基于状态空间模型的子频带语音转换算法

徐 宁[1],杨 震[1],张玲华[2]
(1. 南京邮电大学信号处理与传输研究院,江苏南京 210003;2. 南京邮电大学通信与信息工程学院,江苏南京 210003)

摘 要: 语音转换是一项改变说话人声音特征的技术,该领域主流方法——基于高斯混合模型的全频带参数映射,会导致转换后的语音频谱产生帧间不连续性.本文针对以上问题提出了改进方案:首先引入状态空间模型来模拟语音动态变化特性,其次利用离散小波变换对语音低频和高频部分的参数分为子频带处理.文章最后用主观和客观实验对提出的算法进行的实验仿真和验证.

关键词: 语音转换;高斯混合模型;状态空间模型;全频带转换;子频带转换
中图分类号: TN925 **文献标识码:** A **文章编号:** 0372-2112 (2010) 03-0646-08

Sub-Band Voice Morphing Algorithm Based on State-Space Model

XU Ning[1], YANG Zhen[1], ZHANG Ling-hua[2]
(1. *Institute of Signal Processing and Transmission of Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*;
2. *College of Telecommunication & Information Engineering of Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*)

Abstract: Voice morphing is a technique to modify a source speaker's speech to sound as if it was spoken by some designated target speaker. The Gaussian mixture model (GMM) based transformations combined with full-band extracted feature parameters have been commonly studied. However, these methods often introduce problems such as artifacts and discontinuities. In order to resolve the problem mentioned above, state-space model (SSM) is first used to describe the relationship between the source speech and the target speech in the spectral domain. Then Discrete Wavelet Transform (DWT) is applied to decompose speech signals into sub-bands in order to improve the quality of the converted speech. Finally, experiments using both objective and subjective measurements are conducted to validate the effectiveness of the proposed method.

Key words: voice morphing; Gaussian mixture model; state-space model; full-band conversion; sub-band conversion.

1 引言

语音转换是一种通过改变源说话人语音的个性特征,同时保留说话者语义信息,并将其转换为具有目标说话人个性特征的语音的技术.在过去的若干年中,一些语音学工作者在这方面做了许多的工作,初步形成了一些体系,其中主要包括:码本矢量量化法(Vector Quantization, VQ)[1,2]、人工神经网络法(Artificial Neural Network, ANN)[3,4]、统计映射法(Statistical Transformation, ST)[5~8]等.码本矢量量化法在语音转换领域的早期被广泛采用.这种算法的特点在于算法复杂度比较低,但由于码书的大小有所限制,即码字的总数是有限的,所以最后转换出来的语音特征参数缺乏多样性,本应该连续变化的语音参数之间常常呈现出离散的过渡特性[9,10];人工神经网络凭借其优异的学习能力,也被引入进语音转换领域.但是 ANN 也存在着一个很大的缺点——"泛化性能"差[11],即它能比较准确地模拟已经被训练过的数据之间的映射关系,但对于从未"见过"的新的测试数据,它常常无能为力;近年来,统计映射算法受到很大的关注.Stylianou 等人将高斯混合模型(Gaussian Mixture Model, GMM)用于特征参数的映射,取得了很大的成功[5].然而基于 GMM 的转换算法也存在着自身的缺点——转换后的特征参数"过于平滑"和相邻语音帧间参数"跳变"[12,13].究其根源,主要有以下两点原因:(1) 理论上,如果要对语音信号进行准确的描述,需要高维的特征参数来表征(同时描述低频信息和高频信息).而高维的参数获取时往往由于训练数据不足带来矩阵奇异、矩阵的逆求解困难以及方程组无法求解等一系列数学问题,因此实际操作时,常常利用低维参数来近似地表征语音信号,即"丢弃"一些细节信息,"保留"轮廓包络信息,而这最终导致了语音转换后特征参数"过于平滑"的现象;(2) GMM 转换算法成立的前提是

收稿日期:2008-09-24;修回日期:2009-01-14
基金项目:国家 863 重点项目(No.2006AA010102);国家自然科学基金(No.60872105,60971129);江苏省普通高校研究生科研创新计划(No.CX08B_079Z)

假设各个语音参数帧之间是统计独立的,因此必然忽视了语音相邻帧之间存在着很强的相关性这样一个事实,从而产生了转换后的语音相邻帧的特征参数出现“跳变”,而不是平滑过渡的现象.现在也有一些学者针对这两方面的问题展开了研究并提出了不同的改进方法:Toda 等人通过引入动态频率规整和计算全局标准差的技术来避免过平滑问题[14].Chung-Hsien 等人提出了采用隐马尔科夫模型(Hidden Markov Model,HMM)来跟踪语音相邻帧之间参数的变化特性[15].

本文针对以上两方面的问题,提出了一套全新的解决方案:一方面利用小波分析呈现多分辨率的特点[16,17],对语音信号进行“分带”处理(即低频部分用较少的系数表示,高频部分用较多的系数表示),用这种既保留了低频包络信息,又不丢失高频细节信息的方法来克服转换后参数“过于平滑”的问题;另一方面,首次提出将状态空间模型(State-Space Model,SSM)[18~20]用于语音转换领域,用来模拟语音帧间参数连续变化的特性,基本解决了参数“跳变”的问题.

2　语音转换系统框架

图 1 是本文 提出的基于状态空间模型的子频带语音转换系统框图.整个系统分为训练阶段和转换阶段两个部分.在训练阶段,首先将语音信号通过 STRAIGHT 模型[21]分析,提取源和目标的语音参数,包括语谱参数和相对应的基频轨迹,然后对语谱参数进行小波分析,并提取低频系数.当源和目标小波系数经过动态时间规整(Dynamic Time Wrapping,DTW)算法[22]对齐后,就可以用于训练 SSM 了.在转换阶段,待转换的源语音同样经过 STRAIGHT 分析和小波分析,并提取小波低频系数用 SSM 进行转换,而同时保留小波高频系数不变,用于和转换后的小波低频系数重构语谱参数.最终修改后的基频轨迹和转换后的语谱参数通过 STRAIGHT 模型合成语音.

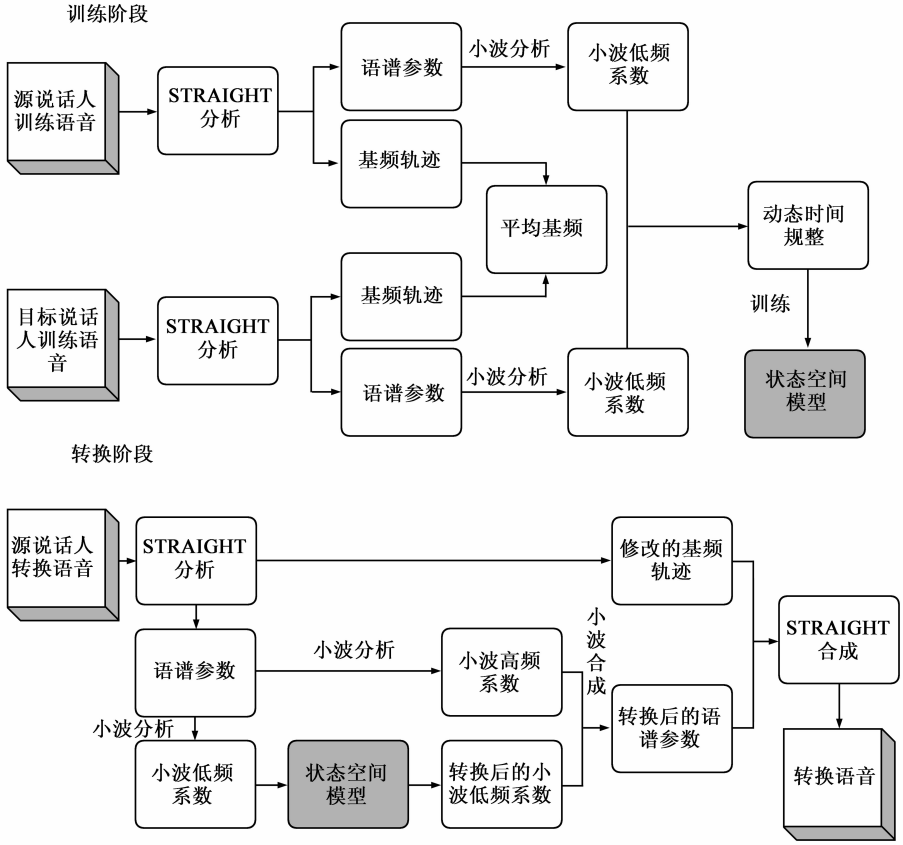


图1　基于SSM的子频带语音转换系统框图

3　语音转换算法

语音转换算法从本质上说就是建立多维矢量之间的映射关系.假设源说话人的训练特征集为 $\boldsymbol{U}=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_T\}$,目标说话人训练特征集为 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N\}$,其中每一个 $\boldsymbol{\mu}_i$ 和 $\boldsymbol{y}_j$ 都是 p 维特征矢量.由于训练模型之前一般要采用 DTW 算法进行时间对齐,所以本文令 $T=N$.

3.1　传统的 GMM 转换算法

GMM 本质上是若干个高斯函数的线性组合,即:

$$p(\boldsymbol{\theta}) = \sum_{q=1}^{Q} \alpha_q N(\boldsymbol{\theta};\boldsymbol{\lambda}_q;\boldsymbol{\Sigma}_q),\ \sum_{q=1}^{Q} \alpha_q = 1, \alpha_q \geqslant 0 \quad (1)$$

其中 $N(\boldsymbol{\theta};\boldsymbol{\lambda}_q;\boldsymbol{\Sigma}_q)$是高斯分布的概率密度函数，$\boldsymbol{\lambda}_q$ 和 $\boldsymbol{\Sigma}_q$ 是该分布的均值和协方差矩阵，α_q 是各个高斯函数的混合权重，Q 是高斯函数的总个数.

用 GMM 设计转换算法时[23]，首先将源和目标对齐的特征矢量组合到一起：$\boldsymbol{z}_k = [\boldsymbol{\mu}_k, \boldsymbol{y}_k]^{\mathrm{T}}$，上标 T 表示矩阵的转置.接着对扩展矢量集 $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_k, \cdots\}$用 GMM 训练，得到 GMM 的参数$(\boldsymbol{\theta};\boldsymbol{\lambda}_q;\boldsymbol{\Sigma}_q)$，其中联合协方差矩阵和均值分别表示为：

$$\boldsymbol{\Sigma}_q = \begin{bmatrix} \boldsymbol{\Sigma}_q^{UU} & \boldsymbol{\Sigma}_q^{YU} \\ \boldsymbol{\Sigma}_q^{UY} & \boldsymbol{\Sigma}_q^{YY} \end{bmatrix}, \boldsymbol{\lambda}_q = \begin{bmatrix} \boldsymbol{\lambda}_q^{U} \\ \boldsymbol{\lambda}_q^{Y} \end{bmatrix}, q = 1,2,\cdots Q \quad (2)$$

最后根据文献[23]，转换函数可表示为：

$$\hat{\boldsymbol{y}}_t = F(\boldsymbol{\mu}_t) = E[\boldsymbol{y}_t | \boldsymbol{\mu}_t]$$

$$= \sum_{q=1}^{Q} p_q(\boldsymbol{\mu}_t)[\boldsymbol{\lambda}_q^{Y} + \boldsymbol{\Sigma}_q^{YU}(\boldsymbol{\Sigma}_q^{UU})^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\lambda}_q^{U})], t = 1,2,\cdots N \quad (3)$$

$$p_q(\boldsymbol{\mu}_t) = \frac{\alpha_q N(\boldsymbol{\mu}_t;\boldsymbol{\lambda}_q^{U};\boldsymbol{\Sigma}_q^{U})}{\sum_{p=1}^{Q} \alpha_p N(\boldsymbol{\mu}_t;\boldsymbol{\lambda}_p^{U};\Sigma_p^{U})} \quad (4)$$

式(3)清楚地说明，传统的 GMM 转换算法是一帧一帧"独立"地转换源特征矢量的，前后帧数据之间没有任何关联，作者认为这也是导致合成语音质量下降的一个主因.

3.2 本文的语音转换算法

3.2.1 基于小波分析的子带参数提取算法

小波分析(Wavelet Analysis)，是一种窗口面积固定但其时间窗和频率窗形状都可改变的时频局域化分析方法，即在低频部分具有较高的频率分辨率和较低的时间分辨率，在高频部分性质恰恰与之相反.正是这种特性，使小波分析对信号具有自适应能力，呈现出所谓的"多分辨率分析"的特点.对于语音信号而言，其低频部分常常被认为包含了语义和说话人个性特征等基本信息，因此在语音信号处理过程中，显得尤为重要.所以，本文提出利用离散小波变换算法来提取语音信号低频部分的参数，同时保留高频部分参数并用于后期转换.

根据离散小波变换(Discrete Wavelet Transform，DWT)理论[16,17]，某离散信号 $s(n)$的 j 层分解可表示为：

$$s(n) = \sum_{k=0}^{\frac{n}{2^j-1}} R_{j,k}(n)\phi_{j,k}(n) + \sum_{t=1}^{j}\sum_{k=0}^{\frac{n}{2^j-1}} S_{t,k}(n)\varphi_{t,k}(n) \quad (5)$$

其中 $\phi(n) = \sum_k h(k)\phi(2n-k)$ 和 $\varphi(n) = \sum_k g(k)\phi(2n-k)$ 分别被称为尺度函数和小波函数，并且它们满足瑞斯条件(Riesz Basis，RB)，即 $R_{j,k}(n) \leqslant \langle s, \phi_{j,k}(n)\rangle$，$S_{t,k}(n) \leqslant \langle s, \varphi_{t,k}(n)\rangle$，也就是说，在满足 RB 时，原离散信号能够被无失真重构.另外，式(5)中的 $h(k)$和 $g(k)$是一对正交镜像滤波器(Quadrature Mirror Filter，QMF)，即 $g(k) = (-1)^k h(N-k-1)$.

基于 DWT 的语音参数提取和合成过程如图 2 所示.在分解信号时，原始信号首先通过一对 QMF 分解为高频信号 H_1 和低频信号 L_1，然后低频信号 L_1 经过"抽取"运算后，又通过另一对 QMF 进一步分解为 H_2 和 L_2，如此迭代下去，直到规定的分解层数；DWT 重构信号的过程恰恰与之相反.

值得注意的是，在原始信号(如一帧语音信号)经过 DWT 分解过后，第 j 层(假设一共 j 层)的低频系数往往是几维至十几维的矢量(矢量的维数取决于原始信号的长度)，因此，用 DWT 分解提取语音特征参数完全符合"维数低"这个要求；同时，DWT 重构信号时，又没有丢弃 $H_1, H_2, \cdots, H_j$ 等高频成分，达到了保留信号"细节"的目的.

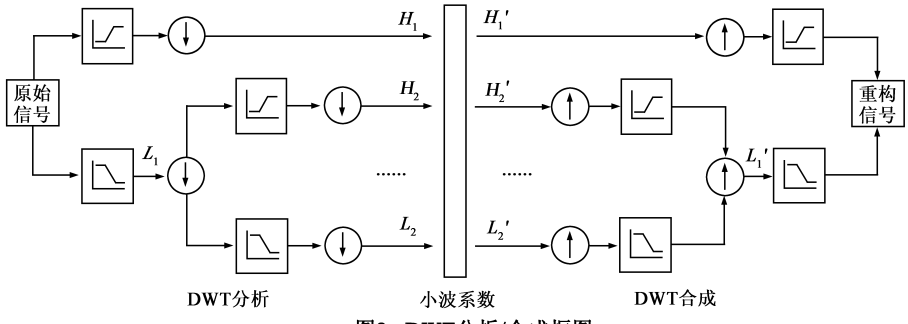


图2 DWT分析/合成框图

3.2.2 基于 SSM 的语音转换核心算法

SSM 假设观测变量(Observed Variable，OV) $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$是由一系列被称之为隐变量(Hidden Variable，HV)的 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_T\}$以某种方式产生，即：

$$\boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t \quad (6)$$

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t \quad (7)$$

其中 $\boldsymbol{y}_t$ 为 p 维矢量，$\boldsymbol{x}_t$ 为 k 维矢量，$\boldsymbol{F}$ 和 $\boldsymbol{H}$ 分别为状态—信号和状态—状态转移矩阵，且 $\boldsymbol{x}_1 \sim N(\boldsymbol{\omega}, \boldsymbol{\Lambda})$，$\boldsymbol{\varepsilon}_t \sim N(\boldsymbol{v}, \boldsymbol{C})$，$\boldsymbol{\eta}_t \sim N(\boldsymbol{w}, \boldsymbol{D})$($\boldsymbol{\omega}, \boldsymbol{v}, \boldsymbol{w}$ 为均值向量，$\boldsymbol{\Lambda}, \boldsymbol{C}, \boldsymbol{D}$ 是协方差矩阵).因此待估计的参数集 $\boldsymbol{\Theta} = \{\boldsymbol{\omega}, \boldsymbol{v}, \boldsymbol{w},$

$\boldsymbol{\Lambda}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{F}, \boldsymbol{H}\}$.

在本文中,我们用期望最大化法(Expectation Maximum, EM)[24]来估算 $\boldsymbol{\Theta}$.为了使推导更加清晰,我们首先考虑一种特殊情况——隐变量 $\boldsymbol{X}$ 已经被观测到,即 $\boldsymbol{X}$ 是普通的随机变量.

由 SSM 的定义式(6)、(7), $\boldsymbol{Y}$ 和 $\boldsymbol{X}$ 的联合概率分布可以被定义为:

$$\begin{aligned} \Gamma(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{X}) &= p(\boldsymbol{Y}, \boldsymbol{X} \mid \boldsymbol{\Theta}) \\ &= p(\boldsymbol{x}_1 \mid \boldsymbol{\Theta}) \prod_{t=2}^{N} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \Theta) \prod_{t=1}^{N} p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \Theta) \end{aligned} \tag{8}$$

其中 N 表示总语音帧数,且 $p(\boldsymbol{x}_1 \mid \boldsymbol{\Theta}) = \frac{1}{\sqrt{(2\pi)^k \mid \boldsymbol{\Lambda} \mid}} \cdot \exp\{-\frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{\omega})^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} (\boldsymbol{x}_1 - \boldsymbol{\omega})\}$.显然, $\boldsymbol{x}_t$ 和 $\boldsymbol{y}_t$ 亦应该服从高斯分布:

$$\boldsymbol{x}_t \sim N(\boldsymbol{F}\overline{\boldsymbol{x}_{t-1}} + \boldsymbol{w}, \boldsymbol{F}\boldsymbol{\Sigma}_{t-1}^{xx}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{D}) \tag{9}$$

$$\boldsymbol{y}_t \sim N(\boldsymbol{H}\overline{\boldsymbol{x}_t} + \boldsymbol{v}, \boldsymbol{H}\boldsymbol{\Sigma}_t^{xx}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{C}) \tag{10}$$

其中 $\overline{\boldsymbol{x}_t}, \boldsymbol{\Sigma}_t^{xx}$ 分别是 t 时刻 $\boldsymbol{x}_t$ 的均值和协方差矩阵.根据式(9)、(10),有以下条件概率成立:

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\Theta}) = \frac{1}{\sqrt{(2\pi)^k \mid \boldsymbol{D} \mid}} \cdot \exp\{-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_{t-1} - \boldsymbol{w})^{\mathrm{T}} \boldsymbol{D}^{-1} (\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_{t-1} - \boldsymbol{w})\} \tag{11}$$

$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\Theta}) = \frac{1}{\sqrt{(2\pi)^p \mid \boldsymbol{C} \mid}} \cdot \exp\{-\frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t - \boldsymbol{v})^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t - \boldsymbol{v})\} \tag{12}$$

将式(11)、(12)代入式(8),并令 $\gamma(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{X}) = \log\Gamma(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{X})$,可得:

$$\begin{aligned} \gamma(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{X}) = & -\frac{1}{2}\sum_{t=1}^{N}\left\{\log \mid \boldsymbol{C} \mid + (\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t - v)^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t - \boldsymbol{v})\right\} \\ & -\frac{1}{2}\sum_{t=2}^{N}\{\log \mid \boldsymbol{D} \mid + (\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_{t-1} - \boldsymbol{w})^{\mathrm{T}} \boldsymbol{D}^{-1} (\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_{t-1} - \boldsymbol{w})\} \\ & -\frac{1}{2}\log \mid \boldsymbol{\Lambda} \mid - \frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{\omega})^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} (\boldsymbol{x}_1 - \boldsymbol{\omega}) - \frac{N(p+k)}{2}\log(2\pi) \end{aligned} \tag{13}$$

而我们的要解决的问题可表述为:

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\Theta} \gamma(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{X}) \tag{14}$$

求解最大似然问题的常用方法是用式(13)对 $\boldsymbol{\Theta}$ 中每个待求参数求偏导数,并令其等于零,于是最终估算的结果为(步骤省略):

$$[\hat{\boldsymbol{H}}, \hat{\boldsymbol{v}}] = \left[\sum_{t=1}^{N} \boldsymbol{y}_t \boldsymbol{x}_t^{\mathrm{T}} \quad \sum_{t=1}^{N} \boldsymbol{y}_t\right] \begin{bmatrix} \sum_{t=1}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^{\mathrm{T}} & \sum_{t=1}^{N} \boldsymbol{x}_t \\ \sum_{t=1}^{N} \boldsymbol{x}_t^{\mathrm{T}} & 1 \end{bmatrix}^{-1} \tag{15}$$

$$\hat{\boldsymbol{C}} = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{y}_t \boldsymbol{y}_t^{\mathrm{T}} - \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{y}_t \boldsymbol{x}_t^{\mathrm{T}} \hat{\boldsymbol{H}}^{\mathrm{T}} - \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{y}_t \boldsymbol{v}^{\mathrm{T}} \tag{16}$$

$$[\hat{\boldsymbol{F}}, \hat{\boldsymbol{w}}] = \left[\sum_{t=2}^{N} \boldsymbol{x}_t \boldsymbol{x}_{t-1}^{\mathrm{T}} \quad \sum_{t=2}^{N} \boldsymbol{x}_t\right] \begin{bmatrix} \sum_{t=2}^{N} \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}^{\mathrm{T}} & \sum_{t=2}^{N} \boldsymbol{x}_{t-1} \\ \sum_{t=2}^{N} \boldsymbol{x}_{t-1}^{\mathrm{T}} & 1 \end{bmatrix}^{-1} \tag{17}$$

$$\hat{\boldsymbol{D}} = \frac{1}{N-1}\sum_{t=2}^{N} \boldsymbol{x}_t \boldsymbol{x}_t^{\mathrm{T}} - \frac{1}{N-1}\sum_{t=2}^{N} \boldsymbol{x}_t \boldsymbol{x}_{t-1}^{\mathrm{T}} \hat{\boldsymbol{F}}^{\mathrm{T}} - \frac{1}{N-1}\sum_{t=2}^{N} \boldsymbol{x}_t \boldsymbol{w}^{\mathrm{T}} \tag{18}$$

$$\hat{\boldsymbol{\omega}} = \boldsymbol{x}_1 \tag{19}$$

$$\hat{\boldsymbol{\Lambda}} = \boldsymbol{x}_1 \boldsymbol{x}_1^{\mathrm{T}} - \boldsymbol{x}_1 \boldsymbol{\omega}^{\mathrm{T}} \tag{20}$$

现在,让我们来考虑完整的 EM 算法.以上估算 $\boldsymbol{\Theta}$ 的结论是在假设 $\boldsymbol{X}$ 为可观测变量的前提下推导出来的,而实际中 $\boldsymbol{X}$ 是隐变量,那么直接求解上述最大似然问题就存在困难.于是 EM 算法采用了先计算充分统计量的数学期望,将其含隐变量的问题转换为普通随机变量的问题,然后再按照常规的最大似然方法求解.针对本文的 SSM 模型而言,这些充分统计量是指 $\boldsymbol{y}_t, \boldsymbol{y}_t\boldsymbol{y}_t^{\mathrm{T}}, \boldsymbol{y}_t\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{x}_t, \boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{x}_t\boldsymbol{x}_{t-1}^{\mathrm{T}}$,根据 SSM 的定义式(6)、(7)和式(9)、(10),上述充分统计量的数学期望可用一种迭代的方式得到,即在第 i 次迭代时有:

$$E[\boldsymbol{x}_t \mid \boldsymbol{\Theta}]^{(i)} = \boldsymbol{F}^{(i-1)} E[\boldsymbol{x}_{t-1} \mid \boldsymbol{\Theta}^{(i)}] + \boldsymbol{w}^{(i-1)} \tag{21}$$

$$\begin{aligned} E[\boldsymbol{y}_t \mid \boldsymbol{\Theta}^{(i)}] &= E[\boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t \mid \boldsymbol{\Theta}^{(i)}] \\ &= \boldsymbol{H}^{(i-1)} E[\boldsymbol{x}_t \mid \boldsymbol{\Theta}^{(i)}] + \boldsymbol{v}^{(i-1)} \end{aligned} \tag{22}$$

$$\begin{aligned} E[\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] &= \boldsymbol{\Sigma}_t^{xx,(i-1)} + E[\boldsymbol{x}_t \mid \boldsymbol{\Theta}^{(i)}] E[\boldsymbol{x}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] \\ &= \boldsymbol{F}^{(i-1)} \boldsymbol{\Sigma}_{t-1}^{xx,(i-1)} \boldsymbol{F}^{(i-1),\mathrm{T}} + \boldsymbol{D}^{(i-1)} \\ &\quad + E[\boldsymbol{x}_t \mid \boldsymbol{\Theta}^{(i)}] E[\boldsymbol{x}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] \end{aligned} \tag{23}$$

$$\begin{aligned} E[\boldsymbol{y}_t\boldsymbol{x}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] &= E[(\boldsymbol{H}\boldsymbol{x}_t + \varepsilon_t)\boldsymbol{x}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] \\ &= \boldsymbol{H}^{(i-1)} E[\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] + \boldsymbol{v}^{(i-1)} E[\boldsymbol{x}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] \end{aligned} \tag{24}$$

$$\begin{aligned} E[\boldsymbol{y}_t\boldsymbol{y}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] &= E[(\boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t)(\boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t)^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] \\ &= \boldsymbol{H}^{(i-1)} E[\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] \boldsymbol{H}^{(i-1),\mathrm{T}} + 2\boldsymbol{H}^{(i-1)} \\ &\quad \cdot \boldsymbol{E}[\boldsymbol{x}_t \mid \boldsymbol{\Theta}^{(i)}] \boldsymbol{v}^{(i-1),\mathrm{T}} + \boldsymbol{C}^{(i-1)} \end{aligned} \tag{25}$$

$$\begin{aligned} E[\boldsymbol{x}_t\boldsymbol{x}_{t-1}^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] &= E[(\boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t)x_{t-1}^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] \\ &= \boldsymbol{F}^{(i-1)} E[\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] + \boldsymbol{w}^{(i-1)} E[\boldsymbol{x}_{t-1}^{\mathrm{T}} \mid \boldsymbol{\Theta}^{(i)}] \end{aligned} \tag{26}$$

其中 $E[\boldsymbol{x}_1 \mid \boldsymbol{\Theta}^{(0)}], \boldsymbol{\Sigma}_1^{xx,(0)}, \boldsymbol{\Theta}^{(0)} = \{\boldsymbol{\omega}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{w}^{(0)}, \boldsymbol{\Lambda}^{(0)}, \boldsymbol{C}^{(0)}, \boldsymbol{D}^{(0)}, \boldsymbol{F}^{(0)}, \boldsymbol{H}^{(0)}\}$ 是初始值.然后用这些数学期望值替代式(15)~(20)中相对应的充分统计量,这样就完成了第 i 步.完整的 EM 算法估计 SSM 模型参数的步骤总结如下:

表 1 本文 SSM 模型参数的估算流程

初始化:给定 $E[\boldsymbol{x}_1 | \boldsymbol{\Theta}^{(0)}]$, $\boldsymbol{\Sigma}_1^{xx,(0)}$, $\boldsymbol{\Theta}^{(0)}$ 的初始值和阈值 ξ,计算似然函数式(8)的值,令其为 $\Gamma^{(0)}$,并令 $i=0$;

第 1 步:$i=i+1$. 对于 $t=1,2,\cdots N$,计算式(21)~(26),然后用这些值带入式(15)~(20)(代替它们相对应的充分统计量),得到 $\Theta^{(i)}$;

第 2 步:用 $\Theta^{(i)}$ 计算式(8),得到 $\Gamma^{(i)}$;

第 3 步:比较 $\Gamma^{(i)}$ 和 $\Gamma^{(i-1)}$ 的值,即计算 $\vartheta = \| \Gamma^{(i)} | - | \Gamma^{(i-1)} \|$,若 $\vartheta \leqslant \xi$,则停止迭代,否则转第 1 步.

由上述的分析过程可以看出,将 SSM 用于语音转换领域,有独特的优点:(1)从局部来看,状态变量之间服从一阶马尔可夫过程,因此特别适合描述语音帧间参数连续变化的动态特性;(2)从整体来看,如果把 $\boldsymbol{X}$ 看作语义信息,将 $\boldsymbol{Y}$ 看作语音信号本身,则式(6)、(7)表达了这样一个事实:语义信息 $\boldsymbol{X}$ 通过与说话人个性相关的某种"操作" $\boldsymbol{H}$ 变换后,呈现出我们熟悉的语音信号 $\boldsymbol{Y}$. 因此,结合 SSM 在语音转换上的物理意义和自身的特点,本文提出了一套区别于以往所有语音转换算法的新的训练/转换算法,其框架如表 2、3 所示:

表 2 训练算法

初始化:源和目标的语音(语义内容一样)经过 DWT 提取参数后,用 DTW 对齐得 $\boldsymbol{Y}_{\text{train}}^{\text{source}}$, $\boldsymbol{Y}_{\text{train}}^{\text{target}}$;

第 1 步:用式(21)~(26)对 $\boldsymbol{Y}_{\text{train}}^{\text{source}}$ 建模,得到模型参数 $\boldsymbol{\Theta}^{\text{source}}$ 和 $\boldsymbol{X}_{\text{train}}^{\text{source}}$;

第 2 步:假设 $\boldsymbol{X}$ 表示语义信息,则 $\boldsymbol{X}_{\text{train}}^{\text{source}} = \boldsymbol{X}_{\text{train}}^{\text{target}}$. 由于此时 $\boldsymbol{Y}_{\text{train}}^{\text{target}}$, $\boldsymbol{X}_{\text{train}}^{\text{target}}$ 均是观测变量,所以对 $\boldsymbol{Y}_{\text{train}}^{\text{target}}$ 用式(15)~(20)求 Θ^{target}.

表 3 转换算法

初始化:待转换的源语音经过 DWT 提取参数得 $\boldsymbol{Y}_{\text{convert}}^{\text{source}}$;

第 1 步:用训练阶段得到的 $\boldsymbol{\Theta}^{\text{source}}$ 和 $\boldsymbol{Y}_{\text{convert}}^{\text{source}}$ 反估计得 $\boldsymbol{X}_{\text{convert}}^{\text{source}}$;

第 2 步:根据公式(6),结合 $\boldsymbol{\Theta}^{\text{target}}$ 和 $\boldsymbol{X}_{\text{convert}}^{\text{source}}$,可得最终的转换结果 $\boldsymbol{Y}_{\text{convert}}^{\text{target}}$,即对于每一个 $\boldsymbol{y}_t \in \boldsymbol{Y}_{\text{convert}}^{\text{target}}$,有 $\boldsymbol{y}_t = H^{\text{target}} \boldsymbol{x}_t^{\text{source}} + \boldsymbol{\varepsilon}_t^{\text{target}}$.

4 仿真实验

在本节中,将用主观和客观两套方法对传统的 GMM 转换结果和本文提出的基于 SMM 的子频带转换结果进行对比评测. 实验结果表明:本文提出的创新算法不仅极大地提高了转换后说话人个性与目标说话人个性的相似度,而且使得转换后语音音质相对传统算法有了质的飞跃。

4.1 仿真环境

本文所用的语音库是在我校语音消声室中录制的纯净语音,其中包括两男两女以标准普通话口音朗读的各 146 组音节词和 531 条句子. 这些语音均采用 16kHz 采样率采样,并以 16bit 的精度存储于计算机内以供使用. 本文采用 STRAIGHT 模型作为语音分解/合成模型,其帧长为 20ms,帧偏移 5ms,同时采用 DWT 算法提取特征参数(在本文中,采用了 5 阶 DWT 分解,取小波二叉树中最左下角的系数作为待转换的低频系数,由于选取小波的不同而系数维数略有不同,本文实验中范围为:16~22 维),在将参数经过 DTW 算法对齐之后,最终利用 SSM 模型进行语音特征参数的映射转换.

值得一提的是,STRAIGHT 模型分解语音信号后,产生两组不同类型的参数:基频参数和语音声道谱参数. 众多研究文献表明[25~27]:语音声道谱参数包含了绝大多数的说话人个性特征,而基频参数只是起到一个辅助的作用,因此本文 SSM 模型映射的参数主要是指语音声道谱参数. 但考虑到基频参数对最终转换结果亦有一定的影响,因此本文采用了文献[28]的算法,对基频参数做简单的处理,即对源和目标语音的基频进行如下转换:

$$f^{\text{target}}(t) = af^{\text{source}}(t) + b \tag{27}$$

其中 $a = \sqrt{\dfrac{\sigma_t^2}{\sigma_s^2}}$, $b = \mu_t - a\mu_s$, $f^{\text{target}}(t)$ 表示转换后的基频参数轨迹, $f^{\text{source}}(t)$ 表示源基频参数轨迹, σ_t 和 σ_s 分别表示目标和源说话人基频参数的方差,同理, μ_t 和 μ_s 分别表示目标和源的基频参数的均值.

4.2 客观评测

本文采用式(28)来评价转换系统的谱参数失真值:

$$\varepsilon = 10\log \frac{\dfrac{1}{N}\sum_{t=1}^{N} \| \boldsymbol{y}_t^{\text{target}} - \Upsilon(\boldsymbol{y}_t^{\text{source}}) \|^2}{\dfrac{1}{N}\sum_{t=1}^{N} \| \boldsymbol{y}_t^{\text{target}} \|^2} \tag{28}$$

其中 $\boldsymbol{y}_t^{\text{source}}$ 表示待转换的源语音帧, $\Upsilon(\cdot)$ 表示基于 SSM 模型的转换算法, $\boldsymbol{y}_t^{\text{target}}$ 表示转换后的语音帧, N 表示总的语音帧数.

图 3 显示了 GMM 模型、SSM 模型以及采用了子频带转换方案的 SSM 模型的转换结果:首先,基于 SSM 模型的转换结果明显好于基于 GMM 模型的转换结果,在训练参 数相同的情况下,平均谱失真约小 2dB;其次,用小波变换提取子频带参数,并用 SSM 转换后的结果又要比用传统的全频带参数(例如用 LSF 参数)的结果要好,同样,在训练参数相同的情况下,平均谱失真约小 2dB;最后,图 3 直观地告诉我们,三种转换算法的转换效果的好坏,都与训练数据的多少有着直接的关系,即随着训练数据的增多而谱失真减小,同时,对于基于 SSM 模型的转换算法而言,在训练数据从 500 帧达到 2000 帧时,效果有一个很大的飞跃,而后随着数据量的增多,性能明显减缓. 从原理上来分析,SSM 模型优于

GMM 模型的关键在于:SSM 模型能够描述相邻语音帧间的强相关性,即帧间变化的动态特性,而 GMM 模型则一开始就假设语音帧间相互独立.同时,子频带 SSM 又因为充分利用了小波变换"多分辨率"的特性——对信号的轮廓部分加以处理,对细节部分又不予以丢弃,所以转换得到的效果必然要优于利用全带 LSF 参数的 SSM 模型.

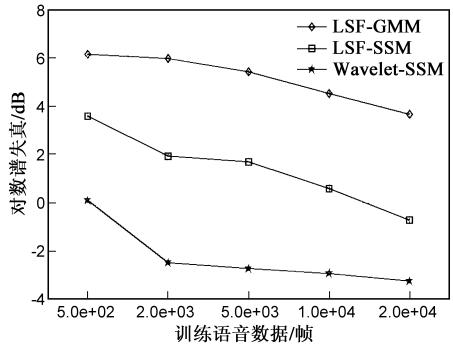


图3　三种转换算法的比较

图 4 是从语音波形和语谱的方面来研究比较以上三种转换算法的效果.从图中可以直观地看到:一、(*b*)图的语音波形(即 GMM 模型转换效果)带有很大的噪声,(*c*)图和(*d*)图中噪声均有减弱的趋势;二、(*f*)图和(*g*)图在语音的高频部分信息丢失的较多,模模糊糊的一片,缺乏细节,而(*h*)图高频部分的细节则相当丰富.这两点现象证明了本文提出的基于子频带参数转换的 SSM 模型的性能的确优于传统的语音转换算法.

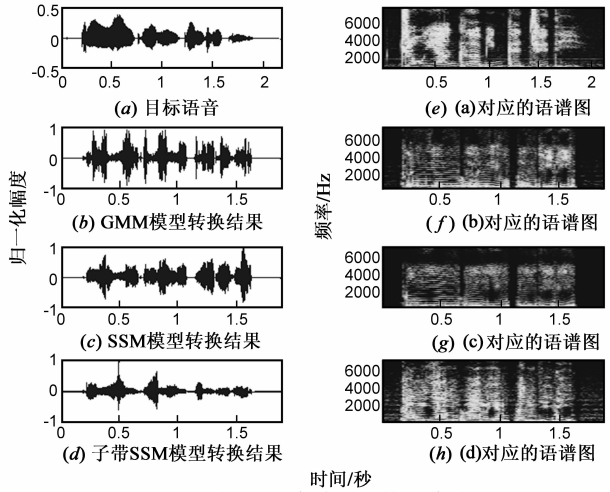


图4　语音波形及其语谱图

图 5 分析了在子频带 SSM 转换算法中,小波类型的变化对转换效果的影响.本文分别采取了 Haar 小波、Daubechies 小波、Biorthogonal 小波、Coiflets 小波和 Symelets 小波这五种小波进行实验,其结果表明:采用不同的小波得到的谱失真值还是有比较明显的不同的.其中 Haar 小波表现得最好,并且随着训练数据量的变化,性能浮动不大.由此可以得出这样的结论:针对不同的说话人,需要选择最优的小波以使转换效果达到最佳.

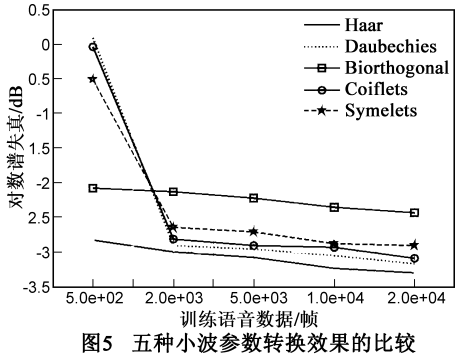


图5　五种小波参数转换效果的比较

4.3　主观评测

主观评测就是让评测人直接听语音,进而对转换语音效果进行评定等级的测试方法.主观评测要达到两个目的:一、评测转换后语音中包含的说话人个性特征;二、评测转换后语音的音质.在语音转换领域中,对于前者,常常采用一种叫做"ABX"测试的方法来实现.表 4 给出了"ABX"测试的结果(一共 6 位评测人,对 50 个词和 30 句话进行评测,表中统计了他们的平均分).其中的百分比反映了转换后的语音"X"与相关类别的相似度情况,由表 4 可以明确地看到 GMM、SSM、子频带 SSM 在转换说话人个性特征的环节上,性能呈现递增的状况.在同样训练数据(20000 帧)的情况下,GMM 算法刚刚达到 46% 与目标人相似,而 SSM 和子带 SSM 则分别达到 70% 和 82%.可见,"刻画帧间动态特性"和"基于多分辨率分析的子频带转换"两种技术极大地提高了传统语音转换算法的精确度.

主观评测的第二项任务常常用评定主观意见分(Mean Opinion Score, MOS)来完成.与 ABX 测试一样,需要若干测试者来听转换后的语音,并评定它们的音质情况.在本文的实验中,为了更准确而客观地评定音质,采用了 ITU-T P.563[29] 标准来代替人工评测.ITU-T P.563 能对一句独立的输入语音进行分析并评定它的音质情况,因此使实验测试结果既准确又高效.该环节

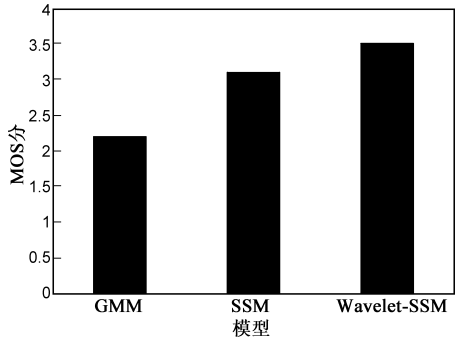


图6　主观MOS分评价

仍然对 50 个词和 30 句话进行实验测试,并统计平均分,如图 6 所示.从图中可以看出,本文提出的转换算法,在语音的音质上比传统的转换算法有了很大的提高,MOS 分从 2.2 分别提高至 3.1 和 3.5.该项实验结果说明:注重研究语音参数的细节特征和动态特性,能有效地提高语音转换的最终效果.

表 4 ABX 测试

训练数据(单位:帧)	GMM			SSM			Wavelet-SSM		
	源	目标	都不像	源	目标	都不像	源	目标	都不像
500	31.5%	35.4%	33.1%	25.1%	51.2%	23.7%	18.6%	66.3%	15.1%
1500	32.7%	38.4%	28.9%	24.6%	57.9%	17.5%	13.5%	75.5%	11.0%
2500	29.7%	40.1%	30.2%	21.5%	62.9%	15.6%	11.9%	75.3%	12.8%
5000	27.0%	43.6%	29.4%	19.8%	65.4%	14.8%	10.1%	79.1%	10.8%
15000	27.3%	45.1%	27.6%	17.3%	68.0%	14.7%	9.5%	81.4%	9.1%
20000	26.9%	46.5%	26.6%	16.7%	70.8%	12.5%	9.5%	82.2%	8.3%

5 总结

传统语音转换算法中存在两大问题:(1)没有考虑语音帧间过渡的动态变化特性;(2)为了便于数学求解,采用较低维数的特征参数,使得高频细节信息在转换过程中严重丢失.本文研究设计了两种方案来解决以上问题,即首先引入 SSM 模型来模拟语音信号帧间参数随着时间缓慢变化的动态特性,然后又利用离散小波变换的多分辨率分析的特点,对语音信号分"子频带"进行处理——提取参数的低频部分用于训练转换函数,并保留参数的高频部分用于最后合成语音.实验证明,本文提出的算法,无论是在区别说话人个性特征方面,还是在主观语音音质方面,都比传统的基于 GMM 模型、采用全频带的低维特征参数的语音转换算法有了质的突破.

参考文献:

[1] ABE M, NAKAMURA S, SHIKANO K, KUWABARA H. Voice conversion through vector quantization[A]. Proceedings of International Conference on Acoustics, Speech, and Signal Processing[C]. New York: IEEE Press, 1988. 655 – 658.

[2] SHIKANO K, NAKAMURA S, ABE M. Speaker adaptation and voice conversion by codebook mapping[A]. Proceedings of IEEE International Symposium on Circuits and Systems [C]. New York: IEEE Press, 1991. 594 – 597.

[3] GUOYU Zuo, WENJU Liu, XIAOGANG Ruan. Genetic algorithm based RBF neural network for voice conversion[A]. Proceedings of IEEE World Congress on Intelligent Control and Automation[C]. New York: IEEE Press, 2004. 4215 – 4218.

[4] IWAHASHI N, SAGISAKA Y. 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.

[5] STYLIANOU Y, CAPPE O. Continuous probabilistic transform for voice conversion[J]. IEEE Transactions on Speech and Audio Processing, 1998, 6(2): 131 – 142.

[6] KAIN A. High Resolution Voice Transformation[D]. Portland: Oregon Health and Sci Univ, 2001.

[7] HUI Ye, STEVE Young. Perceptually weighted linear transformations for voice conversion[J]. Eurospeech, 2003, 8(2): 2409 – 2412.

[8] KUN Liu. High quality voice conversion through combining modified GMM and formant mapping for Mandarin[A]. Proceedings of International Conference on Digital Telecommunications[C]. New York: IEEE Press, 2007. 1038 – 1042.

[9] HIDEYUKI M, MASANOBU A. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt[J]. Speech Communication, 1995, 16(2): 153 – 164.

[10] KNAGENHJELM H, KLEIJN W. Spectral dynamics is more important than spectral distortion[A]. Proceedings of International Conference on Acoustics, Speech, and Signal Processing [C]. New York: IEEE Press, 1995. 732 – 735.

[11] BISHOP C. Neural Networks for Pattern Recognition[M]. UK: Oxford Univ Press Inc, 1995.

[12] YINING Chen. Voice conversion with smoothed GMM and MAP adaptation [A]. Eurospeech [C]. England: Elsevier Press, 2003. 2413 – 2416.

[13] KAIN A, MACON M. Spectral voice conversion for text-to-speech synthesis[A]. Proceedings of International Conference on Acoustics, Speech, and Signal Processing[C]. New York: IEEE Press, 1998. 285 – 288.

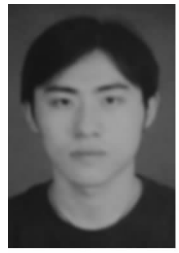
[14] TODA T, BLACK A, TOKUDA K. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory[J]. IEEE Transactions on Audio, Speech and language Processing, 2007, 15(8): 2222 – 2235.

[15] CHUNGHSIEN W. Voice conversion using duration-embedded Bi-HMMs for expressive speech synthesis [J]. IEEE Transactions on Audio, Speech, and Language Processing. 2006, 14(4): 211 – 219.

[16] ADDISON P. The Illustrated Wavelet Transform Handbook:

Introductory Theory and Applications in Science[M]. Edinburgh: Institute of Physics Publishing, 2002.
[17] STRANG G, NGUYEN T. Wavelets and Filter Banks[M]. Wellesley, USA: Wellesley-Cambridge Press, 1997.
[18] ZHENG LI, STEPHEN M. Using a state space model with hidden variables to infer transcription factor activities[J]. Bioinformatics, 2006, 22(6): 747 – 754.
[19] FRANKLIN G, POWELL J, WORKMAN M. Digital Control of Dynamic Systems[M]. NJ, USA: Englewood Cliffs, 1998.
[20] TANIZAKI H. Nonlinear Filters: Estimation and Applications[M]. Germany: Springer Verlag, 1996.
[21] KAWAHARA H, MASUDA-KATSUSE I, DE CHEVEIGNé A. Restructuring speech representations using a pitch adaptive time-frequency-based F0 extraction: Possible role of a repetitive structure in sounds[J]. Speech Communication, 1999, 27(3): 187 – 207.
[22] XUEDONG Huang, ACERO A, HSIAO-WUEN Hong. Spoken Language Processing: A Guide to Theory, Algorithm and System Development[M]. NJ, USA, Prentice Hall PTR, 2001.
[23] HUI Ye, STEVE Young. Quality-enhanced voice morphing using maximum likelihood transformations[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 518 – 530.
[24] DEMPSTER A, LAIRD N, RUBIN D. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, Series B (Methodological), 1977, 39(14): 1 – 38.
[25] KUWABARA H, SAGISAKA Y. Acoustic characteristics of speaker individuality: control and conversion[J]. Speech Communication, 1995, 16(2): 165 – 173.
[26] O TüRK. New Methods for Voice Conversion[D]. Istanbul, Turkey: Boazici University, 2003.
[27] KI SEUNG Lee. Statistical approach for voice personality transformation[J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(2): 641 – 651.
[28] LEVENT ARSLAN M. Speaker transformation algorithm using segmental codebooks (STASC)[J]. Speech Communication, 1999, 2(8): 211 – 226.
[29] Single-ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications[S]. ITU-T Rec. P. 563, 2004.

作者简介:



徐　宁　男,1981 年生,江苏省常州人,南京邮电大学通信与信息工程学院 07 级博士研究生,主要研究方向为语音转换技术和语音识别技术.
E-mail: xuningdlts@gmail.com

杨　震　男,1961 年生,江苏省武进人,南京邮电大学校长、教授、博士生导师.长期从事信号与信息处理、通信理论与技术的教学科研工作.主持和参加完成了国家科技支撑计划、国家“863”、自然科学基金、省部级、校级和合作科研项目近 20 项,在国内外学术刊物和会议上发表学术论文 170 多篇.
Email: yangz@njupt.edu.cn

张玲华　女,1967 年生,南京邮电大学电子信息科学系主任、教授、硕士生导师.曾获厅级科技进步一等奖 1 项,获得江苏省“青蓝工程”优秀青年骨干教师等荣誉称号.曾参与并很好完成了国家自然科学基金(69601001)、863 高科技发展计划(2001AA143070)及多项省部级科研项目.
E-mail: zlh@njupt.edu.cn