶段得到的语音转换规则进行转换得到转换的语音特征，由这些转换的语音特征合成出最终的转换语音。

如前面介绍，表征语音个性化的特征有很多，而共振峰的位置、共振峰的带宽和频谱倾斜都可以在频谱包络上体现出来，因此，下面语音特征的转换分为频谱包络和韵律两类来分别介绍其所采用的方法和目前发展现状。

4 频谱包络转换

频谱包络的语音转换是语音转换中最重要的一个方面，因为与频谱包络相关的共振峰位置、共振峰带宽和频谱倾斜对语音的个性特征贡献很大。频谱包络转换的主要方法将在下面进行介绍。

对于基于 LPC 语音模型的转换，LPC 系数都要先等效的转换为其推演参数 LSF、伪对数面积比、对数面积比、PARCOR 系数、LPC 倒谱等，然后再对这些 LPC 的推演参数进行转换，来实现频谱包络的转换。而对于基于倒谱包络语音模型的转换，则先要通过对语音进行频谱分析，得到倒谱系数或者 MFCC 系数，再进行转换。

4.1 向量量化法

Abe[1, 2]较早采用基于向量量化的码书映射方法来进行频谱包络的转换，Arslan[5,7]也采用这种方法。其实现过程如下：

首先对源说话人和目标说话人的语音频谱参数空间进行量化，使源语音和目标语音的码向量一一对应，分别得到 M 个源语音的码向量 u_{sk} (k=1,2,…M)和 M 个目标语音的码向量 u_{tk} (k=1,2,…M)。

然后在训练阶段通过训练得到由每一个源语音码向量 u_{sk} 到 M 个目标语音码向量 u_{tk} (k=1,2,…M)的映射码书 $\boldsymbol{H}$，$\boldsymbol{H}$ 为 $M\times M$ 的矩阵。映射码书的建立过程如下：

(1) 由源和目标说话人产生学习单词集，然后所有的单词逐帧进行向量量化。

(2) 用动态时间规整技术（DTW）对两个说话人的相同的单词向量进行对齐。

(3) 两说话人之间的向量对应关系累积成柱状图。应用柱状图作为加权系数，映射码书就为目标语音向量的线性合成时的加权系数。

在转换阶段，先将源语音的谱包络系数量化为源语音向量空间的第 l 个码向量。则转换的码向量 $\hat{y}$ 由式（1）得到

$$\hat{y}=\sum_{k=1}^{M}h_{lk}u_{tk} \tag{1}$$

其中，h_{lk} 为映射码书 H 的元素，满足 $\sum_{k=1}^{M}h_{lk}=1$，u_{tk} 为目标语音码向量。

4.2　说话人插值法

Iwahashi 和 Sagisaka[6]提出采用基于事先存储的多个说话人频谱包络进行插值的方法得到转换语音的频谱包络。频谱包络通过慢变化的插值率来进行平滑的转换。对于这种方法，少数几个说话人的语音频谱包络参数首先预存在合成系统中，多个说话人的相同语音的频谱序列首先进行动态时间规整（DTW），用式（2）进行插值实现转换

$$\hat{\boldsymbol{y}}^i = \sum_{k=1}^{M} w_k \boldsymbol{x}_k^i \tag{2}$$

其中，

$$\sum_{k=1}^{M} w_k = 1$$

$\boldsymbol{x}_k^i$ 表示第 k 个说话人的第 i 帧的频谱包络参数向量；M 表示预先存储的说话人数；w_k 表示第 k 个说话人插值率；$\hat{\boldsymbol{y}}^i$ 表示插值后得到的第 i 帧的频谱包络参数向量。

最佳插值率 $w_1, w_2, \cdots w_M$ 由式（3）函数的最小化来得到

$$F(w_1, w_2, \cdots w_M) = \sum_i (\hat{y}^i - y^i)^2 \tag{3}$$

其中，$\boldsymbol{y}^i$ 表示目标说话人的第 i 帧频谱包络参数向量。

4.3　线性多变量回归法

Valbret[18]提出采用线性多变量回归法（LMR, linear multivariate regression）来进行频谱包络转换。首先应用标准的 DTW 法将源语音和目标语音中提取的频谱包络特征参数进行对齐；然后应用标准的非监督分类技术将源说话人和目标说话人的声学空间分成非叠加的类；对每一类，由 LMR 得到一个简单的线性转换函数。

$\{\boldsymbol{x}\}_1^{N_k}$ 和 $\{\boldsymbol{y}\}_1^{N_k}$ 分别表示源语音和目标语音码书的第 k 类源语音频谱向量集和目标语音向量集。用 LMR 法估计一个 p 乘 p 阶矩阵 $\boldsymbol{P}_k$，它满足使归一化的源向量 $\tilde{x}_i$ 和目标向量 $\tilde{y}_i$ 的平方差最小。

$$\tilde{\boldsymbol{x}}_i = \frac{\boldsymbol{x}\grave{\boldsymbol{\imath}} - \quad x_k}{\boldsymbol{s}\ _{x_i}} \tag{4}$$

$$\tilde{\boldsymbol{y}}_i = \frac{\boldsymbol{y}\grave{\boldsymbol{\imath}} - \quad y_k}{\boldsymbol{s}\ _{y_i}} \tag{5}$$

其中，$\grave{\boldsymbol{\imath}}_{x_k}$、$\grave{\boldsymbol{\imath}}_{y_k}$、$\boldsymbol{s}_{x_i}$、$\boldsymbol{s}_{y_i}$ 分别表示该类的经验均值向量和方差。

矩阵 $\boldsymbol{P}_k$ 通过使误差最小化来得到

$$e = \sum_{i=1}^{N_k} \left\| \boldsymbol{P}_k \tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{y}}_i \right\|^2 \tag{6}$$

在转换阶段，首先将源语音频谱包络参数向量进行量化归类，确定所用的转换矩阵 $\boldsymbol{P}_k$、均值向量 $\grave{\boldsymbol{\imath}}_{x_k}$ 和方差 $\boldsymbol{s}_{x_i}$；再将源语音频谱包络参数向量进行“归一化”得归一化的频谱包络

参数向量，乘以转换矩阵；最后将得到的向量进行“解归一化”，即得转换的频谱包络参数向量。

4.4 动态频率规整（DFW）

Valbret[18]还提出了用 DFW（dynamic frequency warping）法来实现频谱包络的转换。实现过程为：首先应用标准的 DTW 法将源语音和目标语音中提取的频谱包络特征参数进行对齐；然后应用标准的非监督分类技术将源说话人和目标说话人的声学空间分成非叠加的类；然后在每一类内计算源语音和目标语音的对数幅度谱，去掉频谱倾斜，记录幅度值；根据源语音和目标语音频谱得到频率规整函数，它满足使式（7）中的频率归一化距离最小

$$D(\boldsymbol{S}^{t},\boldsymbol{S}^{s})=\min_{C}\left\{\left[\sum_{k=1}^{P}d^{t,s}(i(k),j(k))w(k)\right]\left[\sum_{k=1}^{P}w(k)\right]^{-1}\right\} \tag{7}$$

其中，C 为频率规整路径 $C=\{C_k\}(k=1,2,\cdots P)$，$C_k=(i(k),w(i(k)))$；$w(k)$ 为加权系数，用来对距离进行归一化；$d^{t,s}(i,j)=|S^{t}(i)-S^{s}(j)|$ 为频谱距离。

规整路径可以用规整曲线表示。规整函数的数目等于类中的频谱向量对的数目。对于每一类得到一个平均的规整函数，它可用一个三阶多项式来表示。

在转换阶段，首先将频谱包络参数归类，确定采用的转换函数；然后计算对数幅度谱包络，去掉频谱倾斜，进行转换，最后将目标语音的频谱倾斜加上，即得转换的频谱包络。

4.5 神经网络法

Narendranath[16]提出用神经网络来实现共振峰的转换。首先进行共振峰提取，采用最小相位群延迟函数（minimum phase group delay functions）法来提取前三个共振峰频率；对于三个共振峰频率的转换关系用神经网络来捕捉；在训练阶段将源语音和目标语音的三个共振峰频率参数分别作为神经网络的三个输入和三个输出，采用含有 8 个神经元的两个中间隐含层；在转换后合成时，将转换的共振峰频率和平均基音频率通过共振峰合成器合成出最终的转换语音。

4.6 高斯混合模型法（GMM）

Stylianou[4]和 Alexander[3]采用高斯混合模型法来实现频谱包络的转换。

在训练阶段用 GMM 法分别对源说话人和目标说话人的声学空间分 Q 类建模，如式（8）为对源说话人声学空间进行建模

$$P_{\mathrm{GMM}}(\boldsymbol{x}|\lambda)=\sum_{q=1}^{Q}a_q N(\boldsymbol{x};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q),\ \sum_{q=1}^{Q}a_q=1,\ a_q\geq 0 \tag{8}$$

其中，
$$N(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

a_q 表示 x 由 q 类产生的先验概率，$N(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma})$ 表示具有均值向量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 的 n 维正态分布。对于 GMM 参数 $\{a,\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ 用传统的 EM 法进行估计。

在转换阶段，首先求源语音特征向量 x 由第 q 类产生的概率 $p(c_q|x)$，其计算可由贝叶斯

准则得到，如式（9）

$$p(c_q \mid \boldsymbol{x}) = \frac{a_q N(\boldsymbol{x}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)}{\sum_{p=1}^{Q} a_q N(\boldsymbol{x}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)} \tag{9}$$

转换函数为

$$F(\boldsymbol{x}) = \sum_{q=1}^{Q} \left(\boldsymbol{\mu}_q^{Y} + \boldsymbol{\Gamma}_q \boldsymbol{\Sigma}_q^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_q^{X}) \right) p(c_q \mid \boldsymbol{x}) \tag{10}$$

$$\boldsymbol{\Gamma} = \mathrm{E}[(\boldsymbol{y} - \boldsymbol{v})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}]$$

$$\boldsymbol{\Sigma} = \mathrm{E}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{\mu} - \boldsymbol{\mu})^{\mathrm{T}}]$$

其中，v 为目标函数的均值向量。

上面列出了目前语音转换中主要的频谱包络转换方法，语音的频谱包络对语音的个性特征贡献很大，但对其转换却是一项很复杂的工作。

向量量化法是一种比较早的方法，它将转换规则根据源语音的特征空间首先进行分类，再对每一类通过训练建立相应的转换关系，是符合实际的，这种方法取得了较好的效果，但由于这种方法是将语音的频谱特征空间进行量化“硬”归类，则必然使转换的语音特征存在不连续性，从而使转换的语音质量下降；高斯混合模型法则对语音的频谱特征空间采用概率的方法进行“软”分类，就有效的克服向量量化法的不连续性，但同时也使频谱包络进行了平滑处理，使共振峰特性下降；Arslan[5,7]采用加权向量量化码书映射的方法，也在一定程度上克服了向量量化法不连续性；说话人插值法的转换性能要受到所选的参考说话人的影响，其优点是当训练语料很少时仍能得到较好的语音转换效果；DFW 法能改变语音频谱包络的共振峰的位置和带宽，但对频谱幅度却无能为力；LMR 法在转换共振峰位置时，同时影响了共振峰的幅度和带宽，而且在对声学空间进行分类时也引入了频谱信息的不连续性。上面提到的转换方法都对语音频谱包络转换起到积极的作用，但转换的频谱包络与目标语音频谱包络仍有一定的差别，高效的频谱包络转换法仍需要进一步研究。

在转换频谱包络中，Mizuno[15]和 Narendranath[16]提取共振峰频率和频谱倾斜等谱包络参数来分别进行转换；而大多数研究者则基于表征整个语音频谱包络的参数进行转换，这些频谱包络参数如 LSF、PARCOR、PLAR、对数面积比、LPC 倒谱、倒谱和 Mel 倒谱等；Arslan[5,7]采用调整 LSF 距离的方法来调整共振峰带宽。

5　韵律转换

韵律特征是表征语音个性化的重要特征，韵律转换也是语音转换的重要内容，韵律的转换内容主要包括基音频率的转换、时长的转换和能量的转换等。

基音频率的转换也即改变基音频率，不仅是语音转换的重要内容，也是文—语转换中得到高质量合成语音难度很大的一项关键技术。在文—语转换中要改变合成单元的基音频率，以使其具有不同的音调特性且要与包含这个单元的语音段的基音频率变化轮廓相匹配，但在这个工作中，研究者们发现当基音周期改变较大时，往往会导致合成语音听起来很机械或有

回声和杂音，从而导致语音质量的下降[19,20]。语音转换要转换源语音的基音频率为目标语音基音频率，基音频率改变常常会较大，所以，有效的基音周期改变算法是得到高质量转换语音的保证。

研究者们对基音周期转换的实现做了大量的工作，提出了很多实现方法，其实现方法总体上说主要有时域法和频域法，也有结合时域法和频域法来实现的；由于在原始语音信号的时域或频域上直接进行基音周期改变常常会引起谱包络的变化，且会发生较大失真，有效的基音周期变换法常常是基于激励-滤波模型，将激励源与声道的滤波特性分开，只改变激励源的基音周期来实现基音周期的改变。在时域中最简单易行的方法是 TD-PSOLA 法；George[21]给出了基于分析/合成-叠接相加语音模型的波形不变基音周期转换方法；Stylianou[4]采用的是基于 HNM 的基音周期转换法；Moulines[22]和 Türk [17]中给出了谐波删除-复制和频谱压缩-扩展两种 FD-PSOLA 基音频率转换方法；还有很多基音周期变换算法，在此不一一列举。

对于基音频率值的转换，Arslan[5,7]采用的是基于单高斯模型的均值/方差模型；Chappell[23]提出建立在句子基础上的基音轮廓码书的方法来确定转换的基音频率值。

对于时长的转换一般都是对基音周期的语音进行删除或复制来实现，同时伴随着幅度即能量的调整。在对韵律进行转换时，常常是同时进行基音周期、时长和能量转换的。

对于韵律值如基音频率变化轮廓、能量轮廓和语音速率的建模、控制和转换，在目前语音转换中主要是针对平均韵律值进行转换。没有对这些超音段特征进行详细的提取和转换的原因在于目前的语音技术还很难操作这些高层信息，虽然有关于时长模型[24,25]、声调模型等的报道，但仍不是很准确，且存在一定的争议。

6 总结

语音转换是一门比较新的语音信号处理领域的分支，语音转换有着很重要的意义。目前的语音转换取得了许多成果。在语音转换效果的评估中，ABX 测试法是应用较多的，ABX 测试即为说话人的区分测试，A 和 B 分别代表源语音和目标语音，X 代表转换语音，在测试中，要求测听者判断 X 更接近与 A 还是更接近于 B。

Abe[1]用 ABX 测试方法对实验结果进行测试，结果是 57%~65%的转换语音被认为更加接近目标语音（12 个测听者判断来自 3 个男声的 40 个句子）；Kain 和 Macon[26]研究与 TTS 相关联的 VC 系统，结果是男声→女声的转换中 97.5%的转换语音更加接近目标语音，男声→男声的转换中 52%的转换语音更加接近目标语音（20 个测听者判断 20 个句子）；Arslan[5,7]报道的结果为，男声→女声的转换为 100%，男声→男声的转换为 78%（3 个测听者判断 2~3 个句子）；Stylianou[4]报道的结果为 97%（20 个测听者判断 3 个句子）。但也应该注意，ABX 测试是有局限性的，即使是 100%的结果也不表示转换的语音与目标语音已没有区别。实际上转换的效果常常是转换的语音比原语音更加接近目标语音，而转换的语音还不能识别为目标说话人语音，通常被认为是第三人的语音。

现在的语音转换主要是基于音段的转换，对于超音段的转换研究的还比较少，对于语言特征的转换就更少，这样得到的转换效果是“粗”的，不能反映说话人的长时间动态特性。由于语音特征的修改，常常使转换后的语音质量下降，引入许多噪音和声音的失真等。

对于语音转换效果的提高，一方面要进一步研究语音产生的原理；另一方面要研究更加精确有效的语音模型，以便于更加方便和有效的对语音的个性特征进行控制和转换；语音转换得到的语音质量还有待提高。

参考文献：

[1] ABE M, NAKAMURA S, SHIKANO K, *et al*. Voice conversion through vector quantization[A]. ICASSP [C]. New York, 1988. 655-658.

[2] ABE M. A segment-based approach to voice conversion[A]. ICASSP[C].Toronto, Canada, 1991. 765-768.

[3] KAIN A. High Resolution Voice Transformation[D]. OGI School of Science and Engineering at Oregon Health and Science University, 2001.

[4] STYLIANOU Y. Harmonic Plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification[D]. École Nationale Supèrieure des Télécommunications, Paris, France,1996.

[5] ARSLAN L M. Speaker transformation algorithm using segmental codebooks (STASC)[J]. Speech Communication, 1999, 28(6): 211-226.

[6] IWAHASHI N, SAGISAKA Y. Speech spectrum transformation by speaker interpolation[A]. Proc IEEE Int onf Acoust, Speech, Signal Processing[C]. 1994. 461-464.

[7] ARSLAN L M, TALKIN D. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum[A]. Proceedings of the EUROSPEECH[C]. Rhodes, Greece, 1997. 1347-1350.

[8] MATSUMOTO H, HIKI S, SONE T, *et al.* Multidimensional representation of personal quality of vowels and its acoustical correlates[J]. IEEE Trans, 1973, AU-21: 428-436.

[9] FURUI S. Research on individuality features in speech waves and automatic speaker recognition techniques[J]. Speech Communication, 1986, 5(2): 183-197.

[10] ITOH K, SAITO A. Effects of acoustical feature parameters of speech on perceptual identification of speaker[J]. IECE Trans, 1982, J65-A: 101-108.

[11] KUWABARA H, SAGISAKA Y. Acoustic characteristics of speaker individuality: control and conversion[J]. Speech Communication, 1995, 16(2): 165-173.

[12] CORVELEYN S, COOSE B, VERHELST W. Voice modification and conversion using PLAR-parameters[A]. Proc 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA-2002)[C]. Leuven, Belgium, 2002. MPCA02-1-4.

[13] VERHELST W, MERTENS J. Voice conversion using partitions of spectral feature space[A]. ICCASSP[C]. Atlanta USA, 1996. 365-368.

[14] LEE K S, DOH W, YOUN D H. Voice conversion using low dimensional vector mapping[J]. IEICE Trans Inf & Syst, 2002, E85-D(8): 1297-1305.

[15] MIZUNO H, ABE M. Voice conversion algorithm based on piecewise linear conversion rules for formant frequencies and spectrum tilt[J]. Speech Communication, 1995, 16(2): 153-164.

[16] NARENDRANATH M, MURTHY H A, RAJENDRAN S. Transformation of formants for voice conversion using artificial neural networks[J]. Speech Communication, 1995, 16(2): 207-216.

[17] TÜRK O. New methods for voice conversion. Master Degree Thesis of Science[D]. Bogaziçi University, 2003.

[18] VALBRET H, MOULINES E, TUBACH J P. Voice transformation using PSOLA technique[J]. Speech Communication, 1992, 11(6): 175-187.

[19] TANAKA K, ABE M. A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F_0[A]. IEEE ICASSP[C]. 1997. 951-954.

[20] **初敏. 韵律研究与合成语音的自然度**[A]. **第五届现代语音学学术会议文集**[C].2001. 295-301.

[21] GEORGE E B, SMITH M J T. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model[J]. IEEE Transactions on Speech and Audio Processing, 1997, 5(5): 389-406.

[22] MOULINES E, CHARPENTIER F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones[J]. Speech Communication, 1990, 9: 453-467.

[23] CHAPPELL D T, HANSEN J H L. Speaker-specific pitch contour modeling and modification[A]. Proceedings of the IEEE ECASSP[C]. Seattle, Washington, 1998. 885-888.

[24] SHIH C, GU W, VAN SANTEN J P H. Efficient adaptation of TTS duration model to new speakers[A]. Proceedings of ICSLP[C]. Wydney, Australia, 1998. 177-180.

[25] GU W, SHIH C, VAN SANTEN J P H. An efficient speaker adaptation method for TTS duration model[A]. Proceedings of Eurospeech[C].Budapest, Hungary, 1999.

[26] KAIN A, MACON M. Spectral voice conversion for text-to-speech synthesis[A]. Proceedings of ICASSP[C]. Seattle, Washington, 1998. 285-288.

作者简介：

李波（1974-），男，山东青岛人，国防科学技术大学博士生，主要研究方向为语音信号处理。

王成友（1966-），男，四川东山人，博士，国防科学技术大学副教授、硕士生导师，主要研究方向为语音信号处理。

蔡宣平（1961-），男，福建尤溪人，国防科学技术大学教授、硕士生导师，主要研究方向为信号处理。

唐朝京（1962-），男，江苏武进人，国防科学技术大学教授、博士生导师，电子科学与工程学院副院长，主要研究方向为信号处理、编码。

张尔扬（1941-），男，北京人，国防科学技术大学教授、博士生导师，主要研究方向为信号处理。