约束条件下的结构化高斯混合模型及非平行语料语音转换

车滢霞,俞一彪
(苏州大学电子信息学院, 江苏苏州 215006)

摘 要: 提出一种约束条件下的结构化高斯混合模型及非平行语料语音转换方法.从源与目标说话人的原始非平行语料中提取出少量相同音节,在结构化高斯混合模型的训练过程中,利用这些相同音节包含的语义信息及声学特征对应关系对 K 均值聚类中心进行约束,并在(Expectation Maximum,EM)迭代过程中对语音帧属于模型分量的后验概率进行修正,得到基于约束的结构化高斯混合模型(Structured Gaussian Mixture Model with Constraint condition,C-SGMM).再利用全局声学结构(Acoustic Universal Structure,AUS)原理对源和目标说话人的约束结构化高斯混合模型的高斯分布进行匹配对准,推导出短时谱转换函数.主观和客观评价实验结果表明,使用该方法得到的转换后语音在谱失真,目标倾向性和语音质量等方面均优于传统的结构化模型语音转换方法,转换语音的平均谱失真仅为0.52,说话人正确识别率达到95.25%,目标语音倾向性指标ABX平均为0.82,性能更加接近于基于平行语料的语音转换方法.

关键词: 语音转换;结构化高斯混合模型;非平行语料;约束条件
中图分类号: TN912.33 **文献标识码:** A **文章编号:** 0372-2112 (2016) 09-2282-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.09.37

Non-parallel Corpora Voice Conversion Based on Structured Gaussian Mixture Model Under Constraint Conditions

CHE Ying-xia, YU Yi-biao
(*School of Electronic and Information Engineering, Soochow University, Suzhou, Jiangsu* 215006, *China*)

Abstract: This paper proposes a structured Gaussian mixture model with constraint conditions (C-SGMM) for non-parallel corpora voice conversion. A small number of voice signals with the same syllables from the source and target non-parallel corpus are extracted as constraint conditions, then the correspondence between acoustic features of source and target corpus formed by these syllables are applied in the process of statistical acoustic model training. The constraint conditions are used to restrict the cluster centers of *K*-means clustering process, and they are also used in EM algorithm to adjust the voice frame's posterior probability belonging to a Gaussian distribution component for model training. Then Gaussian distributions in source and target structured Gaussian mixture models are aligned using acoustic universal structure principle and the conversion function can be derived. Results of both subjective and objective experiments indicate that the conversion performance obtained by the proposed method are advanced to that of the traditional structured method in cepstrum distortion, target tendency and speech quality aspects. The average cepstrum distortion of converted speech is only 0.52, the speaker recognition rate of the converted speech reaches 95.25%, and the performance closer to the conventional parallel corpora GMM based method is achieved.

Key words: voice conversion; structure Gaussian mixture model; non-parallel corpora; constraint conditions

1 引言

语音转换是指将 A 说话人的语音进行转换并使其听起来像 B 说话人的语音,且保持语义内容不变的一种技术[1].语音转换、尤其是基于非平行语料训练的语音转换是目前语音研究领域比较新的课题,对于具有

收稿日期:2015-02-08;修回日期:2015-08-10;责任编辑:梅志强
基金项目:国家自然科学基金(No.61271360);江苏省自然科学基金(No.BK20131196)

表现力的语音合成、语音伪装通信、多媒体配音和残疾人发声等发面都有很广泛的应用价值,因此近年来得到越来越多研究者的关注.

传统的说话人语音转换方法大多采用平行语料的联合训练获得转换函数[2~4].但由于平行训练语料在实际应用中难以获得,近年来一些学者在基于非平行语料的语音转换方面进行研究,并取得了一定的成果. Mouchtaris 等利用说话人自适应技术,通过特定人的平行语料训练推导出非平行语料下的转换函数[5],但首先要对参考说话人语音进行充分的联合训练获得参考模型,再分别进行自适应获得转换函数,计算量大且过程复杂;Erro 等将最近邻搜索算法(N-N)与语音转换相结合[6],通过不断迭代转换过程后达到理想转换效果,这一方法不仅迭代过程计算量大,而且最近邻搜索算法难以保证正确的声学特征对准处理;Saito 等通过建立噪声信道模型实现基于少量平行语料训练的非平行语料下的语音转换[7],但联合模型的训练仍然需要少量的高质量平行语料;曾道建等提出了结构化语音转换方法[8],通过对结构化高斯混合模型(Structured Gaussian Mixture Model,SGMM)在特征空间的对准实现说话人转换,由于非平行语料的语音成份对应关系难以正确保证,转换效果受到很大的影响.

本文提出一种约束条件下的结构化高斯混合模型并将其应用于非平行语料语音转换. 首先从源与目标说话人非平行语料中提取出少量相同音节作为约束信息,利用其包含的语义信息及声学特征对应关系,在结构化高斯混合模型(SGMM)的训练过程中,约束 K 均值聚类的聚类中心以及修正 EM 过程中语音帧属于某高斯分布的后验概率,得到基于约束的结构化高斯混合模型(C-SGMM). 再利用全局声学结构(AUS)原理对源说话人和目标说话人的 C-SGMM 进行高斯分布对准,得到短时谱转换函数,实现语音转换. 通过主观和客观评价准则对转换后的语音进行实验测评,使用该方法得到的转换后语音相比于传统的结构化语音转换方法[8]降低了谱失真,提高了目标倾向性和语音质量,转换性能更加接近于传统的基于平行语料的语音转换方法[2~4].

2　系统构成

图 1 描述了约束条件下的结构化高斯混合模型应用于非平行语料语音转换的系统构成.

语音转换系统分为训练与转换两个部分. 在训练阶段,对源语音和目标语音进行 STRAIGHT 分析,提取出短时谱及基频,从短时谱中提取出线性预测倒谱系数(LPCC). 与此同时,从源说话人及目标说话人的非平行训练语料中提取出相同的音节,经过相同的 STRAIGHT 分析过程,其特征参数进行联合训练得到联合分布的高斯混合模型,该联合模型包含了源和目标语音成份的对应关系. 然后,对该模型各个高斯分布进行标记,标明源与目标高斯分布之间的对应关系,作为约束信息指导源与目标语音 LPCC 特征参数的聚类和带约束的 SGMM 建模,得到源与目标说话人语音各自的 C-SGMM. 而 F0 则由单高斯分布描述. 训练好的源与目标 C-SGMM 在保证由联合模型得到的高斯分布对应关系的前提下,通过 AUS 原理进行其它高斯分布的匹配对准,继而推导出非平行语料语音短时谱转换函数.

转换阶段与传统的基于高斯混合模型(GMM)的平行语料语音转换类似,利用转换公式分别对 LPCC 特征参数和 F0 转换并合成后,得到转换后的语音.

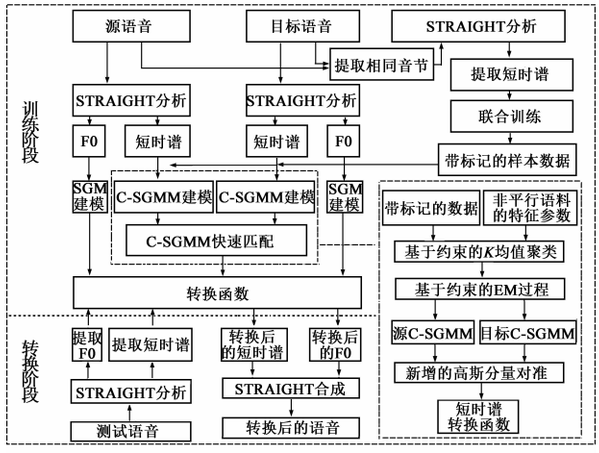


图1　基于约束条件的结构化高斯混合模型的非平行语料语音转换系统构成

3　约束条件下的高斯混合模型

C-SGMM 是从源和目标说话人的原始非平行训练语料中提取少量相同音节作为约束信息,在 SGMM 的训练过程中约束 K 均值聚类中心的产生,同时修正 EM 过程中某语音帧对应某高斯分量的后验概率进行迭代训练. 因此,本节将首先对 SGMM 进行简要描述及分析,然后针对 C-SGMM 的训练过程,从约束信息引入 K 均值聚类和 EM 过程两个方面进行介绍.

3.1　结构化高斯混合模型及其分析

高斯混合模型(GMM)是单一高斯概率密度函数的延伸,由于 GMM 能够平滑地近似任意形状的密度分布,因此近年来常被用在语音识别,语音转换等方面[9]. 结构化高斯混合模型[8]将高斯混合模型结构化,描述出高斯混合模型的各个单高斯分布之间的结构特性,如图 2. 首先用 EM 算法估计出高斯混合模型的参数 $(\boldsymbol{\pi}_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$,然后采用 Bhattacharyya 距离测度计算高斯混合模型中各个单高斯分布之间的相似性,定义为

两个单高斯分布的距离. Bhattacharyya 距离测度(BD)计算公式如下:

$$\mathrm{BD}(p_i(\boldsymbol{x}),p_j(\boldsymbol{x})) = -\ln\int_{-\infty}^{\infty}\sqrt{p_i(\boldsymbol{x})p_j(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \quad (1)$$

结构化高斯混合模型不仅描述了说话人语音特征的统计分布,而且描述了这些特征分布之间的结构关系. 由于高斯混合模型的每一个分量对应一个可分辨的语音特征分布,并且相同的语音成份具有相似的语音特征分布,因此,如果源和目标说话人的训练语料足够充分,使各语音成份能够相对平衡,那么将其语音各自进行结构化高斯混合模型建模之后,即使训练语料不平行,相同的语音成份及其特征分布也能通过不断调整结构化高斯混合模型中高斯分量在其中的位置达到相对对准,实现高斯分布的一一对应,推导出语音短时谱转换函数[8].

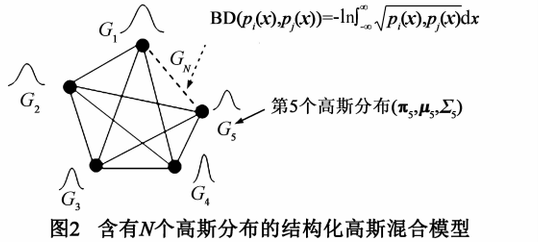


图2 含有*N*个高斯分布的结构化高斯混合模型

但是在实际情况中,非平行训练语料往往是非理想的. 不同说话人发音习惯各有不同,那么源和目标语音的声学特征之间则存在较大离散度,其 SGMM 之间的成份很可能没有潜在的一一对应关系. 再加上源和目标说话人的 SGMM 训练是独立进行的,缺少监督信息,那么强制对准建立的源-目标语音成份声学特征间的对应关系是不够精确的,从而影响整体转换性能.

而在基于平行语料的说话人语音转换中,由于平行训练语料本身就存在语音成份上的对应关系,联合训练正是利用了这种对应关系,因此转换语音从清晰度,可懂度和目标倾向性等方面均表现出较好的性能,但完全平行的训练语料难以获得. 然而,在源和目标说话人原始的非平行训练语料中,少量相同的音节很容易得到并且被提取出来. 这些相同的音节包含了一定的语义信息[10],也包含了源和目标语音成份及其声学特征的对应关系. 显然,这种对应关系可以作为约束信息加以有效利用,在 SGMM 训练和匹配对准时起指导作用,也就是说 SGMM 的建模和对准以源和目标的少量相同音节指示的对应关系为基础,其它大量非平行语料在此基础上对模型进行分量的扩展和微调,充分利用语料本身存在的对应关系,使最终语音转换的效果更接近于平行语料语音转换.

在对源和目标语音分别进行 C-SGMM 建模时,需要先用 EM 算法估计出高斯混合模型的参数($\boldsymbol{\pi}_k$,$\boldsymbol{\mu}_k$,$\boldsymbol{\Sigma}_k$),约束信息对 C-SGMM 建模的指导作用即体现在 EM 算法中. 由于 EM 算法对初始值敏感,其初始迭代值可由基于密度的 K 均值聚类算法[11]产生,所以,将约束信息引入 EM 算法的同时也包括了将约束信息引入 K 均值聚类算法. 以下将从基于约束的 K 均值聚类和基于约束的 EM 算法两方面来进行阐述.

3.2 基于约束的 *K* 均值聚类

在对所有训练语料声学特征参数进行统计建模时,需要提取出源与目标说话人训练语料中少量相同音节联合训练作为约束信息指导 C-SGMM 建模. 基于这些音节样本联合训练得到 GMM 的模型参数,进一步计算得到每个高斯分量对应训练数据中哪些样本. 属于同一个高斯分量的样本集可以看作一个簇,该簇内的样本具有相同的簇标记,包含了相似的语音特征参数,这些样本在聚类时必须被聚到同一类中.

定义带 M 个簇标记的数据样本集 X 为

$$\{\boldsymbol{x}_1^1,\boldsymbol{x}_2^1,\cdots,\boldsymbol{x}_a^1,\boldsymbol{x}_1^2,\boldsymbol{x}_2^2,\cdots,\boldsymbol{x}_b^2,\cdots,\boldsymbol{x}_1^M,\boldsymbol{x}_2^M,\cdots,\boldsymbol{x}_m^M\}$$

$\{\boldsymbol{x}_1^1,\boldsymbol{x}_2^1,\cdots,\boldsymbol{x}_a^1\}$ 属于同一个簇,带有簇标记 1,簇均值为 $\boldsymbol{\mu}_{x_1}$,$\{\boldsymbol{x}_1^2,\boldsymbol{x}_2^2,\cdots,\boldsymbol{x}_b^2\}$ 属于同一个簇,带有簇标记 2,簇均值为 $\boldsymbol{\mu}_{x_2}$,$\{\boldsymbol{x}_1^M,\boldsymbol{x}_2^M,\cdots,\boldsymbol{x}_m^M\}$ 属于同一个簇,带有簇标记 M,簇均值为 $\boldsymbol{\mu}_{x_M}$.

由于相同音节数目的有限性,其特征参数联合训练的分布不能充分表达该说话人语音成份的特征分布,这就需要通过大量非平行训练语料对模型所包含的分布进行适当的扩充,使得模型能充分描述该说话人完整的语音声学特征.

定义原始非平行训练语料的特征参数集 Y 为 $\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_I\}$,由联合训练得到的带标记的样本集 X 为 $\{\boldsymbol{x}_1^1,\boldsymbol{x}_2^1,\cdots,\boldsymbol{x}_\alpha^1,\boldsymbol{x}_1^2,\boldsymbol{x}_2^2,\cdots,\boldsymbol{x}_b^2,\cdots,\boldsymbol{x}_1^M,\boldsymbol{x}_2^M,\cdots,\boldsymbol{x}_m^M\}$,则对 Y 进行约束的半监督 K 均值聚类过程描述如下:

步骤 1 确定初始聚类中心和聚类数目 $S(S>M)$. 将 $\{\boldsymbol{\mu}_{x_1},\boldsymbol{\mu}_{x_2},\cdots,\boldsymbol{\mu}_{x_M}\}$ 作为前 M 个初始值,将 $\boldsymbol{\mu}_{x_1},\boldsymbol{\mu}_{x_2},\cdots,\boldsymbol{\mu}_{x_M}$ 的 ε 邻域(实验结果表明,本实验使用的语料库中女性说话人的 $\varepsilon=1.1$,男性说话人的 $\varepsilon=1.3$ 时可使 K 均值聚类的误差相对较小)以外的 $\boldsymbol{y}_i\in Y$ 按基于密度的 K 均值聚类算法扩充出(S-M)个初始值[11,12];

步骤 2 对 Y 进行聚类. 计算所有 $\boldsymbol{y}_i\in Y$ 与聚类中心 $C\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_S\}$ 的距离,根据距离将 $\boldsymbol{y}_i$ 划分到最近的簇 $\boldsymbol{c}_s$ 中;

步骤 3 对 X 进行聚类. X 中的数据本身带标记,以 $\boldsymbol{\mu}_{x_1},\boldsymbol{\mu}_{x_2},\cdots,\boldsymbol{\mu}_{x_M}$ 为均值的的簇分别被聚类到 $\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_M\}$ 中;

步骤 4 更新聚类中心. 第 t 次迭代时,为了避免由于相同音节数少而导致的训练不充分性,第 j 个聚类中心为

$$\boldsymbol{C}_{jt}=\frac{\sum_{i=1}^{|\boldsymbol{C}_j|}\boldsymbol{y}_{ji}+N_j\times\boldsymbol{\mu}_{x_j}}{|\boldsymbol{C}_j|+N_j} \tag{2}$$

其中,$|\boldsymbol{C}_j|$为第 j 类中无标记样本 $\boldsymbol{y}_{ji}$ 的总数,N_j 为聚到第 j 类中的带标记数据的总数,$\boldsymbol{\mu}_{x_j}$ 为其均值,在 $j>M$ 时,N_j 和 $\boldsymbol{\mu}_{x_j}$ 均为 0;

步骤 5　重复步骤 2 ~ 步骤 4 直至收敛或达到最大迭代次数.

3.3　基于约束的 EM 算法

EM 算法是一种迭代算法,包含了计算期望(E 步)和最大化(M 步)两步,E 步估计未知参数的期望值,给出当前的参数估计,M 步重新估计分布参数,使得数据的似然性最大,给出未知变量的期望估计. C-SGMM 的参数$(\boldsymbol{\pi}_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$使用带约束条件的 EM 算法进行估计,在带约束 K 均值聚类的基础上使用 EM 算法迭代计算,并在迭代过程中融入小样本平行语料约束信息.

首先估计样本数据对应每个高斯分量的后验概率. 对于每个样本数据 $\boldsymbol{y}_i$,由第 k 个分量产生的后验概率为

$$\gamma(i.k)=\frac{\boldsymbol{\pi}_k N(\boldsymbol{y}_i\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j-1}^{k}\pi_j N(\boldsymbol{y}_i\mid\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} \tag{3}$$

其中 $N(\boldsymbol{y}_i\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ 表示第 k 个高斯分布的概率密度函数.

然后通过最大似然估计可得到模型参数更新公式

$$\pi_k=\frac{N_k}{N} \tag{4}$$

$$\boldsymbol{\mu}_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)\boldsymbol{y}_i \tag{5}$$

$$\boldsymbol{\Sigma}_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)(\boldsymbol{y}_i-\boldsymbol{\mu}_k)(\boldsymbol{y}_i-\boldsymbol{\mu}_k)^{\mathrm{T}} \tag{6}$$

其中 $N_k=\sum_{i=1}^{N}\gamma(i,k)$,$N$ 为样本总数. 重复上述过程,直至收敛.

通过其模型参数迭代更新的公式可知,$\boldsymbol{y}_i$ 及其后验概率不仅影响其本身所属分量的参数更新,同时也参与了其它分量的参数更新,也就是说,本应该属于第 k 个分量的样本会以其后验概率对其它模型分量的参数产生或大或小的干扰[12],EM 算法迭代的过程即是减小这种干扰的过程,使样本以更大的后验概率属于某一分布,以更小的后验概率属于其它分布.

根据前面的论述,带有相同簇标记的样本应该属于同一个分量. 根据簇平滑思想,通过加入了少量带标记的数据样本训练得到的 SGMM 模型参数应该使得带有相同簇标记的样本的各个后验概率相似,其不相似度在迭代的过程中可作为修正后验概率的因子[13].

带标记的数据样本 $\boldsymbol{x}_{sj}$ 属于第 s 个分布的后验概率为 $p(s\mid\boldsymbol{x}_{sj},\boldsymbol{\theta})(s=1,2,\cdots,S)$,对于该分布中的其它样本,其 $p(s\mid\boldsymbol{y}_{sn},\boldsymbol{\theta})$ 与 $p(s\mid\boldsymbol{x}_{sj},\boldsymbol{\theta})$ 应该尽可能的相似,其相似度函数可定义为

$$L_{sn}=\frac{1}{N_S}\times\sum_{j=1}^{N_s}\left|p(s\mid\boldsymbol{y}_{sn},\boldsymbol{\theta})-p(s\mid\boldsymbol{x}_{sj},\boldsymbol{\theta})\right| \tag{7}$$

该值越大,说明本应相同的值却相差很大,$p(s\mid\boldsymbol{y}_{sn},\boldsymbol{\theta})$ 应作出较大的使之降低的调整;反之说明应该相同的值相差很小,为保证联合训练得出的对应关系,$p(s\mid\boldsymbol{y}_{sn},\boldsymbol{\theta})$ 不作调整或者不作很大的调整. 可定义修正因子为

$$\rho_{sn}=\frac{1}{\sqrt{1+L_{sn}}} \tag{8}$$

调整后的 $p_l(s\mid\boldsymbol{y}_{sn},\boldsymbol{\theta})=p(s\mid\boldsymbol{y}_{sn},\boldsymbol{\theta})\cdot\rho_{sn}$. 因此,每次 EM 迭代时都使用调整后的后验概率进行模型参数的更新,则模型参数的更新公式为式(3) ~ 式(5),其中

$$\gamma(i.k)=\gamma'(i.k)=\frac{\boldsymbol{\pi}_k N(\boldsymbol{y}_i\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j-1}^{k}\pi_j N(\boldsymbol{y}_i\mid\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}\cdot\rho \tag{9}$$

理论上该修正能加快 EM 算法的收敛速度,使 EM 算法得到的模型参数更符合用户期望.

由式(2)和(9)可知,第 j 类中的样本个数 N_j 影响着 $\boldsymbol{C}_j$ 和 $\gamma'(i.k)$,也就是相同音节的个数影响着基于约束的 K 均值聚类中聚类中心的产生和基于约束的 EM 过程中样本属于某分量的后验概率. N_j 越大,相同音节对 C-SGMM 训练的约束性越强,得到的 C-SGMM 中的分量越接近于平行训练语料的 GMM 的分量,理论上转换效果越好.

在本文的实验中,考虑到实际情况中非平行训练语料的局限性,体现“少量约束信息”,从原始非平行训练语料中提取出 54 个相同音节,每个音节约为 30 帧. 同时,为保证基于相同音节联合训练得出的对应关系不发生太大的偏离,通过观察相似度函数值的统计直方图和 EM 迭代过程中似然值,将女性说话人进行后验概率调整的阈值设为 4.5,男性说话人进行后验概率调整的阈值设为 4.3. 后验概率在相似度函数值达到所设阈值才进行调整,否则不作调整.

最后,求得模型参数$(\boldsymbol{\pi}_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$后再计算每个高斯分布之间的 Bhattacharyya 距离,则 C-SGMM 训练完成.

4　基于 C-SGMM 的非平行语料语音转换

C-SGMM 不仅描述了说话人语音特征分布,而且描述了这些特征分布之间的结构关系,其应用于语音转换的过程是通过不断调整源说话人 C-SGMM 中高斯分布的位置,使源说话人 C-SGMM 调整之后与目标说话人 C-SGMM 在高斯分布上有正确的一一对应的关系,即语音成份上的对应关系,从而实现语音转换.

不同说话人发音的倒谱特征参数变化表现为一种

线性映射关系[13],理论上,不同说话人发出的相同语音,倒谱特征参数在 AUS 中的 Bhattacharyya 距离是相同的,即

$$\mathrm{BD}(p_i(\boldsymbol{x}'),p_j(\boldsymbol{x}'))=\mathrm{BD}(p_i(\boldsymbol{x}),p_j(\boldsymbol{x})) \quad (10)$$

其中,$\boldsymbol{x}$ 是源语音的到谱特征参数,$\boldsymbol{x}'$是相同语音的目标语音的倒谱特征参数.

上式表明,尽管不同说话人发出的同一语音具有不同的声学特征分布,但在倒谱空间中其结构是相同的,只是位置发生了变化. AUS 描述一个语音的内在声学特征结构关系[14],通过 AUS 不变性原理将两个 C-SGMM 中的声学特征分布进行对准,从而可推导出语音转换公式.

两个 C-SGMM 之间的距离定义为:

$$D(S,T)=\sqrt{\frac{1}{N}\sum_{i<j}(s_{ij}-t_{ij})^2},\ 1\leqslant i,j\leqslant N \quad (11)$$

其中,s_{ij},t_{ij}分别表示在源 C-SGMM 中和目标 C-SGMM 中连接节点 i,j 的边,即式(1)所表示的 Bhattacharyya 距离. 当下式成立时,表明源说话人的 C-SGMM 与目标说话人的 C-SGMM 之间的差异最小,此时两模型之间高斯分量实现对准,即语音声学特征已经对准.

$$D(S',T)=\sqrt{\frac{1}{N}\sum_{i<j}(s'_{ij}-t_{ij})^2}<D(S,T) \quad (12)$$

$$s'_{ij}=s_{f^*(i)f^*(j)},1\leqslant i,j\leqslant N,1\leqslant f^*(\cdot)\leqslant N \quad (13)$$

$$f^*(\cdot)=\mathrm{argmin}_{f(\cdot)}\sqrt{\frac{1}{N}\sum_{i<j}[s_{f(i)f(j)}-t_{ij}]^2}$$

$$1\leqslant i,j\leqslant N,1\leqslant f(\cdot)\leqslant N \quad (14)$$

则短时谱的转换函数可以表示为:

$$F(\boldsymbol{x})=\sum_{i=1}^{N}p_i(\boldsymbol{x})\left[\boldsymbol{\mu}_{f(i)}^{\mathrm{T}}+\frac{\boldsymbol{\Sigma}_{f(i)}^{\mathrm{T}}}{\boldsymbol{\Sigma}_i^S}(\boldsymbol{x}-\boldsymbol{\mu}_i^S)\right] \quad (15)$$

$p_i(\boldsymbol{x})$是源说话人 C-SGMM 中第 i 个高斯分布的概率,$\boldsymbol{\mu}_i^S$ 和 $\boldsymbol{\Sigma}_i^S$ 是该高斯分布的均值和协方差矩阵,$f(i)$是源说话人 C-SGMM 中第 i 个高斯分布对应的目标说话人 C-SGMM 的高斯分布序列号,$\boldsymbol{\mu}_{f(i)}^{\mathrm{T}}$和 $\boldsymbol{\Sigma}_{f(i)}^{\mathrm{T}}$是该高斯分布的均值和协方差矩阵.

基于 C-SGMM 的语音转换的核心是特征参数高斯分布的对准. 由于每个高斯分布对应着一个特定的语音成份,如果由于算法的不精确导致模型的对准有偏差,则语音成份的对准也存在偏差,那么合成的语音就会词序混乱,表意不明,所以 C-SGMM 中的高斯分量的对准是十分重要的,直接影响语音的可懂性. 若要得到使对准式(11)~(13)成立的全局最优解,理论上需要对模型中的高斯分布进行全排列,但这个搜索过程过于庞大,对计算机的性能要求极高,因此需要考虑能平衡搜索时间与模型匹配精度矛盾的局部最优算法. 所以,在 C-SGMM 的对准过程中使用基于爬山算法的快速模型匹配算法[15].

由于 C-SGMM 的训练不改变基于少量相同音节联合训练得到的源和目标语音在语音成份上的对应关系,所以在模型匹配过程中只需对新增的高斯分量使用快速匹配算法进行对准,由少量相同音节确定的高斯分布默认已经一一对准,不再进行匹配对准.

除了短时谱,另一个表征说话人个性特征的参数是基音频率 F0. 本文使用单高斯分布描述源说话人和目标说话人的基音频率分布特性,由此得到基音频率的转换公式为

$$\mathrm{F0}^{\mathrm{T}}=\mu^{\mathrm{T}}+\frac{\sigma^{\mathrm{T}}}{\sigma^S}(\mathrm{F0}^S-\mu^S) \quad (16)$$

其中,μ^S 和 σ^S 表示源说话人基频的均值和方差,μ^{T} 和 σ^{T} 表示目标说话人基频的均值和方差.

5 实验与分析

设计了四个实验来评价本文提出的方法是否有效. 其中包括两个客观评价实验和两个主观评价实验,并且与基于 GMM 的平行语料语音转换及传统的结构化非平行语料语音转换结果进行比较. 本实验中采用的训练和测试语料为均在安静的环境下录制的语料库 SUDA-3[8],训练语料包括两段男声和两段女声(记为 F1,F2,M1,M2),包含语音成份相对平衡,持续时间均为约 3 分钟. 测试语料为 F1、F2、M1 和 M2 的各 40 段语音,每段语音持续时间约 3 秒钟. 录制训练和测试语料时采样率均设为 16kHz,量化位均为 16. 联合训练的高斯分布数为 64. 非平行训练语料在使用 C-SGMM 建模时,均采用 128 个高斯分布,即在原 64 个高斯分布的基础上新增 64 个高斯分布. LPCC 特征参数均设为 39 阶.

5.1 客观评价

本小节设计了两个客观评价实验,分别为说话人识别(SR)测试和谱失真(CD)测度. 下列表格中,GMM 表示使用传统的基于 GMM 的平行语料语音转换方法,SGMM 表示传统结构化方法,C-SGMM 表示基于约束的结构化高斯混合模型方法. 客观测评结果如表 1 和表 2.

说话人识别测试主要是通过测试转换后语音属于源说话人和目标说话人语音的似然度来评价转换性能. 建立四个说话人(F1,F2,M1,M2)的高斯混合模型,高斯混合模型中的高斯分布数为 16,特征矢量为 12 维 MFCC 参数以及 12 维一阶 MFCC 差分值.

表 1 客观测评结果(SR)

转换方向	SR(%)		
	GMM	SGMM	C-SGMM
M1-M2	92	86	88
M1-F1	100	100	100
F1-M1	100	100	100
F1-F2	95	92	93
平均值	96.75	94.50	95.25

表 2　客观测评结果(CD)

转换方向	CD		
	GMM	SGMM	C-SGMM
M1-M2	0.38	0.54	0.52
M1-F1	0.36	0.61	0.53
F1-M1	0.36	0.54	0.47
F1-F2	0.34	0.63	0.56
平均值	0.36	0.58	0.52

谱失真测度测评转换后语音的倒谱与目标语音倒谱之间的差异性,表示为

$$\mathrm{CD} = \frac{\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{M}(T_{i,j}-S'_{ij})^2}}{N} \tag{17}$$

其中,N 为语音的帧数,M 为特征参数 LPCC 的阶数,T 表示目标语音的 LPCC,而 S'表示转换后语音的 LPCC.

从表 1 的实验结果可知,采用基于 C-SGMM 的语音转换方法,转换后语音的平均正确识别率达到 95.25%,比传统的结构化方法高了 0.75%,更加接近于传统的基于 GMM 的平行语料语音转换方法. 由表 2 可知,使用该方法得到的平均谱失真测度相比于传统的结构化方法降低了 10.3%,充分说明了使用该方法得到的谱包络更加接近于目标谱包络.

5.2　主观评价

设计了两个主观评价实验来测试转换后语音的质量,分别为 ABX 测试和 MOS 测试,测试人数为 20 人. 实验结果如表 3 和表 4.

在 ABX 测试中,听者判断转换后的语音更加接近与源语音还是目标语音,接近源语音则给出评分 0 分,接近目标语音则给出评分 1 分. 在 MOS 测试中,听者根据听到的语音的质量对该语音进行打分,评分分为 5 个等级:1 分表示很差,2 分表示较差,3 分表示一般,4 分表示较好,5 分表示很好.

表 3　主观测评结果(ABX)

转换方向	ABX		
	GMM	SGMM	C-SGMM
M1-M2	0.62	0.58	0.60
M1-F1	1	1	1
F1-M1	1	1	1
F1-F2	0.70	0.66	0.68
平均值	0.83	0.81	0.82

表 4　主观测评结果(MOS)

转换方向	MOS		
	GMM	SGMM	C-SGMM
M1-M2	3.8	3.5	3.6
M1-F1	3.5	3.3	3.4
F1-M1	3.4	2.9	3.3
F1-F2	3.6	3.4	3.5
平均值	3.575	3.275	3.45

从表中各个方法的 ABX 与 MOS 得分可以看出,采用 C-SGMM 方法得到的实验结果相比于传统的结构化方法,从目标倾向性和语音质量两方面更加接近于基于 GMM 的平行语料语音转换方法的性能.

6　总结

本文论述了基于约束条件的结构化高斯混合模型及基于该模型的非平行语料语音转换方法. 该方法不需要平行训练语料,克服了传统结构化方法的语音成份对应关系的问题并且计算量相对较小. 约束信息从源与目标非平行语料中提取并应用到 SGMM 的训练中,通过对 K 均值聚类初始值的选取和聚类中心迭代的约束,将约束信息引入 K 均值聚类;通过相似度函数对样本所属类别的后验概率进行调整,将约束信息引入 EM 算法中,最终得到 C-SGMM. C-SGMM 中新增的高斯分布利用基于 AUS 原理的快速匹配算法进行匹配对准,而由约束信息确定的高斯分布默认对准,从而得到短时谱转换函数进行语音转换. 主观和客观实验结果表明,使用该方法得到的语音转换性能相比于传统的结构化语音转换方法有较大提高,更加接近于传统的基于平行语料的语音转换方法.

由于基于约束的 EM 算法中进行后验概率调整的阈值随不同说话人改变,与基频相关性未知,故后续研究中需对此进行进一步探究.

参考文献

[1] Stylianou, Y. Voice transformation: A survey[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. Taipei: IEEE, 2009. 3585 – 3588.

[2] 康永国,双志伟,陶建华,张维. 基于混合映射模型的语音转换算法研究[J]. 声学学报,2006,31(6):555 – 562.
Kang Yongguo, Shuang Zhiwei, Tao Jianhua, et al. A hybrid method to convert acoustic features for voice conversion [J]. Acta Acustica, 2006, 31(6): 555 – 562. (in Chinese)

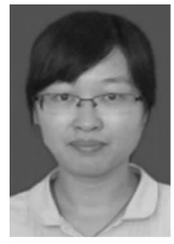
[3] 徐宁,杨震,张玲华. 基于状态空间模型的子频带语音转换算法[J]. 电子学报,2010,38(3):646 – 653.
Xu Ning, Yang Zhen, Zhang Ling-hua. Sub-and voice morphing algorithm based on state-space model[J]. Acta Electronica Sinica, 2010, 38(3): 646 – 653. (in Chinese)

[4] Gu Hung-yan, Tsai Sung-fung. Improving segmental GMM based voice conversion method with target frame selection [A]. International Symposium on Chinese Spoken Language Processing (ISCSLP)[C]. Singapore: IEEE, 2014. 483 – 487.

[5] Mouchtaris A, Van der Spiegel J, Mueller P. Nonparallel training for voice conversion based on a parameter adaptation approach[J]. IEEE Transactions on Audio, Speech,

and Language Processing,2006,14(3):952 – 963.
[6] Erro D,Moreno A,Bonafonte A. INCA Algorithm for training voice conversion systems from nonparallel corpora[J]. IEEE Transactions on Audio, Speech, and Language Processing,2010,18(5):944 – 953.
[7] Saito D,Watanabe S,Nakamura A,et al. Statistical voice conversion based on noisy channel model[J]. IEEE Transactions on Audio,Speech,and Language Processing,2012,20(6):1784 – 1794.
[8] 俞一彪,曾道建,姜莹. 采用独立说话人模型的语音转换[J]. 声学学报,2012,37(3):346 – 352.
Yu Yibiao,Zheng Daojiang,Jiang Ying. Voice conversion based on isolated speaker model[J]. Acta Acustica,2012,37(3):346 – 352. (in Chinese)
[9] Li Xian,Wang Zeng-fu. Frame correlation based autoregressive GMM method for voice conversion[A]. International Symposium on Chinese Spoken Language Processing (ISCSLP)[C]. Singapore:IEEE,2014. 221 – 225.
[10] Li Yan-ping,Zhang Ling-hua,Ding Hui. Nonparallel voice conversion based on phoneme classification and eigenvoices[A]. IEEE International Conference on Communication Technology (ICCT)[C]. Nanjing: IEEE, 2010. 662 – 665.
[11] Oliva G,La Manna D,Fagiolini A,et al. Distance-constrained data clustering by combined k-means algorithms and opinion dynamics filters[A]. Mediterranean Conference of Control and Automation (MED)[C]. Palermo: IEEE,2014. 612 – 619.
[12] 於跃成. 基于半监督学习的分布式和演化聚类研究[D]. 南京:南京航空航天大学,2012.
Yu Yuecheng. Distributed clustering and evolutionary clustering algorithm based on semi-supervised learning [D]. Najing: Nanjing University of Aeronautics and Astronautics,2012. (in Chinese)
[13] Michael Pitz,Hermann Ney. Vocal tract normalization equals linear transformation in cepstral space[J]. IEEE Trans on Audio,Speech,and Language Processing,2005,13(5):930 – 944.
[14] Minematsu N. Mathematical evidence of the acoustic universal structure in speech[A]. IEEE International Conference on Acoustics,Speech,and Signal Processing,(ICASSP)[C]. Philadelphia:IEEE,2005. 889 – 892.
[15] Che Yingxia,Yu Yibiao. Fast matching algorithm between statistical acoustic models of source-target speaker in structured approach of non-parallel corpora voice conversion[A]. IEEE International Conference on Information Science and Technology (ICIST)[C]. Shenzhen: IEEE, 2014. 88 – 92.

作者简介



车滢霞 女,1989 年生,江苏常州人,苏州大学电子信息学院硕士,研究方向为语音信号处理.

俞一彪(通信作者) 男,1962 年生,江苏无锡人,苏州大学电子信息学院教授,主要研究领域为语音信号处理、多媒体通信、信息隐藏.