声音转换技术的研究与进展

左国玉[1,2],刘文举[1],阮晓钢[2]
(1. 中科院自动化所模式识别国家重点实验室,北京 100080;2. 北京工业大学电子信息与控制工程学院,北京 100022)

摘 要: 声音转换是一项改变说话人声音特征的技术,可以将一人的语音模式转换为与其特性不同的另一人语音模式. 声音转换算法的目标是确定一个什么样的模式转换规则,使转换语音保持第一个说话人原有语音信息内容不变,而具有第二个说话人的声音特点. 本文介绍了当前声音转换技术领域的研究状态,主要分析现有声音转换技术中各种转换算法的实现原理,描述声音转换系统性能的各种评估方法,最后给出了对声音转换技术的简要评述和展望.

关键词: 声音转换; 语音频谱; 基频曲线; 声门激励; 码本映射; 人工神经网络; 高斯混合模型; 隐马尔科夫模型

中图分类号: TN912.3 **文献标识码**: A **文章编号**: 0372-2112 (2004) 07-1165-08

Voice Conversion Technology and Its Development

ZUO Guo-yu[1,2], LIU Wen-ju[1], RUAN Xiao-gang[2]
(1. *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080, *China*;
2. *School of Electronics Information and Control Engineering, Beijing University of Technology, Beijing* 100022, *China*)

Abstract: Voice conversion technology transforms one person's speech pattern into another pattern with distinct characteristics. The goal of a voice conversion algorithm is to achieve a transformation that makes the speech of the first speaker sounds as though it were uttered by the second speaker giving it a new identity, while preserving the original meaning. This paper introduces some current studies on voice conversion technology, which focus on various types of algorithms implemented in voice conversion area. Different evaluation methods for voice conversion performance are described. A technological outlook for this speech technique is given in the last section.

Key words: voice conversion; speech spectrum; pitch contour; glottal excitation; codebook mapping; artificial neural network; Gaussian mixture model; hidden Markov model

1 引言

在人们的日常生活交流中,一个人的声音往往就是他的身份名片,也就是通常所说的说话人身份(speaker identity). 说话人身份使人们仅从说话人的声音就能辨认出自己的亲戚朋友,在广播节目中听出是否是自己熟悉的主持人在主持节目. 《红楼梦》中王熙凤的出场描写可谓是"未见其人,先闻其声". 这些现象成为人们研究声音转换技术的最初出发点. 声音转换或声音个性化是一项改变说话人声音特性的技术,使得一人的声音听起来像是由另一人说出来的[1].

声音转换技术属于语音识别和语者识别技术的范畴. 自动语音识别预处理过程中的说话人自适应方法被广泛用在声音转换技术中[2,3]. 这项语音技术发展和延伸了说话人识别技术[4,5]. 1970 年代初,Atal 等人[6]就研究了使用 LPC 声码器改变声音特性的可行性. Seneff[7]研究了一种改变激励和声道参数的方法. Childers 等人[8,9]检验了男声变女声、女声变男声的方法. Abe 等人[10]提出了一种基于矢量量化(VQ)的码本映射技术. Iwahashi 等人[11]提出用频谱插值法增强码本映射技术的鲁棒性. Rinscheid[12]使用时变滤波器和拓扑特征映射实现了声音的改变. Valbret 等人[13]使用基音同步叠加法(PSOLA)调整激励信号中的韵律特征来改善转换性能. Narendranath[14]和 Watanabe[15]分别用 BP 和 RBF 等人工神经网络方法实现共振峰特性和 LPC 频谱包络的变换. 近年来,更多的研究人员致力于语音特征的统计分布来实现声音的转换[2,16-19]. 国内也开始出现这项语音技术的研究[3,20]. 所有这些先驱的工作极大地推动了声音转换技术的发展.

声音转换技术是对语音合成技术的丰富和延拓,有着良好的技术发展前景. 这项技术首先期望应用于特定文语合成系统[21]. 未来的系统会在人们接收 E-mail 或手机短信息时自动将信件内容用发信人的声音读出来. 扩展自然对话系统功能是这种应用的一种延伸. 特别是在娱乐和教育领域,产生多说话人特征的语音显示出很高的需求性,如戏剧、广播剧和电

收稿日期:2002-12-13;修回日期:2003-11-05
基金项目:国家自然科学基金项目(No. 60172055;No. 60121302);北京市自然科学基金(No. 4042025).

影里的角色配音(voice dubbing)等[22].

语音数据的采集与传输赋予声音转换技术以新的研究价值.传统的语料采集办法非常耗时费力,使用声音转换技术有可能使这个过程变得比较简单.如图 1 所示,语音合成系统从一个单说话人语音库中提取每一句话输入声音转换系统,分别采用不同目标说话人的模型,使新产生的语音具有所期望的多个目标说话人声音特性,从而建设成为一个由单人语音库生成的多说话人语音库.声音转换技术的优越性也将反映在超低带宽的语音编码领域.当语音编码系统设计的传输速率为 2.4kbs 或更低时,在传输过程中将不再保留说话人的语音特征[23].声音转换技术则有可能在接收方重现解码语音,使其与传送人的说话人特征相匹配.

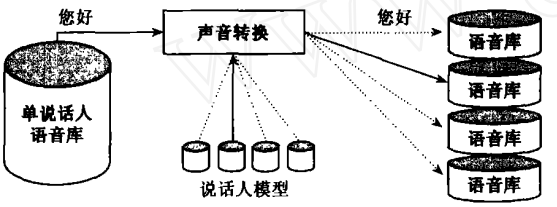

图 1 单人语音库生成多人语音库系统示意图

声音转换的另一个主要用途是用于说话人辨认技术.声音调整是多方会话翻译系统的一个重要技术内容[16,23].系统首先识别一方说话人的每一句话,然后用对方(另一方)语言翻译出来,再用本方说话人声音特征合成新的声音,这样使持不同语言的双方(多方)交流更为方便.在整个会话过程中维持转换语音的自然度是这项应用的重要技术要素.安全系统中的访问控制也激励了声音转换技术的进展[13].

一般来说,声音转换的技术实现主要包括以下几个要素:

(1)语音模型和特征:模型类型规定了系统要调整语音信号的哪方面参数.模型参数或特征由训练和转换过程中的语音分析阶段获得.

(2)映射规则:其作用是将源说话人的声学特征映射到一个近似于目标说话人的特征集上.

(3)语音库:在训练过程中用于训练数据和性能评估时用于测试的语音句子集合.

本文主要从这几个方面阐述当前声音转换技术领域的研究状态,以期能帮助对这项语音技术感兴趣的研究者有一个比较全面的了解,起到抛砖引玉的作用.

2 说话人特征与语音模型及其参数表示

语音信号中含有各种各样的信息,主要载有语音内容信息(what was said)、说话人特征信息(who said it)以及说话环境信息(where it was said).说话人特征描述了与说话人身份相关的声音方面特征,而与具体内容信息和说话环境无关.声音转换的任务就是要改变说话人特征,而其他方面的信息保留不变.一般地,说话人特征可表示为以下三个层面[24~26]:

(1)音段信息:短时声学特征,如共振峰位置、共振峰带宽、频谱倾斜(spectral tilt)、基频(F0)和能量,与音质相关,依赖于发音器官条件和情感状况.

(2)超音段信息:表示声学特征的时变演化,如平均基频、音素时长变化、语调变化和句中重读等,与说话风格和韵律相关,受到社会因素和心理状态的影响.

(3)语言学信息:说话时字词的选取、方言和口音等,在当前声音转换技术研究范围之外.

当某人说话时,超音段特征比较容易改变,说话人可以随意加快放慢语速、提高降低音量以及改变语气的轻重等.音段特征与语音产生器官的生理情况密切相关,因此可认为几乎不变.音段和超音段特征在说话人识别中具有很重要的感知性意义.在所有超音段特征中,平均 F0 和语音速度对说话人识别贡献最多,而在音段特征中,频谱包络和共振峰位置起主要作用.特别地,平均 F0 解释了 55%的辨别说话人能力,而 F0 与前三个共振峰和 F0 与频谱倾斜分别能够表示 71%和 85%的说话人特征变化[27].以短时频谱形式表达的音段特征和超音段特征的平均行为(主要是语速和平均 F0)能充分满足很大程度上的说话人区别,仅频谱包络就包含了丰富的说话人身份信息[28].因此当前声音转换系统主要集中在短时频谱包络参数的变换上,同时调整源说话人的基频、能量和语速从均值上匹配目标说话人的这些参数.

语音模型是语音信号的数学建模.在声音转换系统中,源-滤波器语音模型较好地表示了短时语音频谱,这种模型通过把频谱包络拟合到短时语音幅度谱上,将声道近似为一个缓变滤波器.转换算法常用的模型参数为 LPC[10,15]及其演变形式,如 LPC 倒谱系数[19,29]、线频谱频率(LSF)[17,30]等,以及进一步分析 LPC 频谱获得的共振峰频率和共振峰带宽[14,31,32].其它参数还包括美尔频标倒谱系数(MFCC)[3,16],美尔倒谱系数(MCC)[2,33].将语音信号用相应的 LPC 滤波器反向滤波就得到近似于声门激励波形的 LPC 残差.由于 LPC 残差仍然含有一定的说话人信息.因此,很多转换系统在进行频谱变换时为提高转换语音质量,同时考虑了对残差信号或声门激励波形的变换处理[29,30,34,35].

3 声音转换算法实现

声音转换系统通过改变语音信号的声学特征参数来调整语音.一般地,声音转换过程可以分为训练和转换两个步骤来进行[29,36],如图 2 所示.

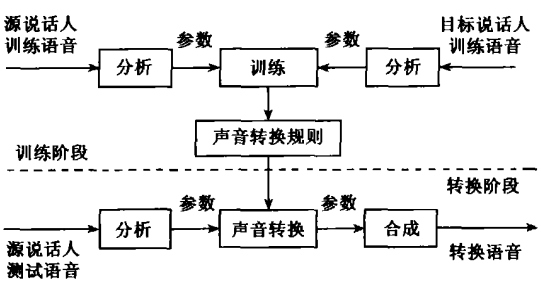

图 2 声音转换算法原理

在训练阶段,系统对源说话人和目标说话人的语音样本进行训练,估计映射规则,获取源语音和目标语音的模型参数之间的关系.在转换阶段,利用转换函数对源语音的音段特征和超音段特征等进行变换,使合成语音具有目标说话人特征.

在训练过程中,系统在一个特定的语音模型假设下分析源语音和目标语音.每一种变换算法都有一个提取模型参数的语音分析过程.在分析完成后,训练过程根据对应的语音将源-目标特征聚类分组,构造训练数据.特征关联属性(feature association)一般可由时间对准和分类过程得到,如动态时间规整(DTW)[10,13,16,29,31]、无监督隐马尔柯夫建模[30]或强制对准(forced-alignment)语音识别[30,37]等过程.经过时间对准后的数据被用来估计转换函数.在当前转换系统中已实现了多种语音频谱的转换算法,其中包括映射码本[10,37]、线性多变量回归[13]、动态频率规整[13]、人工神经网络[14,15]、高斯混合模型[16,17]和隐马尔科夫模型[2,19].转换时,已训练好的转换函数从新输入源语音特征来预测目标语音特征,最后在合成阶段,由预测特征产生最终的转换语音信号.源说话人的韵律特征如 $F0$ 曲线、能量曲线和说话速度也被调整,使之匹配目标说话人的韵律特征.

3.1　语音频谱变换

语音频谱承载了说话人特征的重要信息,调整语音频谱是当前声音转换技术的首要内容.训练频谱变换函数就是为了找到源、目标说话人声学特征之间的映射关系.一般地,训练前,需将源于两个说话人的特征矢量流采用某种算法进行时间对准,然后再根据映射方案训练频谱变换函数.

3.1.1　码本映射　码本映射是声音转换领域比较常用的转换算法.这种转换算法最早是由 Abe 和 Shikano 等人[10]提出来的,源于语音识别过程中的说话人自适应技术[38].图 3 显示了这种基于 VQ 码本映射的声音转换实现原理.

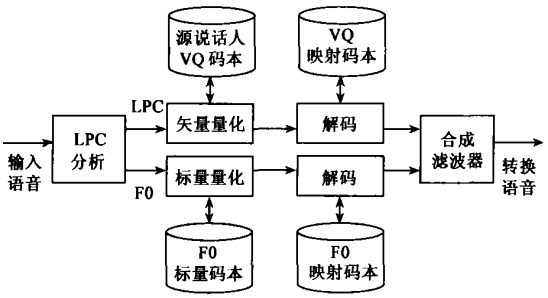

图 3　基于矢量量化码本映射的声音转换系统

在这个方案中,为产生映射码本,首先用矢量量化算法将源说话人和目标说话人的特征空间进行划分,用 DTW 算法将源矢量和目标矢量相关联,产生对应码本矢量的统计直方图.最终的目标码本定义为用直方统计值作为权函数的目标码字的线性组合.可以这样表达:一个与源说话人输入频谱 $X(S)$ 相对应的 VQ 频谱 $V_i(S)$ 经对目标说话人 VQ 频谱 $V_j(T)$ 加权求和(权值 h_{ij} 表示训练数据时 $V_i(S) \rightarrow V_j(T)$ 的对应统计值)转换成与 $V_j(T)$ 目标说话人的频谱 $X_i(T)$,表示如下:

$$X_i(T) = \sum_{j=1}^{N_i} h_{ij} V_j(T) \Big/ \sum_{j=1}^{N_i} h_{ij} \tag{1}$$

这种算法的一个基本问题是由于矢量量化作用造成的频谱的不连续性.为克服这种简单矢量量化方法的缺点,有人提出了模糊矢量量化技术(fuzzy VQ)[39].源说话人输入频谱 $X(S)$ 就不再唯一地量化成 $V_i(S)$,而是表达为 $V_i(S)$ 邻域码矢量的线性组合 $\sum_{i=1}^{M_i} u_i V_i(S)$,其中 u_i 是由 $X(S)$ 的模糊关系函数确定的系数.Abe 等人[22]又提出了分段矢量量化技术(segment VQ)的改进方案,用音素时长大小语音段来代替单帧 VQ 编码,语音段的切分采用了常用的 HMM 音素切分器.通过矢量量化可以获得比较精确的转换.

Arslan 等人[30,37]提出一种基于音素码本和滤波器思想的转换算法,较好地改善了转换信号连续语音帧之间的过渡性能和转换语音质量.这种方法采用语句 HMM (sentence HMM)方法代替 DTW 方法做音素对准,因而对准精度较高,鲁棒性较好.其训练过程如图 4 所示.

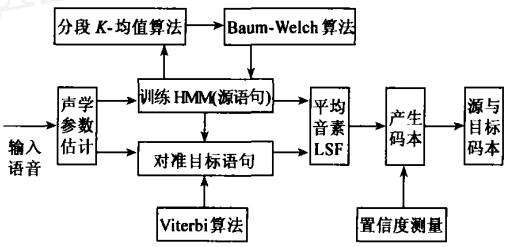

图 4　音素码本训练流程

源、目标码本生成后,与源频谱 $V_s(w)$ 对应的源语音 LSF 矢量被近似为源码本 LSF 矢量的加权组合,与目标频谱 $V_t(w)$ 对应的目标语音 LSF 矢量估计为相同权值加权的目标码本的线性组合,声道滤波器 $H_v(w)$ 就可表示为 $V_t(w)$ 与 $V_s(w)$ 的商.当输入语音的频谱为 $X(w)$ 时,输出语音就可表示为 $Y(w) = H_v(w) \cdot X(w)$.

还有一些改善码本映射算法性能的工作.文献[40]使用一个三层的神经网络实现映射码本.频谱插值法通过几个说话人语音频谱之间插值确定转换语音频谱以提高系统的鲁棒性[11],类似算法在文献[41]中也有描述.

整体上,上述码本映射方案还是受到转换质量不高和鲁棒性不太好等因素的困扰.

3.1.2　线性多变量回归和动态频率规整　有学者提出了不同于全局的码本映射思想的转换算法[13,31].用标准的无监督聚类技术(如 VQ)将说话人声学空间划分为多个不相重叠的类,每一类语音对应于一个转换函数(也称作局部函数),每个转换函数都表述了这一类中源-目标语音之间的映射关系,所以码本映射方案中的全局映射就被这些局部函数所近似.文献[13]中提出了多变量线性回归和动态频率规整算法两种局部转换算法.多变量线性回归(LMR)方法通过最小化每一类中所有源-目标矢量对之间预测误差的均方值来确定各最优线性变换矩阵 M:

$$C_C^k = M c_S^k \tag{2}$$

$$J = \min \sum_{k=1}^{N} (C_T^k - C_C^k)^2$$

最优解为:

$$M = C_T^{\mathrm{T}} C_S (C_S^{\mathrm{T}} C_S)^{-1} \tag{3}$$

其中,C_S^k、C_T^k 和 C_C^k 分别表示源倒谱矢量、目标倒谱矢量和由通过最小化性能指标 J 计算得到的最优化矩阵 M 变换

而得的转换矢量，C^T 代表矩阵 C 的转置，C_S 和 C_T 分别表示 N 个 p 维源矢量和目标矢量序列构成的矩阵. LMR 可解释成在源频谱矢量为高斯联合分布的假设下搜索目标矢量期望值.

动态频率规整算法(DFW)试图在同一声学类中找到源-目标语音频谱的映射路径. 这种方法首先计算每一源、目标说话人的对数幅度谱，并从中去除频谱倾斜(spectral tilt). 对归一化后的源、目标频谱采用一种频率规整算法，获得一条源-目标矢量对应关系的规整曲线. 每一类中规整函数的数量等于这一类的源-目标矢量对(vector pair)数目. 计算这一类中的平均规整函数，并用一个三阶多项式来表示. DFW 算法能在频域改变频谱形状，因此它能调整共振峰频率及其带宽，而其幅度几乎不受影响，但是转换性能稍逊于 LMR 算法[13].

Mizuno 等人[31]提出了类似局部变换的转换算法. 说话人频谱空间由矢量量化划分成许多不同的子空间，由此计算出一个共振峰线性转换规则集.

这种多个局部转换函数方法可以产生无穷个目标特征量. 但是由于选择单个局部转换函数的离散性还存在，所以不连贯性仍然出现在输出语音中.

3.1.3 人工神经网络模型 人工神经网络(ANN)是连续变换函数的一个例子. 理论上，一个具有非线性隐层的 ANN 能够逼近任意映射. 在连续语音中，声道系统特征变化迅速. 为比较真实地变换说话人的声学特征，码本映射方法中的码本尺寸就必须很大. 而在神经网络技术中，即使训练数据量较少但只要选取合适，也能较好地学习一个连续特征映射函数. 神经网络的这种泛化特性有助于降低数据储备要求而能较好完成说话人特性之间的变换. 根据上述原理，Narendranath 等人[14]借助于由 BP 算法训练的人工神经网络实现共振峰频率的变换. 共振峰变换函数的训练算法如下所示：

```
repeat
 for 每一共振峰数据集
  begin
  step1:将对应于源说话人(男声)的共振峰频率(F1, F2, F3)作为网络输入.
  step2:提取目标说话人(女声)语音对应帧的共振峰作为期望输出.
  step3:用 BP 算法调整权值.
  end
until 权值收敛
```

除了采用 BP 网捕获源-目标说话人声学特征之间的关系外，径向基(RBF)网络也可实现说话人之间的频谱变换[15]. 训练时，从训练集中分别抽出以 LPC 频谱表示对应的源说话人和目标说话人的语音音素，分别作为 RBF 网的输入和输出，采用最小二乘法最小化实际输出与期望输出的均方差来调整网络的联结权值.

尽管 ANN 表现出较好的连续性，但是很少有实验数据表明 ANN 方法能取得较佳的转换性能.

3.1.4 高斯混合模型 近年来，很多研究者采用概率方法改善转换语音的自然度和目标说话人特征倾向性. Stylianou 等人[16]用高斯混合模型(GMM)反映源特征分布和目标特征概率分布之间映射关系. 一个高斯混合模型被用来拟合源特征矢量 x 的概率分布，对源特征空间做“软”分类，表示如下：

$$p(x)=\sum_{i=1}^{m}\alpha_i N(x;\mu_i,\sum_i),\quad \sum_{i=1}^{m}\alpha_i=1,\alpha_i\geq 0 \tag{4}$$

其中，$N(x;\mu_i,\sum_i)$ 为第 i 个抽象声学类的正态分布，m 为高斯混合成分(mixture)的数目，α_i 为第 i 类的权系数. 根据贝叶斯理论，给定观察矢量 x，它属于第 i 类的概率为

$$h_i(x)=\frac{\alpha_i N(x;\mu_i,\sum_i)}{\sum_{j=1}^{m}\alpha_j N(x;\mu_j,\sum_j)} \tag{5}$$

参数 $(x;\mu,\sum)$ 由 EM 算法[42]来估计. 转换函数表示为

$$\hat{y}=F(x)=\sum_{i=1}^{m}h_i(x)[v_i+\Gamma_i(x-\mu_i)] \tag{6}$$

在源特征和目标特征相对应的基础上，通过求解最小二乘问题的正态方程估计每一个局部变换函数的参数 v_i 和 Γ_i，使在全部学习数据上的转换误差达到最小：

$$\varepsilon=E[\|y-\hat{y}\|^2] \tag{7}$$

Kain 和 Macon 等人[17]对上述算法做了一些改变，用一个 GMM 拟合由源矢量 x 和目标矢量 x 构成的联合矢量 $z=[x^T,y^T]^T$ 的概率分布 $P(x,y)$. 由给定 x 寻找 $E[y|x]$ 是一个回归：

$$\hat{y}=E[y|x]=\sum_{i=1}^{m}h_i(x)\left[\mu_i^y+\sum_i^{yx}\left(\sum_i^{xx}\right)^{-1}(x-\mu_i^x)\right],$$

$$h_i(x)=\frac{\alpha_i N(x;\mu_i^x,\sum_i^{xx})}{\sum_{j=1}^{m}\alpha_j N(x;\mu_j^x,\sum_j^{xx})} \tag{8}$$

其中 μ_i^x 和 μ_i^y 分别表示源、目标说话人第 i 类的均值矢量，$\sum^{xx}$ 表示源说话人第 i 类的方差，$\sum^{xy}$ 表示源说话人和目标说话人第 i 类的互方差. 联合矢量 z 的第 i 类的协方差和均值分别表示为

$$\sum_i^z=\begin{bmatrix}\sum_i^{xx} & \sum_i^{xy}\\ \sum_i^{yx} & \sum_i^{yy}\end{bmatrix},\quad \mu_i^z=\begin{bmatrix}\mu_i^x\\ \mu_i^y\end{bmatrix} \tag{9}$$

与最小二乘法相比，联合概率方法理论上能使回归问题的高斯混合成分得到更合理的配置，但在进行 EM 算法运算时的计算量要大很多.

试验表明，GMM 方法有效地改善了转换语音的自然度，结合韵律参数调整，可以极大提高转换语音的目标说话人特征倾向性.

3.1.5 隐马尔科夫模型 在有些语音合成系统中，声音转换算法采用了非特定人(SI)语音识别系统中广泛使用的说话人自适应技术[2,3]，如最大似然线性回归(MLLR)[43]、最大后验概率(MAP)[44]、矢量场平滑(VFS)[45]和 HMM 插值法[33]等技术. 在基于 HMM 的 TTS 系统[2]中，这种转换算法的基本原理是，音素 HMM 作为语音合成单元，初始的说话人无关音素 HMM(average voice HMM)在训练阶段由观察矢量训练而成. 使

用目标说话人语音对说话人独立的音素 HMM 做 MLLR 自适应,从而使自适应后的音素模型具有目标说话人特征. 合成时,将待合成的给定文本转换为上下文相关的音素标记序列. 根据标记序列,由经过自适应后的音素 HMM 单元拼接成语句 HMM. 这种算法用平均说话人声音取代了源说话人声音. 可以看到,HMM 转换算法与前文所述各种算法在实现原理架构上明显不同,在语音合成过程中几乎包含了应用自适应技术的语音识别过程,其技术基础是合成语音参数可由音素 HMM 中生成. 尽管由 HMM 做自适应后转换合成的输出语音性能还不是太好,但优点是只需少量数据做自适应就可以方便地合成不同目标说话人的语音.

3.2　激励信号变换

如前文所讨论,在源-滤波器语音模型中,声门激励信号或 LPC 残差信号仍含有一定的说话人特征信息. 当前声音转换系统对于激励信号的处理方法主要存在以下几种.

3.2.1　激励码本　Childers 等人[34]将声门激励信号分成浊音和清音两种,对于浊音部分,激励波形表示为一个六阶多项式模型,以多项式系数为矢量形成有一个 32 项浊音激励的声门激励码本;清音部分的噪声激励信号用一个 256 项随机产生的噪声激励码本表示. 进而类似频谱码本映射,将激励信号源码本映射到目标码本上再做韵律和频谱的调整[9].

在 Arslan[30]的激励码本滤波器方案中,激励映射码本由源、目标语音的 LPC 残差信号训练形成. 目标声门激励由激励码本滤波器 $U_i^t(w)/U_i^s(w)$ 的加权组合构建而成的滤波器 $H_g(w)$ 来估计,$U_i^t(w)$ 和 $U_i^s(w)$ 分别表示第 i 个源和目标的激励码本的幅度谱. 当输入语音为 $X(w)$、声道码本滤波器为 $H_v(w)$ 时,经过语音频谱和激励变换后,输出语音就可表示为 $Y(w)=H_g(w)^{*}H_v(w)^{*}X(w)$.

3.2.2　神经网络预测器　Lee 等人[29]将声学特征被分成线性和非线性两类:抽取了 LPC 倒谱系数作为线性部分的特征,而非线性部分表示激励信号,由一个用神经网络表示的长时延非线性预测器模型表示. LPC 倒谱系数矢量转换规则由正交矢量空间变换[46]确定,用平均基频比和映射码本实现声源信号的转换.

3.2.3　LPC 残差预测　Kain 等人[35]认为,对于语音的浊音部分,变换后的 LPC 频谱可预测目标 LPC 残差. 其基本假设是对于特定说话人语音相近的声学类,其残差信号也相似并且可以预测. 上述激励或残差信号的处理在不同程度上增加了转换语音的目标说话人特征倾向.

3.3　基频曲线变换

语音的韵律特征尤其是基频曲线含有大量的说话人身份信息,对确定说话人身份起了很重要的作用. 但是相对于声道相关的声学特征和激励信号,当前关于韵律变换的研究工作还比较少,其研究也主要集中在 F0 曲线的变换上,而能量曲线和语速等特征只是简单地做均值线性变换. F0 曲线变换主要表现为以下几种方法,

3.3.1　均值线性变换模型　最简单 F0 曲线建模的方法是认为 F0 服从高斯正态分布,由此可估计其均值和方差,从而实现输入基频值到期望基频值的映射[18,29,30,32,47],其映射关系可表示为

$$x_2=\frac{x_1-\mu_s}{\sigma_s}\cdot\sigma_t+\mu_t \tag{10}$$

其中,μ_s 和 σ_s 分别为源说话人基频的均值和标准差,μ_t 和 σ_t 分别为目标说话人基频的均值和标准差.

3.3.2　确定性/随机性混合建模　Ceyssens 等人[48]提出一种确定和随机性混杂建模 F0 曲线的方案. 对每一句话求取对数域上基频的回归拟合直线,确定基频偏置 Po 和下倾斜率 Pd,估计回归线与实际 F0 曲线的方差 V;类似求取 (Po,Pd,V) 中各元素对句子长度 L 的回归关系,分别获得 Po、Pd 和 V 关于 L 的偏置、斜率和方差等九个参数. 基频曲线偏置的变换关系可表示为

$$PCo=(PTo_o+L\cdot PTo_d)+PTo_v\frac{Po-(PSo_0+L\cdot PSo_d)}{PSo_v} \tag{11}$$

式中,Po 表示输入语句的基频偏置,S、T 和 C 分别代表源、目标和转换语句,下标 o,d 和 v 分别表示偏置、斜率和方差. 转换语音基频的下降斜率 PCd 和方差 VC 可类似求出. 式(11)可认为是式(10)的一阶形式.

3.3.3　逐段线性映射　Gillet 提出了一种基频逐段(piecewise)线性映射的方法[49]. 首先确定每一语句的四个基频点:句首高位值 S,非句首峰值 H,重音后谷底值 L 和句末低位值 F. 当分别有一段通过点 (F_s,F_t) 和 (L_s,L_t),一段通过 (L_s,L_t) 和 (H_s,H_t),另一段通过 (H_s,H_t) 和 (S_s,S_t) 时,逐段求取源-目标 F0 的映射关系,表示如下:

$$M(x)=\begin{cases}F_t+\dfrac{(x-F_s)(L_t-F_t)}{(L_s-F_s)}, & x<L_s\\[2ex] L_s+\dfrac{(x-L_s)(H_t-l_t)}{(H_s-L_t)}, & L_s<x<H_s\\[2ex] H_s+\dfrac{(x-H_s)(S_t-H_t)}{(S_s-H_s)}, & x>H_s\end{cases} \tag{12}$$

逐段线性映射方法效果要优于式(10)的基频转换效果.

3.3.4　语句基频曲线码本　用整句水平上生成的语句基频曲线码本建模一个说话人基频特征是可能的[50]. 这种方法使用 DTW 算法来比较测试语句和相同说话人数据集中所有语句,从中找到 F0 曲线匹配最近的语句后,选取期望说话人的相同语句,再将所选的这两条语句以音素或词边界为中间约束做 DTW,所获得的 F0 曲线再规整到测试语句,就产生了一条用于转换语音的新 F0 曲线. 该算法优点在于存在使用真实 F0 曲线进行合成或调整的可能性,但是它只局限于小词汇量的专用场合,生成所有的 F0 曲线则不切实际.

Türk[51]提出了类似的分段(segmental)语音基频曲线映射方法.

3.3.5　高斯混合模型　文献[52]中提出与上述算法都不同的思想,认为 F0 和 LSF 频谱是类相关的,因此可用一个高斯混合模型来描述输入语音这种广义类的分布,在改变语音幅度频谱的同时调整了基频参数. 在 F0 调节因子较大时,这种方法在一定程度上可以改善转换语音的自然度.

从现有的语音技术来看,高层次信息如基频-声调的抽取和控制还存在不少困难,因此还不能很好地对具体的超音

段特征建模和变换.

3.4 语音库设计

语音库的作用是为训练转换函数和使用客观和主观评估方法测试声音转换系统的性能提供必要的语音数据. 语音库的设计是成功的声音转换技术的重要因素,它涉及四个主要方面[36]:

3.4.1 语音库大小 规定了库中可获得的每个说话人数据量的多少. 一个语音库可能包含少到只有五个元音字母的内容[14,34]、一个单词集[10,13,41]、短句[53]或者长于一小时的文章朗读语料[18,30].

3.4.2 语音库内容 评判的重要标准是看所描述的语音覆盖范围(phonetic coverage),也就是语音库中说话人的发音是否覆盖了可能的话语声音空间,例如音素、双音子、三音子等.

3.4.3 说话人数目 在目前关于声音转换技术的语音库中,说话人的数目也从最少两位至最多六位不等. 一般地,一个大的说话人数目有利于声音转换系统的评测.

3.4.4 时间对准 在训练转换系统时,需将具有对等语言学特性的源和目标特征相关联. 为减少训练时间,一般可将语音库中各说话人的对应语句在训练前用相关算法做好对准标注. 时间对准是特征关联的一种有效方法.

4 声音转换系统性能评估

声音转换性能评估构成了声音转换技术的重要组成部分. 合成语音可以根据清晰度、自然度和使用场合的实用性来比较评估[54,55]. 对于转换语音,适用性指标是指说话人的可识别性,即目标说话人特征倾向性. 语音质量是一个多维性术语,还没有一种评估体系能够完全适合或自动评测合成语音或转换语音的质量. 下面简要介绍现有的一些转换语音性能的主观和客观评测方法.

4.1 主观评估

主观性能评估主要有以下三种方法[16,17,31].

4.1.1 ABX 测试 在主观评测中,说话人的 ABX 测试方法被广泛采用. 激励 X 表示转换语音,激励 A 和 B 或是源说话人的语音或是目标说话人的语音,话语内容相同但与 X 不同. 每一个三元组就是这样三句话语的组合. 参与测试的听者要求对 A 和 B 中的哪一个与 X 的声音最相似做出选择. 尽管可能有 100 %的选择正确率,但是转换语音也不能被认为是目标说话人说出来的.

4.1.2 倾向性测试 倾向性测试(preference test)评价由两种不同算法(如码本映射和高斯混合模型)实现的转换语音中,哪一种最接近于期望的说话人声音特征. 测试也是用 ABX 方法实现,其中 X 是自然语音,A 和 B 分别是两种算法实现的转换语音对(speech pair),话语内容相同但与 X 不同. 听者被要求在转换语音对中做出倾向性选择. 倾向测试是对语音转换系统性能的一种横向评估方法.

4.1.3 观点测试 观点测试是为了评估转换算法的整体性能而设计的. 语音对包含所有的源、目标说话人的语音和由不同算法实现的转换语音,不同的句子用来组成这些语音对. 要求听者按照不同的等级对每一对语音的相似性打分,不同的等级用来表示语音对中的语音相似性[16,31]. 观点测试比较好地反映了转换语音的综合质量,包括清晰度和自然度等.

还有一些其他主观评价方案. 如 *AB* 测试法[37]对上述语音对中的语音是否相同所做的选择进行统计,统计值反映了各语音对中的语音相似性.

4.2 客观评估

转换语音的客观性能评估一般建立在语音幅度谱计算的基础上,主要有以下三种评价标准[10,49,30].

4.2.1 频谱失真度 转换语音与目标语音之间的频谱失真距离为一项客观评测内容被广泛采用,其基本形式可表示如下:

$$D = \sum_{i=1}^{p} (C_i - C_i^*)^2 \tag{13}$$

常用频谱距离都可用上式及其各种演变形式来表示. 其它如 Itakura-Saito 距离等似然失真形式也常用来做语音距离的客观评价.

4.2.2 信噪比 语音编码中最常用的信号噪声比(SNR)也被借用到转换语音性能的客观评测中,表示如下:

$$\mathrm{SNR}(s_1, s_2) = 10\log_{10}\frac{\sum |\mathrm{FFT}(s_1(n))|^2}{\sum (|\mathrm{FFT}(s_2(n))| - |FFT(s_1(n))|)^2} \tag{14}$$

上式表明,SNR 值越大,转换效果则越佳.

4.2.3 说话人辨识 其主要思想是,将转换语音作为说话人识别系统的输入,以确定目标说话人辨识的似然性. 在这里,目标与源说话人的对数似然比被用于说话人决策的置信度测量:

$$\theta_{st} = \log\frac{P(X|\lambda_t)}{P(X|\lambda_s)} = \log P(X|\lambda_t) - \log P(X|\lambda_s) \tag{15}$$

式中,λ_t 和 λ_s 分别表示目标和源说话人的模型. X 是观察矢量. 与 λ_s 模型相比,转换语音被 λ_t 模型赋予较高的似然性,θ_{st} 在数值轴上表现出向右较大的移位,转换算法性能就较好.

实验结果表明,客观频谱的差别并非一定存在相应的感官性能差异,说明了转换语音的客观度量与感官评价之间的弱联系性[16,19].

5 总结与展望

本文介绍了声音转换技术领域的研究现状,分析了当前声音转换技术中所采用的各种转换算法原理. 整体上,统计方法中 GMM 转换技术的性能要优于其他方法实现的频谱变换函数. 声门激励信号与韵律特征的调整能有效地改善转换语音性能. 理想系统应能产生高质量、很容易辨别为目标说话人的转换语音. 尽管声音转换技术前景诱人,但至今还未有真正实用的系统投入使用. 因此,语音领域的研究人员仍需提高转换性能以求系统之实用性. 对此本文做如下几点简要评述和展望:

(1) 频谱参数,如共振峰频率、共振峰带宽、频谱包络和频谱倾斜等音段信息主要决定说话人的声音音色,包含了绝大部分说话人信息. 现有转换方法也基本集中在这些参数上. 各

种算法的测试结果均发现转换语音与目标语音的频谱包络吻合性不是太好,转换语音与目标语音的音质差别较大,辨认率不高.在改进现有或提出新转换算法的基础上,发现新的说话人相关的声学特征参数将是提高转换语音质量问题的可能解决方案.

(2)语音超音段特征在辨识说话人身份的过程中起着重要作用,而现有韵律转换算法大多只是将某些韵律特征如 F0 曲线、能量曲线和音素时长变化等做简单地线性化变换,未能有效捕获含有说话人信息的 F0 曲线、能量曲线等韵律特征.因此,转换算法的性能飞跃有可能产生在对说话人相关的具体韵律特征的建模和转换上,如提出高精度基频跟踪算法、对生成 F0 曲线、音素时长变化、协同发音和语调与声调更准确的建模等.

(3)激励信号中含有一定的说话人信息,而且已有不少学者提出了转换激励信号特征的方法,但是研究人的发声机理和听觉机理,进而对激励信号建立更有效的模型仍是提高转换语音性能的一个重要方面.

(4)除朗读风格以外,多种说话风格和情感语音合成[56]也是声音转换技术的研究方向.它不仅涉及到韵律的建模和变换,而且对更高层次上语言学方面的语音风格研究提出了要求.

(5)一个精心设计的转换语音库是声音转换系统具有高性能的重要前提,但是目前的语音库都是根据各自需要设计出来的,其内容、大小不一.建设一个多层次、标注较完整的通用转换语音库是改善转换性能的基础.

参考文献:

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

[2] M Tamura, et al. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR[A]. Proc ICASSP[C]. Salt Lake City, USA: IEEE, May 2001. 805 - 808.

[3] Y Chen, M Chu, et al. Voice Conversion with Smoothed GMM and MAP Adaptation[A]. Proc Eurospeech[C]. Geneva, Switzerland: ISCA, Sept. 2003:2413 - 2416.

[4] I Karlsson. Glottal waveform parameters for different speakers types [A]. Proc Speech '88, 7th FASE Symp[C]. Edinburgh, Scotland: IOA, 1988. 225 - 231.

[5] B S Atal. Automatic recognition of speaker from their voices[J]. Proceedings of the IEEE, April 1976, 64(4):460 - 475.

[6] B S Atal, S L Hanauer. Speech analysis and synthesis by linear prediction of the speech wave[J]. J Acoust Soc Am, 1971, 50(2):637 - 655.

[7] S Seneff. System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction[J]. IEEE Trans Acoust Speech Sig, 30(4), 1982:566 - 578.

[8] D G Childers, et al. Factors in voice quality: Acoustic features related to gender[A]. Proc ICASSP[C]. New York, USA: IEEE, 1987. 293 - 296.

[9] D G Childers, et al. Voice Conversion[J]. Speech Communication, 1989, 8(2):147 - 158.

[10] M Abe, et al. Voice Conversion through Vector Quantization[A]. Proc ICASSP[C]. New York, USA: IEEE, 1988(1). 655 - 658.

[11] N Iwahashi, et al. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks[J]. Speech Communication. 1995, 16(2):139 - 151.

[12] A Rinscheid. Voice conversion based on topological feature maps and time-variant filtering[A]. Proc ICSLP[C]. Philadelphia, USA: ESCA, Oct. 1996. 1445 - 1448.

[13] H Valbret, et al. Voice transformation using PSOLA technique[J]. Speech Communication, 1992, 11(2 - 3):175 - 187.

[14] M Narendranath, et al. Transformation of formants for voice conversion using artificial neural networks[J]. Speech Communication, 1995, 16(2):207 - 216.

[15] T Watanabe, et al. Transformation of spectral envelope for voice conversion based on radial basis function networks[A]. Proc ICSLP '2002[C]. Denver, USA: ISCA, Sept. 2002. 285 - 288.

[16] Y Stylianou, et al. Continuous probabilistic transform for voice conversion[J]. IEEE Transactions on Speech and Audio Processing. March 1998, 6(2):131 - 142.

[17] A Kain, M Macon. Spectral voice conversion for text-to-speech synthesis [A]. Proc ICASSP [C]. Seattle, USA: IEEE, May 1998(1). 285 - 288.

[18] T Toda, et al. STRAIGHT-based voice conversion algorithm based on Gaussian mixture model[A]. Proc ICSLP[C]. Beijing, China: ESCA, Oct. 2000. 279 - 282.

[19] E K Kim, et al. Hidden markov model based voice conversion using dynamic characteristics of speaker[A]. Proc Eurospeech[C]. Rhodes, Greece: ESCA, 1997. 2519 - 2522.

[20] W Zhang, et al. Voice conversion based on acoustic feature transformation[A]. Proc NCMMSC[C]. Shenzhen, China, 2001. 189 - 192.

[21] R Carlson, et al. Experiments with voice modeling in speech synthesis [J]. Speech Communication. 1991, 16(5 - 6):481 - 489.

[22] M Abe. A segment-based approach to voice conversion[A]. Proc ICASSP[C]. Toronto, Canada: IEEE, May 1991. 765 - 768.

[23] A Schmidt-Nielson, D P Brock. Speaker recognizability testing for voice coders[A]. Proc ICASSP[C]. Atlanta, USA: IEEE, May 1996. 1149 - 1152.

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

[25] D Klatt and L C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers[J]. J Acoust Soc Am, 1990, 87(2):820 - 857.

[26] P H Milenkovic. Voice source model for continuous control of pitch period[J]. J Acoust Soc Am, 1993, 93(2):1087 - 1096.

[27] H Matsumoto, et al. Multidimensional representation of personal quality of vowels and its acoustical correlates[J]. IEEE Trans Audio and Electroacoustics, 1973, 21(5):428 - 436.

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

[29] K S Lee, et al. A new voice transformation based on both linear and

nonlinear prediction[A]. Proc ICSLP[C]. Philadelphia, USA: ESCA, 1996. 1401 - 1404.

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

[31] H Mizuno and M Abe. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt [J]. Speech Communication. 1995, 16(2): 165 - 173.

[32] J Gutiérrez, et al. A new multi-speaker formant synthesizer that applies voice conversion techniques [A]. Proc Eurospeech[C]. Aalborg, Denmark: ISCA, 2001: 357 - 360.

[33] T Yoshimura, et al. Speaker interpolation in HMM-based speech synthesis system [A]. Proc. Eurospeech [C]. Rhodes, Greece: ESCA, 1997. 2523 - 2526.

[34] D G Childers. Glottal source modeling for voice conversion [J]. Speech Communication. 1995, 16 (2): 127 - 138.

[35] A Kain, M Macon. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction[A]. Proc ICASSP[C]. Salt Lake City, USA: IEEE, June 2001. 813 - 816.

[36] A Kain. High resolution voice transformation[D]. Portland, USA: Oregon Health & Science University, Oct. 2001.

[37] L M Arslan, D Talkin. Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum[A]. Proc Eurospeech [C]. Rhodes, Greece: ESCA, 1997(3). 481 - 489.

[38] K Shikano, et al. Speaker adaptation through vector quantization[A]. Proc ICASSP[C]. Tokyo, Japan: IEEE, 1986. 2643 - 2646.

[39] S Nakamura, K Shikano. Spectrogram normalization using fuzzy vector quantization[J]. J Acoust Soc Japan, 1989, 45(2): 107 - 114.

[40] M I Savic, I H Nam. Voice personality transformation[J]. Digital Signal Processing Journal. 1991, 1(2): 107 - 110.

[41] M Hashimoto, N Higuchi. Spectral mapping for voice conversion using speaker selection and vector field smoothing[A]. Proc Eurospeech[C]. Madrid, Spain: ESCA, 1995: 431 - 434.

[42] N Kambhatla. Local Models and Gaussian Mixture Models for Statistical Data Processing[D]. Portland, USA: Oregon Graduate of Institute of Science and Technology, 1996.

[43] C J Leggetter, P C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models [J]. Computer Speech and Language, 1995, 9(2): 171 - 185.

[44] Lee Chin-Hui, et al. A study on speaker adaptation of the parameters of continuous density hidden markov models [J]. IEEE Trans on Signal Processing, 1991, 39(4): 806 - 814.

[45] K Ohjura, et al. Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs [A]. Proc ICSLP [C]. Banff, Canada: ESCA, Oct. 1992. 369 - 372.

[46] K S Lee, et al. Voice personality transformation using an orthogonal vector space conversion[A]. Proc Eurospeech [C]. Madrid, Spain: ESCA, 1995(1). 427 - 430.

[47] Y Stylianou, et al. Statistical methods for voice quality transformation [A]. Proc Eurospeech[C]. Madrid Spain: ESCA, 1995. 447 - 450.

[48] T Ceyssens, et al. On the Construction of a Pitch Conversion System [A]. Proc EUSIPCO[C]. Toulouse, France: EUSIP, 2002. 1301 - 1304.

[49] B Gillett, S King. Transforming voice quality[A]. Proc Eurospeech [C]. Geneva, Switzerland: ISCA, 2003. 1713 - 1716.

[50] D Chappell, J Hansen. Speaker-specific pitch contour modeling and modification[A]. Proc ICASSP[C]. Seattle, USA: IEEE, May 1998. 885 - 888.

[51] O Türk. New Methods for Voice Conversion[D]. Istanbul, Turkey: Boğazici University, 2003.

[52] A Kain, Y Stylianou. Stochastic modeling of spectral adjustment for high quality pitch modification[A]. Proc ICASSP[C]. Istanbul, Turkey: IEEE, June 2000. 949 - 952.

[53] L M Arslan, et al. Speaker transformation using sentence HMM based alignments and detailed prosody modification[A]. Proc ICASSP[C]. Seattle, USA: IEEE, May 1998. 289 - 292.

[54] D Klatt. Review of text-to-speech conversion for English [J]. J Acoust Soc Am. 1987, 82(3): 737 - 793.

[55] A Mariniak. A global framework for the assessment of synthetic speech without subjects[A]. Proc Eurospeech[C]. Berlin, Germany: ESCA, 1993(3): 1683 - 1686.

[56] M Schröder. Emotional speech synthesis-A review[A]. Proc Eurospeech [C]. Aalborg, Denmark: ISCA, 2001(1): 561 - 564.

作者简介:

左国玉 男,1971 年 10 月生于安徽,博士研究生,主要研究领域为声音转换、语音信号处理、模式识别、人工智能.

刘文举 男,1960 年 4 月生于北京,副研究员,主持多项国家自然科学基金和 863 计划课题,主要研究领域为汉语连续语音识别、语音合成、听觉计算模型、语音识别神经计算、马尔柯夫类模型实用快速算法及人工智能方法进行规划、决策等.

阮晓钢 男,1958 年 4 月生于四川,教授、博士生导师,多次承担国家和省部级科研项目,主要研究领域为自动控制、模式识别、神经计算学、人工智能、智能信息处理.