基于双因子高斯过程动态模型的声道谱转换方法

孙新建[1] 张雄伟[2] 杨吉斌[2] 曹铁勇[2] 钟新毅[1]

摘 要 针对作者已经提出的双因子高斯过程隐变量模型 (Two-factor Gaussian process latent variable model, TF-GPLVM) 用于语音转换时未考虑语音的动态特征, 并且模型训练时需要估计的参数较多的问题, 提出引入隐马尔科夫模型 (Hidden Markov model, HMM) 对语音动态特征进行建模, 并利用 HMM 隐状态对各帧语音进行关于语义内容的概率软分类, 建立了分离精度更高、运算负荷较小的双因子高斯过程动态模型 (Two-factor Gaussian process dynamic model, TF-GPDM). 基于此模型, 设计了一种全新的基于说话人特征替换的语音声道谱转换方案. 主、客观实验结果表明, 无论是与传统的统计映射和频率弯折转换方法相比, 还是与双因子高斯过程隐变量模型方法相比, 本文方法都获得了语音质量和转换相似度的提升, 以及两项性能的更佳平衡.

关键词 声道谱转换, 高斯过程隐变量模型, 双因子模型, 隐马尔科夫模型, 语音动态特征

引用格式 孙新建, 张雄伟, 杨吉斌, 曹铁勇, 钟新毅. 基于双因子高斯过程动态模型的声道谱转换方法. 自动化学报, 2014, **40**(6): 1198−1207

DOI 10.3724/SP.J.1004.2014.01198

Vocal Tract Spectrum Conversion Using a Two-factor Gaussian Process Dynamic Model

SUN Xin-Jian[1] ZHANG Xiong-Wei[2] YANG Ji-Bin[2] CAO Tie-Yong[2] ZHONG Xin-Yi[1]

Abstract We developed in a previous work a two-factor Gaussian process latent variable model (TF-GPLVM) to perform spectral conversion using a strategy of speaker characteristics replacement. Despite its improved performance compared with traditional mapping-based methods, the model suffers from two drawbacks: 1) it cannot capture the speech dynamical characteristics, and 2) there is a large number of parameters to estimate. To overcome these two drawbacks, we propose in this paper to combine TF-GPLVM with hidden Markov model (HMM), and develop an enhanced two-factor Gaussian process dynamic model (TF-GPDM). In the model, the speech dynamics are modeled by state transition probability of HMM, meanwhile speech frames are categorized into a limited number of phonetic content classes using HMM states. Both subjective and objective evaluations show that, compared with both traditional mapping-based methods, such as Gaussian mixture model (GMM) and FW, and TF-GPLVM based one, the proposed TF-GPDM not only improves the speech quality and identity similarity, but also reaches a better compromise between the two dimensions.

Key words Vocal tract spectrum conversion, Gaussian process latent variable model (GPLVM), two-factor model, hidden Markov model (HMM), speech dynamical characteristics

Citation Sun Xin-Jian, Zhang Xiong-Wei, Yang Ji-Bin, Cao Tie-Yong, Zhong Xin-Yi. Vocal tract spectrum conversion using a two-factor Gaussian process dynamic model. *Acta Automatica Sinica*, 2014, **40**(6): 1198−1207

语音转换 (Voice conversion, VC) 是在保持语义内容信息不变的前提下, 通过改变语音中源说话人声音特征, 使之具有目标说话人声音特征的处理技术[1]. 语音转换在个性化人机交互、信息安全、多媒体娱乐等领域有广泛的应用前景. 例如, 将语音转换和文语转换 (Text-to-speech, TTS) 技术相结合, 可以实现个性化的语音生成系统; 通过语音转换来模仿特定人的声音, 可以实现语音伪装和突破说话人识别准入系统; 通过语音转换还可以再现历史人物演讲和电影配音等.

由于语音的声道谱充分体现了说话人声音特征[2], 因此现有语音转换研究主要集中在对声道谱的建模和转换上. 目前, 国内外学者对声道谱转换研究普遍采用参数映射. 其基本思路是在统一的语音分解–合成模型下, 首先提取语音声道谱特征参数,

收稿日期 2012-12-12 录用日期 2013-05-21
Manuscript received December 12, 2012; accepted May 21, 2013
国家自然科学基金 (61072042), 江苏省自然科学基金 (BK2012510), 解放军理工大学预先研究基金 (20110205, 20110211) 资助
Supported by National Natural Science Foundation of China (61072042), Natural Science Foundation of Jiangsu Province (BK2012510), and Pre-research Foundation of PLA University of Science and Technology (20110205, 20110211)
本文责任编委 宗成庆
Recommended by Associate Editor ZONG Cheng-Qing
1. 解放军理工大学通信工程学院 南京 210007 2. 解放军理工大学指挥信息系统学院 南京 210007
1. College of Communication Engineering, PLA University of Science and Technology, Nanjing 210007 2. College of Command Information Systems, PLA University of Science and Technology, Nanjing 210007

然后通过训练寻找源说话人和目标说话人声道谱之间的映射关系, 最后利用这种映射关系实现输入语音到目标语音的转换. 典型方法包括: 码书映射 (Codebook mapping)[3−4]、人工神经网络 (Artificial neural network, ANN)[5−7]、概率统计映射 (Probabilistic statistical mapping)[8−14]、频率弯折 (Frequency warping, FW)[15−18]. 其中, 基于共振峰的分段线性频率弯折[17] 和基于联合高斯混合模型 (Gaussian mixture model, GMM) 的统计映射[9] 被公认为性能最好而被普遍采用. FW 通过改变频率尺度对不同说话人的声道谱进行映射匹配, 这种方法对频谱结构的破坏小, 转换后的语音自然度保持好, 但是由于频谱幅度调整受限, 因此转换相似度不高. GMM 方法利用软分类实现说话人声道谱之间的连续映射, 在转换相似度上取得较好性能, 但由于过平滑问题, 转换后声道谱细节信息丢失严重, 导致语音质量下降明显. 通过以上分析, FW 和 GMM 在声道谱转换性能上各有利弊, 语音质量和转换相似度难以兼顾[17, 19]. 针对这一问题, 国内外学者也提出了许多改进方法. 例如, Toda 等把 FW 变换后的声道谱加入经 GMM 和 FW 变换后的声道谱差值, 一方面克服 GMM 的过平滑问题, 另一方面调整 FW 的谱能量, 使得转换后语音更贴近目标说话人[20]; Erro 等对经过 GMM 软分类的声道谱空间实行 FW 变换, 并将变换后的频谱进行能量滤波, 这样在基本保持和 GMM 方法一样的转换相似度前提下, 语音质量获得明显提升[19]. 以上融合方法虽然一定程度上提高了语音转换的综合性能, 但没有根本解决由于映射带来的语音质量和转换相似度之间的矛盾.

最近关于语音感知的研究表明, 人类的听觉感知系统对语音中的说话人特征和语义内容实际是分离表示和处理的[21−23]. 依据此理论, 有学者抛开传统的映射思路, 从说话人特征和语义内容分离角度研究声道谱转换. 2009 年, Popa 等利用双线性模型 (Bilinear model, BLM) 把声道谱分解成表示个人特征的形式参数和承载语义信息的内容参数, 并用外推法 (Extrapolation) 实现了一种语音转换方案, 实验结果表明这种方法与 GMM 方法相比具有更好的鲁棒性[24]. 同年, 徐宁等利用状态空间模型 (State-space model, SSM) 对语音声道谱进行建模, 把谱参数分解为状态序列和测量矩阵, 并认为状态序列承载了不同说话人发相同音时声道谱的共性信息, 而测量矩阵表征了说话人之间的差异性信息, 基于此思路提出了一种基于矩阵替换的声道谱转换方法, 这种方法在语音质量和转换相似度上都比 GMM 方法有所提升[25].

然而, BLM 和 SSM 在描述说话人特征和语义内容之间的相互作用时都利用线性关系, 模型建立相对简单. 为提高建模精度, 作者利用包含两个隐变量的高斯过程隐变量模型, 即双因子高斯过程隐变量模型 (Two-factor Gaussian process latent variable model, TF-GPLVM) 对语音的说话人特征和语义内容进行建模, 利用高斯过程核函数描述两者之间复杂的相互作用, 并利用说话人特征替换进行语音转换实验, 取得了较好的转换效果[26]. 但是, TF-GPLVM 模型未考虑语音前后帧的时间相关性, 并且模型训练时需要估计的参数较多, 针对这两个问题, 本文提出通过引入 HMM 对语音的动态特征进行建模, 并利用 HMM 软分类方法减少模型的语义内容参数数量, 提出了一种模型精度更高、计算负荷更小的用于说话人特征和语义内容建模的双因子高斯过程动态模型 (Two-factor Gaussian process dynamic model, TF-GPDM). 在模型训练基础上, 本文利用说话人特征替换实现了一种全新的声道谱转换方法. 主、客观实验结果表明: 无论与传统的 GMM 和 FW 映射方法相比, 还是与 TF-GPLVM 语音转换方法相比, 本文提出的 TF-GPDM 语音转换方法都获得了更高的语音质量和转换相似度, 以及两种转换指标的更佳平衡, 语音转换综合性能明显提升.

1 语音的双因子高斯过程隐变量模型

高斯过程隐变量模型 (Gaussian process latent variable model, GPLVM) 是近年来新兴的一种数据降维工具, 能有效地过滤数据间的误差和个体差异. 若 $Y = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t, \cdots, \boldsymbol{y}_T]$ 表示观测到的数据, $\boldsymbol{y}_t \in \mathbf{R}^D$, $X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$ 表示 Y 对应的隐含变量, $\boldsymbol{x}_t \in \mathbf{R}^q$, 且 $q \ll D$, 则隐变量空间到观测数据空间的映射可以表示为

$$\boldsymbol{y}_t = g(\boldsymbol{x}_t; W) + \boldsymbol{\eta}_t = \sum_i \boldsymbol{w}_i \phi_i(\boldsymbol{x}_t) + \boldsymbol{\eta}_t \tag{1}$$

其中, $W = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots]$ 是函数 g 的参数矩阵, $\boldsymbol{w}_i \in \mathbf{R}^D$; ϕ_i 是一组基函数, 用来描述非线性运算; $\boldsymbol{\eta}_t \in \mathbf{R}^D$ 代表误差, 且 $\boldsymbol{\eta}_t \sim \mathrm{N}(0, \beta^{-1} I)$, β^{-1} 代表误差方差. 对于 GPLVM 而言, 从所有的隐变量到观测数据的每一维建立映射时都设有一个高斯过程先验, 即 W 的行向量服从高斯分布: $W_{d,:} \sim \mathrm{N}(0, I)$[27].

为了适应对语音中说话人特征和语义内容建模需要, 我们在 GPLVM 中再引入一个隐变量对数据间的个体差异进行建模, 得到了 TF-GPLVM 模型[26], 其数学描述如下:

$$\boldsymbol{y}_t = \sum_j \psi_j(\boldsymbol{a})\, g_j(\boldsymbol{x}_t; W_j) + \boldsymbol{\eta}_t =$$

$$\sum_{i,j}\psi_j(\boldsymbol{a})\boldsymbol{w}_{ij}\phi_i(\boldsymbol{x}_t)+\boldsymbol{\eta}_t \tag{2}$$

式 (2) 中, $\boldsymbol{y}_t$ 表示第 t 帧语音观测数据, 本文采用 14 维的线谱频率参数 (Line spectral frequencies, LSF); $\boldsymbol{a}\in\mathbf{R}^p$ 代表说话人特征; $\boldsymbol{x}_t\in\mathbf{R}^q$ 代表第 t 帧的语义内容; ψ_j 是一组对应于说话人特征的基函数; 其他符号定义与式 (1) 相同, 另外假设 W_j 之间相互独立. 实际上, 式 (2) 可以理解为语义内容在不同说话人特征条件下到语音观测数据的映射, 它描述了说话人特征与语义内容之间的相互作用. 注意到, 如果式 (2) 中的基函数 ϕ_i 和 ψ_j 都为线性, 则式 (2) 变为在概率意义下的双线性模型, 因此 TF-GPLVM 模型可以认为是双线性模型在概率意义下的非线性拓展.

如果一个语音集包含 S 名说话人, 且每名说话人的发音内容相同, 时间对准后语音观测数据为 $Y=[Y^1\ \cdots\ Y^s\ \cdots\ Y^S]$, 其中 $Y^s=[\boldsymbol{y}_1^s\ \cdots\ \boldsymbol{y}_t^s\ \cdots\ \boldsymbol{y}_T^s]$. 利用 TF-GPLVM 对 Y 中包含的说话人特征和语义内容建模时, 一个合理假设是同一说话人各帧语音的说话人特征参数固定, 而不同说话人时间对准后的各帧语音具有相同的语义内容. 因此, Y 中的说话人特征参数共有 S 个, 构成特征集 $\mathcal{A}=\{\boldsymbol{a}_1\ \cdots\ \boldsymbol{a}_S\}$, 语义内容参数共有 T 个, 构成内容集 $\mathcal{X}=\{\boldsymbol{x}_1\ \cdots\ \boldsymbol{x}_T\}$, Y 包含的语音数据总数为 $S\times T=N$. 根据式 (2) 建立的双因子高斯过程映射, Y 的似然函数表示为

$$p(Y|\mathcal{A},\mathcal{X},\Omega,\Theta)=\frac{|\Omega|^N}{(2\pi)^{\frac{DN}{2}}|K_Y|^{\frac{D}{2}}}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega^2 Y{K_Y}^{-1}Y^{\mathrm{T}}\right)\right] \tag{3}$$

式 (3) 中, K_Y 为 $N\times N$ 的核矩阵, 矩阵中的元素由式 (4) 表示的核函数计算, 尺度参数 $\Omega=\mathrm{diag}\{\omega_1,\omega_2,\cdots,\omega_D\}$ 用来描述语音数据的各维方差, Θ 是与核矩阵 K_Y 相关的参数集.

$$k_Y(\boldsymbol{a},\boldsymbol{a}';\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{\psi}(\boldsymbol{a})^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{a}')\boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}')+\beta^{-1}\delta_{\boldsymbol{a},\boldsymbol{a}';\boldsymbol{x},\boldsymbol{x}'} \tag{4}$$

式 (4) 中, $\boldsymbol{\psi}(\boldsymbol{a})=[\psi_1(\boldsymbol{a}),\psi_2(\boldsymbol{a}),\cdots]^{\mathrm{T}}$, $\boldsymbol{\phi}(\boldsymbol{x})=[\phi_1(\boldsymbol{x}),\phi_2(\boldsymbol{x}),\cdots]^{\mathrm{T}}$; $\boldsymbol{a}$ 和 $\boldsymbol{a}'$ 来自说话人特征集 $\mathcal{A}$, $\boldsymbol{x}$ 和 $\boldsymbol{x}'$ 来自语义内容集 $\mathcal{X}$; $\delta_{\boldsymbol{a},\boldsymbol{a}';\boldsymbol{x},\boldsymbol{x}'}$ 是 Kronecker delta 函数, 当 $\boldsymbol{a}=\boldsymbol{a}'$ 且 $\boldsymbol{x}=\boldsymbol{x}'$ 时取值为 1, 其他为 0. 由式 (4) 可以看出, 核函数 k_Y 的选择与基函数 $\boldsymbol{\psi}$ 和 $\boldsymbol{\phi}$ 的形式密切相关, 因此我们可以灵活选择不同核函数以适应实际建模需求.

2 语音的双因子高斯过程动态模型

2.1 TF-GPLVM 存在的问题

利用 TF-GPLVM 对语音的说话人特征和语义内容进行建模时, 存在一个明显缺陷是处理的语音数据在时间上是独立的, 忽略了语音前后帧之间的时间相关信息, 而这种信息对于处理语音这样的时序信号来讲非常重要[28]. 图 1 将一段目标语音和 TF-GPLVM 方法转换后语音的第 4 阶 Mel 倒谱系数的时间变化轨迹描绘出来. 从图中可以发现, 转换后语音特征参数与自然语音相比, 两者的变化趋势基本一致, 但是从短时细节来看, 目标语音表现的比较平滑自然, 而 TF-GPLVM 转换语音抖动较多, 另外两个轨迹的局部特征也表现出较大的差异性. 引起这种差异的主要原因就是 TF-GPLVM 进行语音转换时, 没有考虑语音前后帧的时间相关性, 参数轨迹在较小的时间范围内波动较大. 实际上, 科研人员在研究基于映射的语音转换时就已经发现考虑语音动态特征对提升语音转换效果, 尤其是语音质量, 作用非常明显, 先后提出了利用语音前后帧差分特征[10] 和 HMM 模型[29] 等多种适用于语音转换的语音动态特征建模方法. 另外, TF-GPLVM 模型建立时有一个前提假设是所有说话人时间对准后不同时间点的语音帧被 "硬" 归于不同的语义内容, 这样模型中的语义内容参数数量与每名说话人的语音帧数相同, 因此模型训练时需要估计的参数数量过多, 导致算法效率很低.

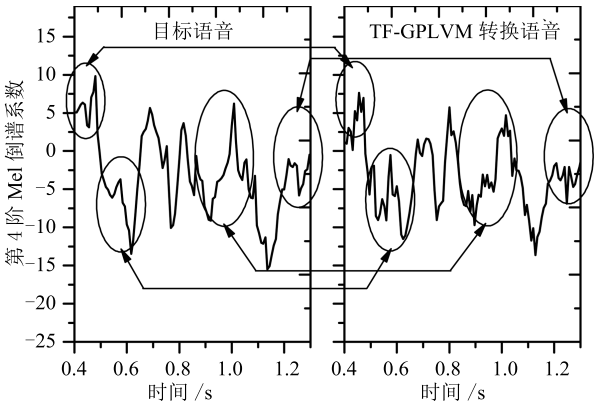

图 1 目标语音和 TF-GPLVM 转换语音参数轨迹对比

Fig. 1 An example of TF-GPLVM converted and natural target parameter trajectories

2.2 TF-GPDM 模型描述

为解决上节提出的 TF-GPLVM 存在的问题, 并且考虑到 HMM 模型是利用隐状态序列对语音进行建模, 而 TF-GPLVM 模型也在隐空间描述说话人特征和语义内容, 两者在处理语音时具有一定

的相似性. 因此, 本文将在利用 TF-GPLVM 进行说话人特征和语义内容建模时引入 HMM 来描述语音的动态特征, 并认为 HMM 的每个隐状态对应一类语义内容, 进而利用每帧语音所处隐状态的后验概率对语音进行概率软分类, 以此建立了用于说话人特征和语义内容建模的双因子高斯过程动态模型 (TF-GPDM), 其数学描述如下式所示:

$$\boldsymbol{y}_t = \sum_{q=1}^{Q} \gamma_q(t) \sum_{i,j} \psi_j(\boldsymbol{a}) \boldsymbol{w}_{ij} \phi_i(\boldsymbol{x}_q) + \boldsymbol{\eta}_t \tag{5}$$

其中, Q 表示 HMM 隐状态数量, $\gamma_q(t)$ 表示第 t 帧语音属于第 q 类语音的分类系数, 它在数值上等于第 t 帧语音处于状态 λ_q 时的后验概率, 由式 (6) 计算[30], 且 $\sum_{q=1}^{Q} \gamma_q(t) = 1$, $\boldsymbol{x}_q$ 代表第 q 类语音的语义内容参数, 其他符号的定义与式 (2) 一致.

$$\gamma_q(t) = p(\lambda_q(t)|Y^s) = \frac{\alpha_q(t)\beta_q(t)}{\sum_{i=1}^{Q} \alpha_i(t)\beta_i(t)} \tag{6}$$

式 (6) 中, $\alpha_q(t)$ 表示第 t 帧语音处于状态 λ_q 条件下 HMM 产生部分观察序列 $\{\boldsymbol{y}^s{}_1, \boldsymbol{y}^s{}_2, \cdots, \boldsymbol{y}^s{}_t\}$ 的概率, 称为前向概率; $\beta_q(t)$ 表示第 t 帧语音处于状态 λ_q 条件下 HMM 产生部分观察序列 $\{\boldsymbol{y}^s{}_t, \boldsymbol{y}^s{}_{t+1}, \cdots, \boldsymbol{y}^s{}_T\}$ 的概率, 称为后向概率. 两种概率的计算可参见文献 [30].

为更直观表达利用 TF-GPLVM 和 TF-GPDM 两种模型对语音说话人特征和语义内容建模的区别, 我们将它们的实现框图在图 2 中表示出来. 其中, 图 2 (a) 是 TF-GPLVM, 图 2 (b) 是 TF-GPDM. 通过对两幅图的比较可以发现, 两种模型最大的区别是 TF-GPDM 通过 HMM 考虑了语音各帧的时序信息, 而 TF-GPLVM 对数据的处理是相对独立的. 另外, 两种模型包含的语义内容参数的数量也不相同, TF-GPLVM 的语义内容数量与语音序列包含的语音帧数目相同, 而 TF-GPDM 的语义内容数量与 HMM 的状态数一致.

2.3 TF-GPDM 模型训练

TF-GPDM 模型训练用的语音数据与第 1 节相同, 都为 Y. 通过第 2.2 节分析, 利用 TF-GPDM 对语音说话人特征和语义内容建模时, 语音的动态特征和分类信息都通过 $\gamma_q(t)$ 体现出来, 因此需要首先建立语音的 HMM 模型, 然后利用式 (6) 求出 $\gamma_q(t)$. 为确保相同时刻不同说话人的语音具有相同的语义内容, 我们把经过时间对准后的不同说话人的 LSF 参数合并成一个向量作为训练样本, 即: $\boldsymbol{y}_t = [(\boldsymbol{y}_t^1)^{\mathrm{T}} \cdots (\boldsymbol{y}_t^s)^{\mathrm{T}} \cdots (\boldsymbol{y}_t^S)^{\mathrm{T}}]^{\mathrm{T}}$, 进行联合隐马尔科夫模型 (Joint-HMM) 训练[29].

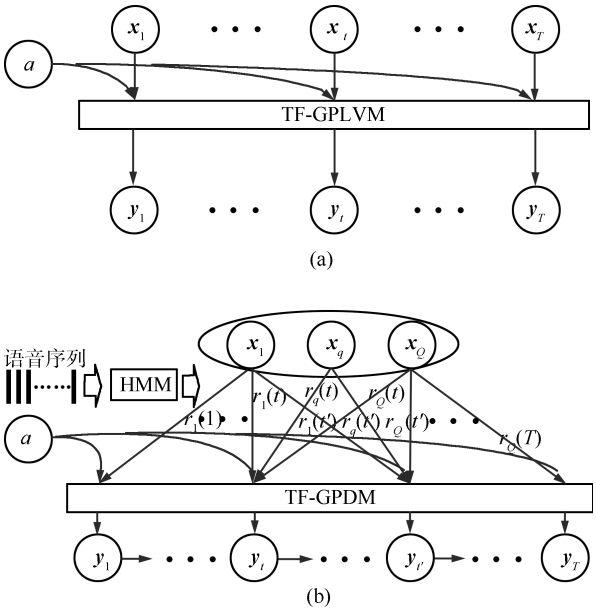

图 2　语音的双因子高斯过程隐变量模型和双因子高斯过程动态模型实现框图

Fig. 2　Speech sequence graphical models of TF-GPLVM and TF-GPDM

Joint-HMM 模型训练完后, TF-GPDM 利用 HMM 隐状态把语音帧进行关于语义内容的软分类, 并假设相同类的语音对应同一个语义内容参数. 如果设定 HMM 隐状态数为 Q, 则 Y 中的语义内容集变为 $\mathcal{X} = \{\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_Q\}$. 这时 Y 的似然概率仍用式 (3) 表示, 但其中核矩阵 K_Y 中的元素变由式 (7) 计算 (推导过程见附录 A):

$$k(\boldsymbol{a}, \boldsymbol{a}'; t, t') = \boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{a})\boldsymbol{\psi}(\boldsymbol{a}') \left[\boldsymbol{\gamma}^{\mathrm{T}}(t)\Sigma\boldsymbol{\gamma}(t')\right] + \beta^{-1}\delta_{\boldsymbol{a},\boldsymbol{a}';t,t'} \tag{7}$$

式 (7) 中, t 和 t' 代表语音序列的不同时刻, 且

$$\boldsymbol{\gamma}(t) = [\gamma_1(t), \cdots, \gamma_Q(t)]^{\mathrm{T}} \tag{8}$$

$$\Sigma = \begin{bmatrix} \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_Q) \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1) & \cdots & \boldsymbol{\phi}(\boldsymbol{x}_Q) \end{bmatrix} \tag{9}$$

训练 TF-GPDM 需要确立的参数包括说话人特征参数 $\mathcal{A} = \{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_S\}$、语义内容参数 $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_Q\}$、尺度参数 Ω 以及核矩阵中的参数 Θ. 采用的方法是最大似然条件下的尺度共轭梯度 (Scale conjugate gradient, SCG) 数值优化算法[31].

3 基于 TF-GPDM 的语音声道谱转换方案

3.1 转换方案描述

在 TF-GPDM 模型建立基础上, 我们采用目标说话人特征替换方式实现了一种全新的语音声道谱转换方案, 实现框图如图 3 所示. 具体的方法流程描述如下:

$$p\left(Y^{\mathrm{Out}}|\boldsymbol{a}_{\mathrm{Tgt}}\mathcal{X},Y\right)=\frac{|\Omega|^{T^*}}{(2\pi)^{\frac{DT^*}{2}}|K_{Y^{\mathrm{Out}}}|^{\frac{D}{2}}}\times \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Omega^2\left(Y^{\mathrm{Out}}-YK_Y{}^{-1}U\right)\times K_{Y^{\mathrm{Out}}}^{-1}\left(Y^{\mathrm{Out}}-YK_Y{}^{-1}U\right)^{\mathrm{T}}\right]\right\} \tag{10}$$

步骤 1. 录制包括源说话人和目标说话人在内的共 S 名说话人平行语音库, 在语音分析和时间对准基础上, 提取每帧语音声道谱的 LSF 参数, 建立模型训练用的语音样本 Y, 假设时间对齐后每名说话人包含的语音帧数为 T, 则训练数据总数为 $S\times T=N$.

步骤 2. 利用 Y 进行联合 HMM 训练, 得到多名说话人的联合 HMM 模型.

步骤 3. 利用联合 HMM 对训练语音进行软分类, 在此基础上利用第 2.3 节方法训练 Y 的 TF-GPDM 模型, 得到目标说话人特征参数 $\boldsymbol{a}_{\mathrm{Tgt}}$ 和语义内容参数 $\mathcal{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_Q\}$, 以及模型参数 Θ. 这里需要说明, TF-GPLVM 模型由于把各语音帧"硬" 归于不同的语义内容来处理, 因此需要估计的语义内容参数和语音帧数相同, 都为 T, 而 TF-GPDM 需要估计的语义内容参数为 Q, 一般情况下 $Q\ll T$, 因此, 与 TF-GPLVM 相比, TF-GPDM 需要估计的参数数量大幅减少, 计算效率明显提升.

步骤 4. 利用联合 HMM 得到源说话人的 HMM 模型[30], 再利用此模型对输入语音 $Y^{\mathrm{In}}=[\boldsymbol{y}_1^{\mathrm{Src}},\cdots,\boldsymbol{y}_\tau^{\mathrm{Src}},\cdots,\boldsymbol{y}_{T^*}^{\mathrm{Src}}]$ 进行软分类.

步骤 5. 在步骤 4 对输入语音进行软分类的基础上, 利用 $\boldsymbol{a}_{\mathrm{Tgt}}$ 和 $\mathcal{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_Q\}$ 重构出转换语音的声道谱参数序列 $Y^{\mathrm{Out}}=[\boldsymbol{y}_1^{\mathrm{Tgt}},\cdots,\boldsymbol{y}_\tau^{\mathrm{Tgt}},\cdots,\boldsymbol{y}_{T^*}^{\mathrm{Tgt}}]$.

3.2 声道谱重构

本节讨论在图 3 转换阶段如何利用说话人特征和语义内容重构出语音声道谱参数. 在 TF-GPDM 已经建立, 并且目标说话人特征 $\boldsymbol{a}_{\mathrm{Tgt}}$ 和分类后的语义内容 $\mathcal{X}$ 已经确定的情况下, 转换后语音声道谱序列 Y^{Out} 的后验概率仍是高斯的, 其密度函数如式 (10) 所示. 因此, Y^{Out} 的估值可以用其均值表示, 即

$$\hat{Y}^{\mathrm{Out}}=\mathrm{E}(Y^{\mathrm{Out}})=YK_Y{}^{-1}U \tag{11}$$

其中, U 是一个大小为 $N\times T^*$ 的矩阵, 矩阵中的元素利用下式求出:

$$k\left(\boldsymbol{a}_{\mathrm{Tgt}},\boldsymbol{a}_s;\tau,t\right)=\boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{a}_{\mathrm{Tgt}})\boldsymbol{\psi}\left(\boldsymbol{a}_s\right)\left[\boldsymbol{\gamma}^{*\mathrm{T}}(\tau)\Sigma\boldsymbol{\gamma}\left(t\right)\right] \tag{12}$$

其中, $\boldsymbol{a}_s\in\mathcal{A}$, $\boldsymbol{\gamma}^*\left(\tau\right)=\left[\gamma_1^*\left(\tau\right),\cdots,\gamma_Q^*\left(\tau\right)\right]^{\mathrm{T}}$ 是输入语音声道谱序列 Y^{In} 在 τ 时刻的软分类系数向量.

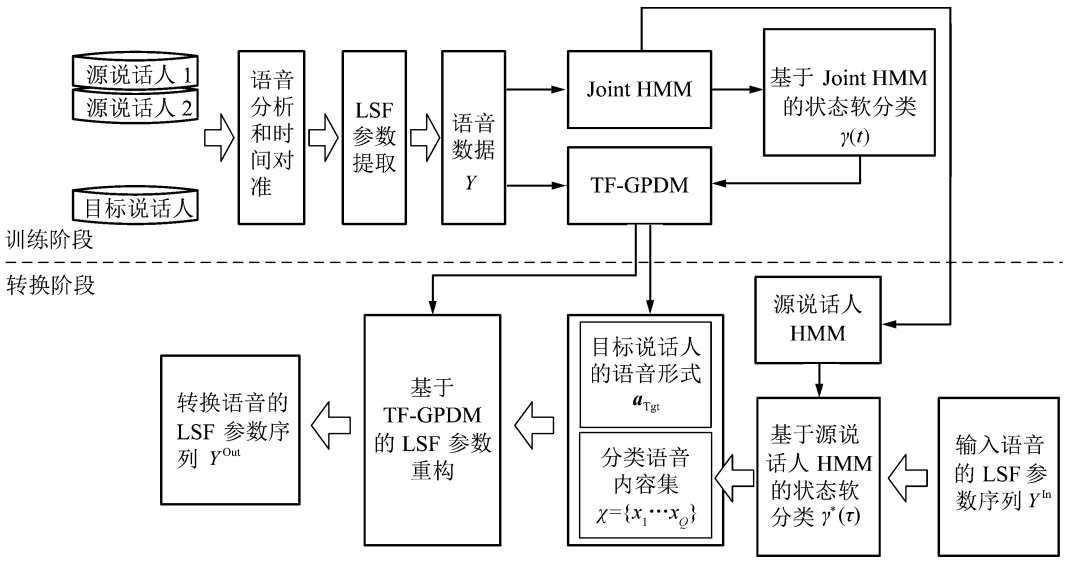

图 3 基于 TF-GPDM 模型的声道谱转换实现框图

Fig. 3 Block diagram of vocal tract spectral conversion using a TF-GPDM

4　实验结果及分析

为了评价本文提出的基于 TF-GPDM 的语音声道谱转换方法, 我们对其性能分别进行了主客观测试, 并与 TF-GPLVM、GMM、FW 等方法进行了比较.

4.1　实验设置

实验语音选自 CMU 的 ARCTIC 语音库[32], 包含 3 名说话人 (bdl、clb 和 slt), 每名说话人共 120 句话, 且语句内容相同. 实验时, 设定 bdl 为源说话人, slt 为目标说话人, 每名说话人的 70 句语音用于训练, 剩下 50 句用于测试. 所有语音利用谐波随机噪声模型 (Harmonic stochastic model, HSM)[33] 进行分析, 每帧语音利用 14 阶的 LSF 参数进行表示, 其中浊音帧的 LSF 参数由谐波谱转换得到, 清音帧的 LSF 参数由随机成分 LPC 谱转换得到. 进行时间对准时, 统一以 bdl 的语音为基准.

利用 TF-GPDM 对语音建模时, 说话人特征和语义内容参数统一设置为 3 维向量, HMM 模型状态数设置为 32, 每个状态的发射概率是单高斯的, 核函数中说话人特征部分采用线性, 而语义内容部分采用径向基的形式, 即

$$\boldsymbol{\psi}(\boldsymbol{a})^{\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{a}'\right)=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}' \tag{13}$$

$$\boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}\left(\boldsymbol{x}'\right)=\exp\left(-\frac{\alpha}{2}\|\boldsymbol{x}-\boldsymbol{x}'\|^{2}\right) \tag{14}$$

其中, α 是径向基宽度参数. 基于 GMM 的映射转换采用文献 [9] 提出的方案, 稍作改变的是把 3 名说话人的 LSF 参数进行联合用于 GMM 训练, 高斯成分设置为 32 个 (GMM-32). 基于 FW 的映射转换采用文献 [17] 提出的共振峰映射方案, 用一个线性分段频率弯折函数对每帧输入语音的频谱进行映射.

另外, 进行主观评价时, 需要合成出转换后的语音波形, 以上所有语音转换方法只针对浊音帧进行转换, 对清音帧不作转换处理. 浊音帧中的随机成分统一采用谐波谱和随机成分 LPC 谱的联合 GMM 概率密度方法进行预测[19]. 假设 GMM 共有 m 个高斯成分, 且第 i 个成分的分布为

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{e}\end{bmatrix}\sim\mathrm{N}\left(\begin{bmatrix}\boldsymbol{\mu}_i^{\boldsymbol{y}}\\ \boldsymbol{\mu}_i^{\boldsymbol{e}}\end{bmatrix},\begin{bmatrix}\Sigma_i^{\boldsymbol{yy}} & \Sigma_i^{\boldsymbol{ye}}\\ \Sigma_i^{\boldsymbol{ey}} & \Sigma_i^{\boldsymbol{ee}}\end{bmatrix}\right) \tag{15}$$

则随机成分的预测函数如下式所示:

$$\hat{\boldsymbol{e}}=\sum_{i=1}^{m}p_i\left(\boldsymbol{y}\right)\cdot\left[\boldsymbol{\mu}_i^{\boldsymbol{e}}+\Sigma_i^{\boldsymbol{ey}}(\Sigma_i^{\boldsymbol{yy}})^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}_i^{\boldsymbol{y}}\right)\right] \tag{16}$$

其中, $p_i\left(\boldsymbol{y}\right)$ 表示谐波谱在第 i 个高斯成分条件下的后验概率. 对于韵律变换统一采用基频的线性对数映射实现, 如下式所示:

$$\log f_0{}'=\mu_{\log f_0}^{\mathrm{Tgt}}+\frac{\sigma_{\log f_0}^{\mathrm{Tgt}}}{\sigma_{\log f_0}^{\mathrm{Src}}}\left(\log f_0-\mu_{\log f_0}^{\mathrm{Src}}\right) \tag{17}$$

其中, $f_0{}'$ 和 f_0 分别代表转换后的基频和源说话人基频, $\mu_{\log f_0}^{\mathrm{Tgt}}$ 和 $\mu_{\log f_0}^{\mathrm{Src}}$ 分别代表目标说话人和源说话人基频对数的均值, $\sigma_{\log f_0}^{\mathrm{Tgt}}$ 和 $\sigma_{\log f_0}^{\mathrm{Src}}$ 分别代表目标说话人和源说话人基频对数的方差.

4.2　客观实验结果及分析

本实验利用两种指标对声道谱转换结果进行客观评价. 第一种是估计转换后的声道谱和目标声道谱之间均方差 (Mean squared error, MSE), 它在一定程度上反映了声道谱的整体转换精度, 用式 (18) 进行计算:

$$\mathrm{MSE}\left(\hat{\boldsymbol{s}},\boldsymbol{s}\right)=\frac{\dfrac{1}{T}\sum\limits_{t=1}^{T}\left\|\hat{\boldsymbol{s}}\left(t\right)-\boldsymbol{s}\left(t\right)\right\|^{2}}{\dfrac{1}{T}\sum\limits_{t=1}^{T}\left\|\boldsymbol{s}\left(t\right)\right\|^{2}} \tag{18}$$

其中, $\hat{\boldsymbol{s}}$ 和 $\boldsymbol{s}$ 分别表示转换后的声道谱和目标声道谱参数向量, T 表示语音帧数, 本实验只考虑浊音帧. 第二种是计算转换后语音数据的方差和目标语音数据的方差之间的平均比率 (Variance ration, VR), 它能反映转换后语音声道谱动态变化和细节特征情况, 用式 (19) 计算[18].

$$\mathrm{VR}=\frac{1}{D}\sum_{d=1}^{D}\left(\frac{\sigma_{\hat{\boldsymbol{s}}}(d)}{\sigma_{\boldsymbol{s}}(d)}\right)^{2} \tag{19}$$

其中, $\sigma_{\hat{\boldsymbol{s}}}\left(d\right)$ 和 $\sigma_{\boldsymbol{s}}\left(d\right)$ 分别表示转换后和目标声道谱向量的第 d 维数据的标准偏差.

图 4 反映的是四种声道谱转换方法的 MSE 结果. 从全局看, 首先, 四种转换方法的性能与训练数据的多少有直接关系, 其 MSE 都随着训练语句的增加而减小. 其次, 基于说话人特征和语义内容建模的转换方法提供的 MSE 曲线的变化趋势比映射方法平缓, 这说明前者的转换性能对训练数据的大小相对不敏感; 最后, 基于说话人特征和语义内容建模的方法提供的 MSE 结果整体上优于映射方法, 这说明前者对声道谱转换的精度得到提升. 从局部看, 当训练语句数较少时, TF-GPLVM 和 TF-GPDM 的性能都不高, 前者略优于后者, 这主要因为在 TF-GPDM 建模中, 需要训练 HMM 模型, 较少的训练语句会造成 HMM 模型性能变差, 进而影响 TF-G-PDM 的建模精度; 随着训练语句的增多, TF-GPDM 性能迅速提升, 并超过 TF-GPLVM, 主要原因是前者利用 HMM 考虑了语音的动态特征,

而 HMM 也随着训练语句增多性能得到改善, 进而提高了对语音说话人特征和语义内容的建模精度.

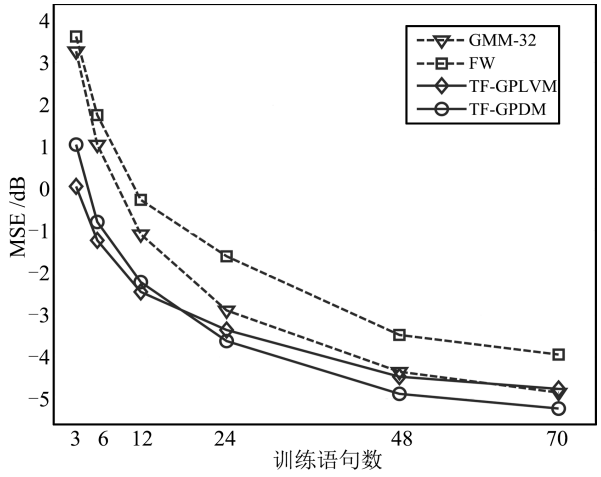

图 4 四种声道谱转换方法的 MSE 结果

Fig. 4 Results of MSE on four different conversion methods

图 5 给出了四种声道谱转换方法以及自然语音的 VR 结果, 训练语句为 70 句. VR 反映的是语音声道谱数据的动态变化情况, 相比于 MSE, 它从自然度上反映语音的转换效果[18]. 理想情况下的 VR 值为 1, 说明转换后的声道谱数据没有受到损坏, 可以得到完全自然的转换语音. VR 值越小说明语音数据的方差越小, 语音越不自然. 从比较结果可以看出: GMM 方法表现最差, 说明转换后语音不自然, 主要原因是概率统计模型带来的过平滑问题; FW 方法 VR 性能较好, 因为其只改变了共振峰的位置, 而保持了声道谱的幅度; TF-GPLVM 和 TF-GPDM 方法也具有较好保持语音自然度的能力; 进一步, TF-GPLVM 方法的 VR 值略高于 TF-GPDM, 可能因为后者利用 HMM 对语音动态特征建模的同时, 由于软分类原因, 计算式 (11) 中的 U 矩阵时也会带来一定的过平滑问题.

4.3 主观实验结果及分析

主观评价包括语音质量和说话人相似度两种听力测试, 共 24 人参加, 训练语句 70 句. 首先, 利用 MOS (Mean opinion score) 评分方法对转换后的语音质量进行评测, 实验结果如表 1 第 2 栏所示. 然后, 仍然利用 MOS 评分规则对给出的 “转换 – 目标” 语句的说话人特征进行相似度打分[19], 实验结果如表 1 第 3 栏所示. 表 1 同时给出了源语音和目标语音的主观评价结果. 从表 1 可以看出, FW 在语音质量这一指标上表现最好, GMM 方法最差, 本文提出的 TF-GPDM 方法性能接近 FW. 而在转换相似度指标上, FW 方法明显劣于其他方法, GMM 和 TF-GPDM 两种方法表现优异. 进一步可以看出, TF-GPDM 在语音质量和转换相似度两项指标上都超过 TF-GPLVM, 再次验证了对语音说话人特征和语义内容建模时需要考虑帧间动态特征的必要性.

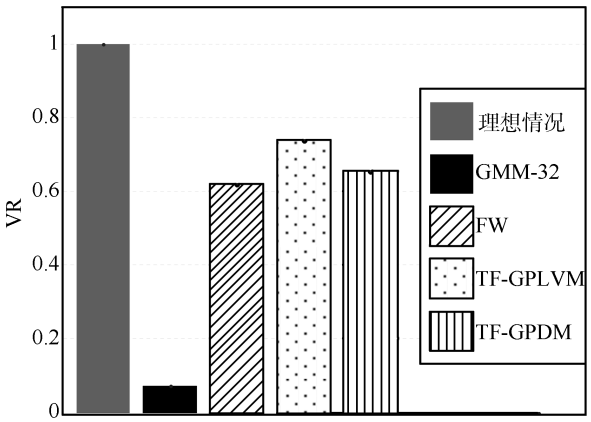

图 5 四种声道谱转换方法以及理想情况的 VR 结果

Fig. 5 Comparison results of VR on four different conversion methods and real speech

表 1 不同转换方法的主观评价结果 (95 % 置信区间)

Table 1 Subjective evaluation results on different conversion methods (at 95 % of confidence intervals)

	语音质量	转换相似度
源语音	4.37 ± 0.04	1.67 ± 0.03
目标语音	4.29 ± 0.03	4.87 ± 0.02
GMM-32	2.27 ± 0.04	3.43 ± 0.03
FW	3.64 ± 0.02	1.87 ± 0.05
TF-GPLVM	2.74 ± 0.03	3.21 ± 0.02
TF-GPDM	2.98 ± 0.03	3.47 ± 0.02

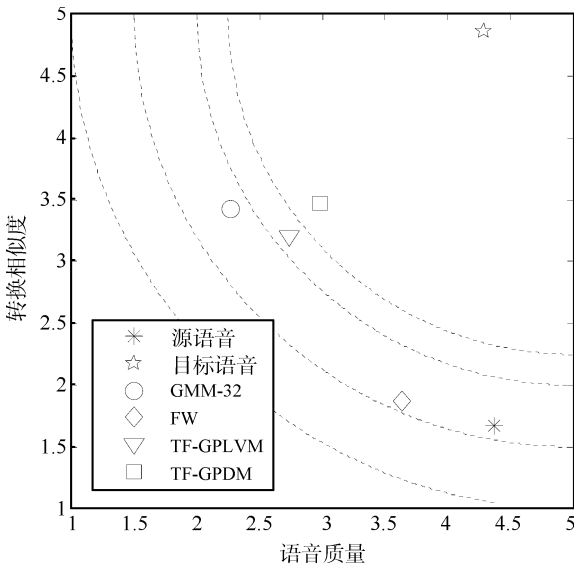

图 6 “语音质量 – 转换相似度” 实验结果

Fig. 6 Results of performance balance plotted according to quality versus similarity

由于实验对语音质量和说话人相似度的评测都采用 5 级 MOS 分, 两项性能指标的评分尺度相同. 为了更直观地比较各种方法的综合性能, 我们将语音质量–转换相似度两个评价结果作为坐标标识在图 6 上. 坐标的相对位置反映了语音转换性能的优劣, 越靠近坐标 (5, 5) 并且距离 (1, 1) 到 (5, 5) 对角线近的位置综合性能越好. 从图 5 可以看出, 相比于映射方法, 基于说话人特征和语义内容建模的方法, 尤其是本文提出的考虑了语音动态特性的 TF-GPDM, 更能兼顾语音质量和转换相似度两项指标, 综合性能更优.

5 结论

本文抛开传统的声道谱参数映射思路, 从说话人特征和语义内容分离角度并利用说话人特征替换方式研究语音转换. 首先, 针对 TF-GPLVM 模型对说话人特征和语义内容建模时未考虑语音动态特征并且算法效率低的问题, 引入 HMM 模型描述语音动态特征信息, 并利用 HMM 隐状态对声道谱参数进行关于语义内容的软分类, 提出了 TF-GPDM 模型. 然后, 利用一个包含多名说话人的平行语音集进行模型训练, 获得语音声道谱的说话人特征参数和语义内容参数. 接着, 在对输入语音进行 HMM 软分类基础上, 利用说话人特征替换实现了一种全新的声道谱转换方案. 最后, 本文对所提方法进行了主客观评价, 并和其他转换方法进行了性能比较. 实验结果表明, 与传统的映射转换方法相比, 本文方法不仅提高了语音质量和说话人相似度, 而且使这两项指标表现更加均衡; 与 TF-GPLVM 转换方法相比, TF-GPDM 也显示出了更优的综合转换性能.

本文对语音说话人特征和语义内容参数的维数以及描述两者之间相互作用的核函数都已固定, 实际上这两者的选择对语音建模精度影响很大. 另外, 本文对 HMM 的状态数设定为 32, 即对语音的内容分了 32 类, 其实这也是一个对实验结果存在较大影响的参数. 由于篇幅有限, 以上问题都未涉及, 值得我们进一步研究.

附录

由于 $\boldsymbol{y}_t$ 的每一维都对应相同的高斯过程映射, 这里只考虑其中一维的情况. 假设 y_t 表示 $\boldsymbol{y}_t$ 中的某一维, 式 (2) 中的向量 $\boldsymbol{w}_{ij}$ 也相应改为标量 w_{ij}. 令 $\boldsymbol{w}_j = [w_{1j}, w_{2j}, \cdots]^{\mathrm{T}}$, 且 $\boldsymbol{w}_j \sim \mathrm{N}(0, I)$, 得到:

$$k(\boldsymbol{a}, \boldsymbol{a}'; t, t') = \mathrm{E}(y_t y_{t'}) = \mathrm{E}\left[\left(\sum_{q=1}^{Q}\gamma_q(t)\sum_{i,j}\psi_j(\boldsymbol{a})w_{ij}\phi_i(\boldsymbol{x}_q)+\eta_t\right)\times\left(\sum_{q=1}^{Q}\gamma_q(t')\sum_{i,j}\psi_j(\boldsymbol{a}')w_{ij}\phi_i(\boldsymbol{x}_q)+\eta_{t'}\right)\right] =$$

$$\mathrm{E}\left[\left(\sum_{q=1}^{Q}\gamma_q(t)\sum_j\psi_j(\boldsymbol{a})\left(\boldsymbol{w}_j{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_q)\right)+\eta_t\right)\times\left(\sum_{q=1}^{Q}\gamma_q(t')\sum_j\psi_j(\boldsymbol{a}')\left(\boldsymbol{w}_j{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_q)\right)+\eta_{t'}\right)\right] =$$

$$\mathrm{E}\left[\left[\sum_{q=1}^{Q}\gamma_q(t)\sum_j\psi_j(\boldsymbol{a})\left(\boldsymbol{w}_j{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_q)\right)\right]\times\left[\sum_{q=1}^{Q}\gamma_q(t')\sum_j\psi_j(\boldsymbol{a}')\left(\boldsymbol{w}_j{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_q)\right)\right]\right]+$$

$$\mathrm{E}(\eta_t\eta_{t'}) = \sum_{q=1}^{Q}\gamma_q(t)\sum_{p=1}^{Q}\gamma_p(t')\times\mathrm{E}\left[\sum_j\psi_j(\boldsymbol{a})\left(\boldsymbol{w}_j{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_q)\right)\sum_l\psi_l(\boldsymbol{a}')\left(\boldsymbol{w}_l{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_p)\right)\right]+$$

$$\mathrm{E}(\eta_t\eta_{t'}) = \sum_{q=1}^{Q}\gamma_q(t)\sum_{p=1}^{Q}\gamma_p(t')\sum_j\psi_j(\boldsymbol{a})\psi_j(\boldsymbol{a}')\times\mathrm{E}\left[\left[\boldsymbol{w}_j{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_q)\right]\left[\boldsymbol{w}_j{}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_p)\right]\right]+\mathrm{E}(\eta_t\eta_{t'}) =$$

$$\sum_{q=1}^{Q}\gamma_q(t)\sum_{p=1}^{Q}\gamma_p(t')\boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{a})\boldsymbol{\psi}(\boldsymbol{a}')\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_q)\boldsymbol{\phi}(\boldsymbol{x}_p)+\beta^{-1}\delta_{\boldsymbol{a},\boldsymbol{a}';t,t'}$$

将式 (8) 和式 (9) 代入, 上式可以写成向量形式, 最后得到:

$$k(\boldsymbol{a}, \boldsymbol{a}'; t, t') = \boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{a})\boldsymbol{\psi}(\boldsymbol{a}')\left[\boldsymbol{\gamma}^{\mathrm{T}}(t)\Sigma\boldsymbol{\gamma}(t')\right]+\beta^{-1}\delta_{\boldsymbol{a},\boldsymbol{a}';t,t'}$$

References

1 Moulines E, Sagisaka Y. Voice conversion: state of the art and perspectives. *Special Issue of Speech Communication. The Netherlands*, 1995, **16**(2): 125−126

2 Furui S. Research of individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication*, 1986, **5**(2): 183−197

3 Abe M, Nakamura S, Shikano K, Kuwabara H. Voice conversion through vector quantization. In: Proceedings of the 1998 IEEE International Conference on Acoustic, Speech, and Signal Processing. New York, USA: IEEE, 1988. 655−658

4 Arslan L M. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*, 1999, **28**(3): 211−226

5 Narendranath M, Murthy H A, Rajendran S, Yegnanarayana B. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 1995, **16**(2): 207−216

6 Guido R C, Vieira L S, Júnior S B, Sanchez F L, Maciel C D, Fonseca E S, Pereira J C. A neural-wavelet architecture for voice conversion. *Neurocomputing*, 2007, **71**(1−3): 174−180

7 Desai S, Black A W, Yegnanarayana B, Prahallad K. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **18**(5): 954−964

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

9 Kain A B. High Resolution Voice Transformation [Ph. D. dissertation], OGI School of Science and Engineering at Oregon Health and Science University, United States, 2001

10 Toda T, Black A W, Tokuda K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(8): 2222−2235

11 Helander E, Virtanen T, Nurminen J, Gabbouj M. Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **18**(5): 912−921

12 Qiao Y, Saito D, Minematsu N. HMM-based sequence-to-frame mapping for voice conversion. In: Proceedings of the 2010 IEEE International Conference on Acoustic, Speech, and Signal Processing. Dallas, TX: IEEE, 2010. 4830−4833

13 Zen H, Nankaku Y, Tokuda K. Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(2): 417−430

14 Helander E, Silén H, Virtanen T, Gabbouj M. Voice conversion using dynamic kernel partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, **20**(3): 806−817

15 Valbret H, Moulines E, Tubach J P. Voice transformation using PSOLA technique. *Speech Communication*, 1992, **11**(2−3): 175−187

16 Sundermann D, Ney H. VTLN-based voice conversion. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology. Darmstadt, Germany: IEEE, 2003. 556−559

17 Shuang Z W, Bakis R, Qin Y. Voice conversion based on mapping formants. In: Proceedings of the 2006 TC-STAR Workshop on Speech-to-Speech Translation. Barcelona, Spain: ISCA, 2006. 219−223

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

19 Erro D, Moreno A, Bonafonte A. Voice conversion based on weighted frequency warping. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **18**(5): 922−931

20 Toda T, Saruwatari H, Shikano K. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City, USA: IEEE, 2001. 841−844

21 Belin P, Zatorre R J, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. *Nature*, 2000, **403**(6767): 309−312

22 Minematsu N. Human speech model based on information separation and its application to speech processing. In: Proceedings of the 7th International Symposium on Chinese Spoken Language Processing. Tainan, China: IEEE, 2010. 477−482

23 Latinus M, Belin P. Human voice perception. *Current Biology*, 2011, **21**(4): R143−R145

24 Popa V, Nurminen J, Gabbouj M. A novel technique for voice conversion based on style and content decomposition with bilinear models. In: Proceedings of the 2009 Interspeech. Brighton, UK: ISCA, 2009. 2655−2658

25 Xu N, Yang Z, Zhang L H, Zhu W P, Bao J Y. Voice conversion based on state-space model for modelling spectral trajectory. *Electronics Letters*, 2009, **45**(14): 763−764

26 Sun X J, Zhang X W, Cao T Y, Yang J B, Sun J. Voice conversion using a two-factor Gaussian process latent variable model. *Przeglad Elektrotechniczny*, 2012, **88**(12a): 318−324

27 Rasmussen C E, Williams C K I. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006. 7−13

28 Knagenhjelm H P, Kleijin W B. Spectral dynamics is more important than spectral distortion. In: Proceedings of the 1995 IEEE International Conference on Acoustic, Speech, and Signal Processing. Detroit, USA: IEEE, 1995. 732−735

29 Duxans H, Bonafonte A, Kain E, van Santen J. Including dynamic and phonetic information in voice conversion systems. In: Proceedings of the 2004 International Conference on Spoken Language Processing. Jeju Island, Korea: ISCA, 2004. 1193−1196

30 Duxans H. Voice Conversion Applied to Text-to-Speech Systems [Ph. D. dissertation], Polytechnic University of Catalonia, Barcelona, 2006

31 Møller M F. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 1993, **6**(4): 525−533

32 Kominek J, Black A W. The CMU ARCTIC speech databases. In: Proceedings of the 5th ISCA Speech Synthesis Workshop. Pittsburgh, USA: ISCA, 2005. 223−224

33 Erro D, Moreno A, Bonafonte A. Flexible harmonic/stochastic speech synthesis. In: Proceedings of the 6th ISCA Workshop on Speech Synthesis. Bonn, Germany: ISCA, 2007. 194−199

孙新建 中国人民解放军理工大学通信工程学院博士研究生. 主要研究方向为语音信号处理, 机器学习. 本文通信作者.
E-mail: sunxinjane@163.com
(**SUN Xin-Jian** Ph. D. candidate at the College of Communication Engineering, PLA University of Science and Technology. His research interest covers speech signal processing and machine learning. Corresponding author of this paper.)

张雄伟 中国人民解放军理工大学指挥信息系统学院教授. 主要研究方向为多媒体信息处理, 智能计算, 压缩感知.
E-mail: xwzhang@public1.ptt.js.cn
(**ZHANG Xiong-Wei** Professor at the College of Command Information Systems, PLA University of Science and Technology. His research interest covers multimedia information processing, intelligent computing, and compressive sensing.)

杨吉斌 中国人民解放军理工大学指挥信息系统学院副教授. 主要研究方向为信号参数估计, 语音识别和转换, 智能信息处理. E-mail: yjbice@sina.com
(**YANG Ji-Bin** Associate professor at the College of Command Information Systems, PLA University of Science and Technology. His research interest covers signal parameter estimation, speech recognition, voice conversion, and intelligent information processing.)

曹铁勇 中国人民解放军理工大学指挥信息系统学院教授. 主要研究方向为智能计算, 图像融合.
E-mail: ctyice@163.com
(**CAO Tie-Yong** Professor at the College of Command Information Systems, PLA University of Science and Technology. His research interest covers intelligent computing and image fusion.)

钟新毅 中国人民解放军理工大学通信工程学院博士研究生. 主要研究方向为信道编码和网络编码.
E-mail: zhongxy2010@gmail.com
(**ZHONG Xin-Yi** Ph. D. candidate at the College of Communication Engineering, PLA University of Science and Technology. His research interest covers channel coding and network coding in wireless communication.)