采用模型自适应的语音转换方法

宋 鹏 王 浩 赵 力
(东南大学 信息科学与工程学院，南京 210096)

摘 要：针对非对称语音库情况下的语音转换，提出了一种有效的基于模型自适应的语音转换方法。首先，通过最大后验概率(Maximum A Posteriori，MAP)方法从背景模型分别自适应训练得到源说话人和目标说话人的模型；然后，通过说话人模型中的均值向量训练得到频谱特征的转换函数；并进一步与传统的 INCA 转换方法相结合，提出了基于模型自适应的 INCA 语音转换方法，有效实现了源说话人频谱特征向目标说话人频谱特征的转换。通过客观测试和主观测听实验对提出的方法进行评价，实验结果表明，与 INCA 语音转换方法相比，本文提出的方法可以取得更低的倒谱失真、更高的语音感知质量和目标倾向度；同时更接近传统基于对称语音库的高斯混合模型(Gaussian Mixture Model，GMM)的语音转换方法的效果。

关键词：模型自适应；语音转换；非对称语音库
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1003－0530(2013)10－1294－06

Voice Conversion Method Based On Model Adaptation

SONG Peng WANG Hao ZHAO Li
(School of Information Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: In order to realize voice conversion using non-parallel corpus, an efficient voice conversion method based on model adaptation is proposed in the paper. Firstly, the source and target speaker models were trained from background model using Maximum a Posteriori (MAP) adaptation algorithm, respectively. Then, a conversion function was trained by using mean vectors of adapted speaker models, and in order to improve the conversion performance, the conversion function was combined with INCA conversion algorithm, and a model adaptation based INCA method was further presented. The proposed method could efficiently transform the spectral features from source speaker to target one. Subjective and objective experiments were carried out to evaluate the performance of the proposed method, the results demonstrate that the proposed method obtains lower cepstral distortion, higher perceptual quality and similarity than INCA method. Meanwhile, compared with INCA algorithm, the proposed method using non-parallel speech corpus can achieve more comparable performance to Gaussian Mixture Model (GMM) based voice conversion method using parallel speech corpus.

Key words: model adaptation; voice conversion; non-parallel speech corpus

1 引言

语音转换指的是通过改变一个说话人的个性特征，使之具有另一个说话人的个性特征，而文本内容保持不变的一种技术。Abe 等学者[1]最早通过码本映射的方法对频谱特征进行匹配转换，但是由于它是在离散的空间上进行的，造成转换后的语音质量比较差；Stylianou[2]和 Kain[3]等针对码本映射方法的不足，提出了基于高斯混合模型(Gaussian Mixture Model，GMM)的语音转换方法，在一定程度上取得了较为满意的转换效果；Toda[4]和 Godoy[5]等将动态频率弯折(Dynamic Frequency Warping，

收稿日期：2013－04－27；修回日期：2013－08－08
基金项目：国家自然科学基金(面向非特定说话人的实用情感语音特征分析与识别的关键技术及应用研究，61273266；汉语数字助听器语音处理核心算法研究，60872073)

DFW) 方法用于频谱特征转换,取得了较其他方法更高的语音质量; 文献 [6] 和 [7] 分别将隐马尔可夫模型(Hidden Markov Model,HMM) 方法、人工神经网络(Artificial Neural Network,ANN) 方法用于语音转换,这些方法都在一定程度实现了说话人个性特征的转换。

但是,上述方法都是针对对称语音库的情况提出的。在现实环境中,对称的语音库很难录制和获取。针对这个问题,有学者提出了基于非对称语音库情况下的语音转换方法。Mouchataris [8] 等给出了基于最大似然双线性回归的语音转换方法,这种方法在很大程度上依赖于基于第三方参考对称语音库训练得到的转换函数的准确性; Popa [9] 等提出了基于说话人个性特征和语义信息分离的语音转换方法,但是从目前的实验结果来看很难对个性特征和语义信息进行有效分离; 俞一彪 [10] 等提出了基于独立说话人模型的语音转换方法,通过对每一个说话人进行 GMM 建模,然后运用全局声学结构(Acoustical Universal Structure, AUS) 进行概率分布的对准,但是这种方法需要大量的训练数据来保证说话人建模的准确性; Erro [11] 等提出了 INCA(Iterative Combination of A Local Nearest Neighbor Search Step and A Conversion Step Alignment) 算法用于频谱特征的对齐与转换,它是建立在源说话人和目标说话人特征空间中距离相近的两点对应着相同的音素的假设的基础上,而这种假设在实际中往往并不十分准确。

不同于上述方法,本文从说话人模型自适应的角度提出了基于模型自适应的语音转换方法。通过自适应方法训练得到源说话人和目标说话人模型,在一定程度上减少了对说话人训练语句数量的需求。通过自适应后的说话人模型参数,提出了基于模型均值的转换方法,可以有效避免对于特征规整对齐的需求。同时进一步与 INCA 算法相结合,提出了模型自适应 INCA 方法,有效提升了语音转换的效果。

2　传统语音转换方法

图 1 给出了经典语音转换的基本流程。从图中可以看出,经典语音转换通常分为训练和转换两个阶段。在训练阶段,首先,运用语音模型分别对具有相同文本内容的源语音和目标语音进行特征提取; 然后,通过动态时间规整(Dynamic Time Warping,DTW) 等方法对特征参数进行时间对齐; 最后,由对齐后的特征参数通过相应的训练方法得到转换函数。在转换阶段,利用训练得到的转换函数对测试语音的特征参数进行转换,并合成转换语音。

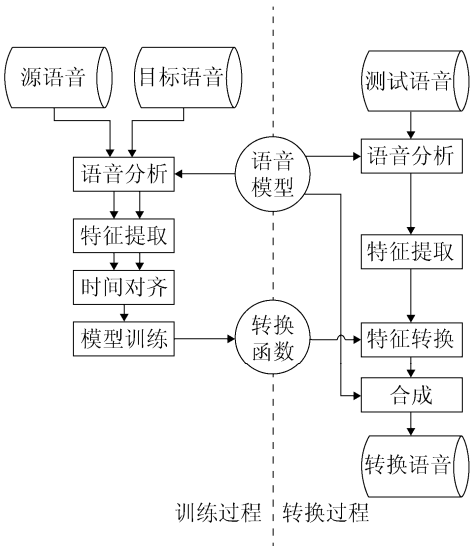

图 1　经典语音转换的基本流程

Fig. 1　Flowchart of classic voice conversion

近年来,GMM 一直是语音转换的研究热点和最常用的模型。它主要包含以下两种训练方法: 最小二乘估计(Least Squares Estimation,LSE) 法 [2] 和联合密度估计(Joint Density Estimation,JDE) 法 [3]。相关实验结果 [3] 表明,这两种方法能够取得相近的实验效果。本文选择基于 JDE 的 GMM 方法作为基线方法,下面简要概述一下其基本原理。

JDE 法通过对源语音和目标语音的频谱特征进行联合建模。设 $X=\{x_1,x_2,\ldots,x_T\}$ 和 $Y=\{y_1,y_2,\ldots,y_T\}$ 分别为通过 DTW 算法对齐后的源语音和目标语音的频谱特征序列,其中 x_t 和 y_t 表示第 t 帧的特征矢量。设 $Z=\{z_1,z_2,\ldots,z_T\}$ 是频谱特征序列对序列,其中 $z_t=[x_t^{\mathrm{T}},y_t^{\mathrm{T}}]^{\mathrm{T}}$(这里 ^{T} 表示转置)。通过 GMM 来对其建模,表示为:

$$p(z)=\sum_{m=1}^{M}\alpha_m N(z,\mu_m,\Sigma_m) \tag{1}$$

其中 $N(z,\mu_m,\Sigma_m)$ 表示高斯概率分布,M 为高斯分量的个数,α_m 表示第 m 个高斯分量的权重,满足 $\sum_{m=1}^{M}\alpha_m=1$,$\mu_m=\begin{bmatrix}\mu_m^x\\ \mu_m^y\end{bmatrix}$ 是均值向量,$\Sigma_m=\begin{bmatrix}\Sigma_m^{xx} & \Sigma_m^{xy}\\ \Sigma_m^{yx} & \Sigma_m^{yy}\end{bmatrix}$ 是

协方差矩阵。根据最小均方误差估计(Minimum Mean-square Error, MMSE)方法,可以得到转换函数:

$$F(\boldsymbol{x}) = E(\boldsymbol{y} \mid \boldsymbol{x}) = \int \boldsymbol{y} p(\boldsymbol{y} \mid \boldsymbol{x})\, dy = \sum_{m=1}^{M} p(m \mid \boldsymbol{x}) \left[\boldsymbol{\mu}_m^{y} + \frac{\Sigma_m^{yx}}{\Sigma_m^{xx}} (\boldsymbol{x} - \boldsymbol{\mu}_m^{x}) \right] \quad (2)$$

其中 $p(m \mid \boldsymbol{x})$ 表示 x 属于第 m 个高斯分量的后验概率,满足:

$$p(m \mid \boldsymbol{x}) = \frac{\alpha_m N(\boldsymbol{x}, \boldsymbol{\mu}_m^{x}, \Sigma_m^{xx})}{\sum_{j=1}^{M} \alpha_j N(\boldsymbol{x}, \boldsymbol{\mu}_j^{x}, \Sigma_j^{xx})} \quad (3)$$

3 基于非对称语音库的语音转换方法

3.1 基于 MAP 方法的模型均值自适应

上述 GMM 方法是基于对称语音库的情况提出的,然而对称语音数据在实际中很难直接获取。针对这种情况,受到语音识别和说话人识别中模型自适应思想的启发,我们通过自适应方法从参考说话人模型中训练得到源说话人和目标说话人的模型。首先通过期望最大化(Expectation Maximization, EM)算法训练得到参考说话人模型 λ_R;然后选择最大后验概率(Maximum A Posteriori, MAP)方法[12]分别训练得到源说话人和目标说话人的模型:λ_S 和 λ_T。给定一组频谱特征观测向量 $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$,设 $P_m(\boldsymbol{o}_t)$ 为第 m 个高斯分量的概率分布,α_R^m 和 $\boldsymbol{\mu}_R^m$ 分别为 λ_R 模型中第 m 个高斯分量的权重和均值向量,计算得到权重和均值的统计量:

$$w_m = \sum_{t=1}^{T} P(m \mid \boldsymbol{o}_t), E_m(\boldsymbol{o}) = \frac{\sum_{t=1}^{T} P(m \mid \boldsymbol{o}_t) \boldsymbol{o}_t}{\sum_{t=1}^{T} P(m \mid \boldsymbol{o}_t)} \quad (4)$$

其中 w_m 和 $E_m(\boldsymbol{o})$ 分别为权重和均值统计量,$P(m \mid \boldsymbol{o}_t)$ 的取值与公式(3)形式相同,满足:$P(m \mid \boldsymbol{o}_t) = \frac{\alpha_R^m P_m(\boldsymbol{o}_t)}{\sum_{j=1}^{M} \alpha_R^j P_j(\boldsymbol{o}_t)}$。则说话人模型均值向量的更新公式如下:

$$\hat{\boldsymbol{\mu}}_R^m = \beta_m E_m(\boldsymbol{o}) + (1 - \beta_m) \boldsymbol{\mu}_R^m \quad (5)$$

其中 β_m 为权重系数,用来对新旧统计量进行平衡。满足 $\beta_m = \frac{w_m}{w_m + \rho}$,其中 ρ 为相关系数,用于描述自适应得到的说话人模型和参考说话人模型的相关度,ρ 通常取值在 8 ~ 20 之间。

3.2 基于模型均值映射的频谱特征转换

由于源说话人和目标说话人的训练语句是非对称的,很难对其通过 DTW 等规整方法进行对齐从而训练得到其转换函数。这里我们研究了自适应说话人模型的均值向量之间的映射关系,并将其用于频谱特征的转换。给定 λ_S 模型的均值向量序列 $\boldsymbol{\mu}_x = \{\boldsymbol{\mu}_x^1, \boldsymbol{\mu}_x^2, \cdots, \boldsymbol{\mu}_x^M\}$,以及对应的 λ_T 模型的均值向量序列 $\boldsymbol{\mu}_y = \{\boldsymbol{\mu}_y^1, \boldsymbol{\mu}_y^2, \cdots, \boldsymbol{\mu}_y^M\}$,模型均值之间的映射函数表示为:

$$F(\boldsymbol{\mu}_x) = \boldsymbol{W}\boldsymbol{\mu}_x + \boldsymbol{b} \quad (6)$$

其中 $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别是转换矩阵和偏差。假定 $\bar{\boldsymbol{\mu}}_x = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\mu}_x^m$, $\bar{\boldsymbol{\mu}}_y = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\mu}_y^m$。运用 LSE 算法分别计算得到 $\boldsymbol{W}$ 和 $\boldsymbol{b}$。形式如下所示:

$$\boldsymbol{W} = \hat{\boldsymbol{\mu}}_y \hat{\boldsymbol{\mu}}_x^{\mathrm{T}} (\hat{\boldsymbol{\mu}}_x \hat{\boldsymbol{\mu}}_x^{\mathrm{T}})^{-1}, \boldsymbol{b} = \bar{\boldsymbol{\mu}}_y - \boldsymbol{W}\bar{\boldsymbol{\mu}}_x \quad (7)$$

其中 $\hat{\boldsymbol{\mu}}_x = \boldsymbol{\mu}_x - \bar{\boldsymbol{\mu}}_x$, $\hat{\boldsymbol{\mu}}_y = \boldsymbol{\mu}_y - \bar{\boldsymbol{\mu}}_y$。

源说话人和目标说话人模型是通过同一参考模型自适应训练得到,因而采用模型均值映射方法可以避免对非对称语句进行对齐的问题。模型均值映射方法在一定程度上反映了源说话人和目标说话人频谱特征之间的映射关系,则频谱特征转换函数可以表达为:

$$F(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \quad (8)$$

3.3 基于模型自适应 INCA 方法的频谱特征转换

尽管公式(8)在某种程度上可以对频谱特征进行转换,但是仅通过均值映射并不能准确的描述不同说话人频谱特征之间的映射关系。我们提出了基于模型自适应 INCA 方法的频谱特征转换方法,不同于 INCA 方法,我们将模型均值方法得到的转换频谱特征作为辅助序列的初值。给定源说话人和目标说话人的非对称频谱特征序列,分别记为 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_J\}$ 和 $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K\}$,则模型自适应 INCA 方法的步骤为:

a) 初始化:定义一个辅助序列 $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_J\}$,不同于文献[11],我们设定初始值 $\hat{\boldsymbol{x}}_j = \boldsymbol{W}\boldsymbol{x}_j + \boldsymbol{b}$。

b) 最近邻对齐:对于 $\hat{\boldsymbol{X}}$ 中的每一帧特征向量 $\hat{\boldsymbol{x}}_j$,找到它在 $\boldsymbol{Y}$ 中对应的最近邻位置 $p(j)$,同样对于 $\boldsymbol{Y}$ 中的每一帧特征向量 $\boldsymbol{y}_k$,

也能找到它在 $\hat{X}$ 中的最近邻位置 $q(k)$。分别表示为:

$$p(j) = \arg\min_k D(\hat{\boldsymbol{x}}_j, \boldsymbol{y}_k),\ q(k) = \arg\min_j D(\boldsymbol{y}_k, \hat{\boldsymbol{x}}_j) \tag{9}$$

其中 $D(\hat{\boldsymbol{x}}_j, \boldsymbol{y}_k)$ 和 $D(\boldsymbol{y}_k, \hat{\boldsymbol{x}}_j)$ 表示频谱特征之间的欧氏距离测度。

c) 转换函数训练: 将对齐后的特征向量如传统 GMM 训练方法类似,训练得到转换函数 $F(\boldsymbol{x})$,形式与式(2)相同。

d) 中间值更新: 通过转换函数更新辅助序列的取值 $\hat{\boldsymbol{x}}_j = F(\boldsymbol{x}_j)$。

e) 收敛性检验: 重复步骤 b) ~ d),直到辅助序列和目标频谱特征序列之间的距离小于阈值 δ。

4　实验结果

4.1　实验环境

我们采用 CMU ARCTIC 语音数据库对本文提出的算法进行评价。选取 BDL(男)和 CLB(女)两位说话人的各 500 条对称语句用于参考模型的训练,并选择说话人 RMS(男)和 SLT(女)各 80 条对称语句用于模型训练和测试。编号记为 1 ~ 80,其中编号 1 ~ 20 的语句用于传统基于对称语音库的 GMM 方法转换函数的训练; 编号 21 ~ 40 和 41 ~ 60 的语句分别用于源说话人和目标说话人模型的训练; 编号 61 ~ 80 的语句用于对结果进行测试。

通过 STRAIGHT 语音模型[13]对特征进行提取,选择 24 阶的美尔倒谱系数(Mel-Cepstrum Coefficients,MCC)用于表示频谱特征,F0 的转换在 log 域采用高斯归一化方法进行转换。自适应相关系数 ρ 取值设为 16,GMM 方法和说话人模型中的高斯分量优化设定为 16。

我们分别进行了男声到女声和女声到男声的语音转换,并对四种不同类型的语音转换方法进行了评价,分别是基于对称语音库的 GMM 方法(GMM)、基于非对称语音库的均值映射方法(MT)、基于非对称语音库的 INCA 方法(INCA)以及基于非对称语音库的模型自适应 INCA 方法(MINCA)。通过 MCD 法和说话人辨识法对转换语音进行了客观评价; 同时通过 ABX 法和平均意见打分(Mean Opinion Score,MOS)法分别对转换语音的目标倾向性和感知度进行了主观测试,其中 6 名语音专业研究人员参与了主观打分。

4.2　客观测试

美尔倒谱距离(Mel Cepstral Distance,MCD)[7]是语音转换领域最为常用的客观评价手段,我们选择其对转换语音的 MCC 失真度进行测试。公式如下:

$$MCD = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{24}(\boldsymbol{c}_d - \boldsymbol{c}_d^t)^2} \tag{10}$$

其中 d 表示 MCC 特征的维度,$\boldsymbol{c}_d$ 和 $\boldsymbol{c}_d^t$ 分别表示转换语音和目标语音的第 d 维 MCC 特征。当 MCD 取值越小时,说明频谱失真度越低,则转换效果越好。

MCD 结果如图 2 所示,我们可以发现: 一、无论是男声到女声还是女声到男声的转换,本文提出的 MINCA 方法都优于传统的 INCA 方法; 二、当训练语音数据较少(如≤4 句)时,本文提出的 MT 方法优于 INCA 方法; 三、随着训练语句数量的增加,本文提出的 MT 方法还是 MINCA 方法都逐渐逼近基于对称语音库的 GMM 方法得到的结果。

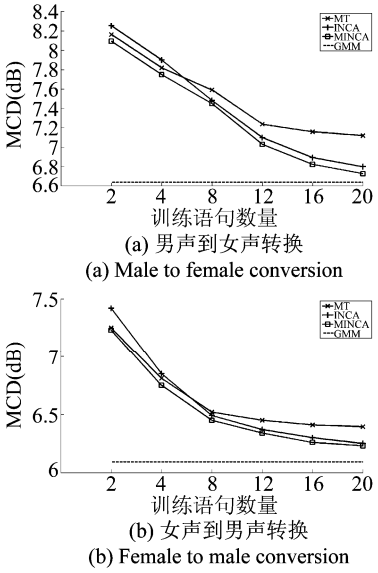

图 2　不同转换方法的 MCD 结果

Fig. 2　MCD results of different conversion methods

我们同样采用说话人辨识法对转换语音的相似度进行了评价。给定源说话人和目标说话人的 GMM 说话人模型 λ_S 和 λ_T,对于测试语句 O,按照下式求取平均相似度得分:

$$\theta_{ST} = \log(\lambda_T \mid O) - \log(\lambda_S \mid O) \quad (11)$$

通过似然度得分来判断转换语音是否被识别为目标说话人。当 $\theta_{ST} > 0$ 时，转换语音被识别为目标说话人；而当 $\theta_{ST} < 0$ 时，转换语音被识别为源说话人。θ_{ST}取值越大则说明转换效果越好。

表 1 给出了说话人辨识法得到的实验结果，实验中我们选择了 20 条语句用于非对称语音库情况下的语音转换函数的训练。从表中我们很容易发现：一、转换前的 θ_{ST}都 <0，说明转换语音都被识别为源说话人，而转换后的 θ_{ST}都 >0，说明转换语音都被识别成目标说话人。二、当训练语句数量相同(20)时，MINCA 方法优于 MT 和 INCA 方法，并且在一定程度上更接近基于对称语音库的 GMM 转换方法得到的结果。

表 1 说话人辨识测试结果

Tab. 1 Results of speaker recognition test

转换类型	转换前 θ_{ST}	转换后 θ_{ST}			
		MT	INCA	MINCA	GMM
男声到女声转换	-5.21	+2.07	+2.16	+2.29	+2.95
女声到男声转换	-5.38	+2.18	+2.25	+2.42	+3.06

4.3 主观测试

为了进一步评价语音质量，我们采用主观评价法对转换语音的感知度以及与目标语音之间的相似度进行评价，我们同样选择 20 条非对称语句用于转换函数训练。首先采用 ABX 测试法对转换语音的倾向度进行测试，这里 A 表示源语音，B 表示目标语音，X 表示转换语音。测听者被要求判断 X 更接近 A 还是 B。表 2 给出了 ABX 测试结果，其中“%”表示 X 被判断为接近 B 的百分比。实验结果如表 2 所示，我们可以发现，相比提出的 MT 方法和经典 INCA 方法，本文提出的 MINCA 方法得到的转换语音更接近于目标语音。这在很大程度上验证了 MCD 和说话人辨识实验等得到的结果。

表 2 ABX 倾向度测试结果(%)

Tab. 2 Results of ABX preference test

转换方法	MT	INCA	MINCA	GMM
倾向度	63	69	72	83

我们接着采用 MOS 打分法对转换语音的感知质量进行评价。选择 5 分制(1～5)的方式对转换语音进行打分，其中 1 分表示很差，5 分表示非常好。表 3 给出了评价得分结果和标准差。可以发现，本文提出的 MINCA 方法相比 MT 方法或 INCA 方法，具有更高的平均得分，同时具有较低的标准差，在很大程度上更接近 GMM 方法的效果。

表 3 MOS 打分结果

Tab. 3 Scores of MOS test

转换方法	MT	INCA	MINCA	GMM
平均得分	2.82	2.95	3.02	3.68
标准差	0.61	0.57	0.55	0.49

5 结论

本文借助说话人自适应的思想，提出了一种有效的基于模型自适应的非对称语音库情况下的语音转换方法。首先通过训练语句自适应训练得到说话人模型；接着利用说话人模型的均值向量，提出了基于均值映射的频谱特征转换方法；同时为了进一步提升转换效果，提出了基于模型自适应 INCA 方法的语音转换。通过主客观实验对本文提出的方法进行了评价，实验结果表明相比传统 INCA 方法，无论是频谱失真度和说话人辨识度、还是转换语音的感知质量和目标倾向度，本文提出的方法都在很大程度上取得了更接近基于对称语音库的 GMM 方法的效果。

参考文献

[1] Abe M, Nakamura S, Shikano K, et al. Voice conversion through vector quantization [C]. In: Acoustics, Speech and Signal Processing (ICASSP), 1988 IEEE International Conference on. 1988: 655-658.

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

[3] Kain A, Macon M W. Spectral voice conversion for text-to-speech synthesis [C]. In: Acoustics, Speech and Signal Processing (ICASSP), 1998 IEEE International Conference on. 1998: 285-288.

[4] Toda T, Saruwatari H, Shikano K. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum [C]. In: A-

coustics, Speech and Signal Processing (ICASSP), 2001 IEEE International Conference on. 2001: 841-844.

[5] Godoy E, Rosec O, Chonavel T. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012: 20(4): 1313-1323.

[6] Qiao Y, Saito D, Minematsu N. HMM-based sequence-to-frame mapping for voice conversion [C]. In: Acoustics, Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. 2010. 4830-4833.

[7] Desai S, Black A, Yegnanarayana B, et al. Spectral mapping using artificial neural networks for voice conversion [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2010, 18(5): 954-964.

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

[9] Popa V, Silen H, Nurminen J, et al. Local linear transformation for voice conversion [C]. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. 2012: 4517-4520.

[10] 徐小峰, 俞一彪. 基于说话人独立建模的语音转换系统研究 [J]. 信号处理, 2009, 25(8A): 171-174
Xu X, Yu Y. The research of voice conversion system based on independent speaker modelling [J]. Signal Processing, 2009, 25(8A): 171-174. (in Chinese)

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

[12] Reynolds D, Quatieri T, Dunn R. Speaker verification using adapted Gaussian mixture models [J]. Digital Signal Processing: A Review Journal, 2000, 10(1): 19-41.

[13] 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 [J]. Speech Communication, 1999, 27(3): 187-207.

作者简介

宋　鹏　男,1983 年出生,东南大学信息科学与工程学院博士研究生,主要研究方向为语音转换及语音合成。
E-mail: pengsongseu@gmail.com

王　浩　男,1985 年出生,东南大学信息科学与工程学院博士研究生,主要研究方向为图像处理、信号处理、低复杂度滤波器设计。
E-mail: klwh2003zhw@qq.com

赵　力　男,1958 年出生,东南大学信息科学与工程学院教授、博士生导师,主要研究方向为语音信号处理。
E-mail: zhaoli@seu.edu.cn