第45卷第4期 中国科学技术大学学报 Vol. 45, No. 4
2015年4月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2015

文章编号:0253-2778(2015)04-0286-08

基于 BIC 和 G_PLDA 的说话人分离技术研究

李 锐,卓 著,李 辉

(中国科学技术大学电子科学与技术系 安徽合肥 230027)

摘要:传统的以贝叶斯信息准则(Bayesian information criterion,BIC)作为相似性度量的说话人分离技术,在短时对话的分离任务中能取得较好的效果,但是随着对话时长的增加,BIC 的单高斯模型不足以描述不同说话人数据的分布,且层次聚类(Hierarchical agglomerative clustering, HAC)时,区分相同说话人和不同说话人的门限值难以划定.针对此问题,提出基于短时 BIC 和长时 G_PLDA 的融合方法,充分利用 BIC 在短时聚类的可靠性和 G_PLDA 在长时段上的优异区分性,在美国国家标准技术局(NIST)08 Summed 测试集上的实验表明,该方法将分类错误率(DER)从 BIC 基线系统的 2.34%降到 1.54%,性能相对提升 34.2%.

关键词:说话人分离; 贝叶斯信息准则; 高斯概率线性判别分析; 分类错误率

中图分类号:TN912.34 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2015.04.005

引用格式:LI Rui,ZHUO Zhu,LI Hui. The research of speaker diarization based on BIC and G_PLDA[J]. Journal of University of Science and Technology of China, 2015,45(4):286-293.
李锐,卓著,李辉. 基于 BIC 和 G_PLDA 的说话人分离技术研究[J]. 中国科学技术大学学报,2015,45(4):286-293.

The research of speaker diarization based on BIC and G_PLDA

LI Rui, ZHUO Zhu, LI Hui

(*Department of Electronic Science and Technology, University of Science and Technology of China, Hefei* 230027, *China*)

Abstract: The traditional technology for speaker diarization(SD), which exploits the Bayesian information criterion(BIC) as the similarity metric, can obtain good results in the short dialogue task, but with the length of the dialogue increasing , single Gaussian model of BIC is insufficient to describe the information distribution of different speakers. Moveover, it is difficult to delineate the threshold between the same speakers and different speakers when using hierarchical clustering (HAC). To solve this problem, a fusion method between BIC and G_PLDA was proposed, so as to make full use of the reliability of BIC in short-term clustering and the excellent discriminating power of G_PLDA in long utterancs. A set of experiments based on NIST 08 Summed shows that this new fusion method reduces the diariazation error rate (DER) from 2.34% of BIC baseline system to 1.54%, improving performance of speaker diarization by 34.2%.

Key words: speaker diariazation; BIC; G_PLDA; DER

收稿日期:2014-11-04;**修回日期**:2014-12-13

作者简介:李锐,男,1991 年生,硕士生. 研究方向:语言信息处理. E-mail: lirui05@mail.ustc.edu.cn

通讯作者:李辉,博士/副教授. E-mail: hli@ustc.edu.cn

0　引言

随着音频处理技术的不断提高，从海量的数据中(如电话录音、新闻广播、会议录音等)获取感兴趣的特定人声已成为研究热点[1]. 另外，如何对这类音频文档进行合理有效的管理，也是目前存在的一个挑战. 美国国家标准局(NIST)从 2002 的丰富转写评测(rich transcription，RT)中正式加入了说话人分离任务[2](speaker diarization，SD)，该任务是指从多人对话中自动地将语音依据说话人进行划分，并加以标记的过程.

与传统“鸡尾酒会”形式的复杂背景下混合语音分离不同，说话人分离主要面向的是多个话者不同时发声的场景，它解决的是“什么时候由谁说”这样一个问题，而前者大多是通过盲源分离(BSS)[3]和计算听觉场景分析(CASA)[4]等方式处理. 目前的说话人分离技术主要包含两个过程：说话人分割(speaker segmentation)和话人聚类(speaker clustering). 分割的过程是指从多人对话的音频中找寻不同说话人身份转变的时间点，然后根据这些变化点可以将语音分割成若干短语音段，理想情况下，经分割后的每个短语音段只会包含一个说话人的信息. 聚类的过程则是将分割后的所有属于同一个说话人的小片段通过一些聚类的方法，再重新组合在一起.

说话人分离技术有着广泛的实际应用意义，如可以利用该技术实现电话和会议数据的自动分离及转写，将分离后的不同说话人声解码后，按敏感词检测和目标人进行抽取；为构建和检索说话人音频档案提供有效的信息. 获得的信息既可以用于音频检索；也可以用来对语音库进行自动标注和自动跟踪等；同时它也是语音识别的基础，直接影响到语音识别的精度.

1　BIC 系统

以 BIC 距离[5]作为相似性度量的 BIC 系统在 NIST04 丰富转写评测中获得了最好的分离效果，该系统主要包含了四个过程：语音端点检测(VAD)、说话人变化点检测(SCD)、基于 BIC 距离的层次聚类(HAC)以及 Viterbi 重分割.

其系统的大致框图如图 1 所示.

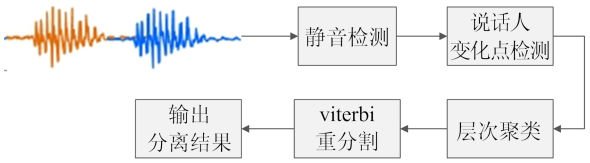

图 1　BIC 基线系统实现框图

Fig. 1　Implementation block diagram of BIC baseline system

1.1　贝叶斯信息准则

贝叶斯信息准则(BIC)作为一种常见的模型选择准则，最早是由 Chen 等[5]将其用于说话人分离. 对于模式识别中常见的模型选择问题，由贝叶斯理论可知，最佳的模型应当具有最大的后验概率，用 θ 表示模型 m_i 的参数集合，$p(\theta \mid m_i)$ 表示 θ 的先验分布，那么边缘似然函数 $p(X \mid m_i)$ 可以表示成为：

$$p(X \mid m_i) = \int p(X, \theta \mid m_i)\mathrm{d}\theta = \int p(X \mid \theta, m_i)p(\theta \mid m_i)\mathrm{d}\theta \quad (1)$$

边缘似然函数有一个很好的性质，即能够对自由参数个数多的模型进行惩罚，这样依据最大的边缘似然选择的模型就不至于太复杂或太简单，而是恰好能够描述观察到的数据. 通过拉普拉斯变换，可以找到边缘似然概率的近似值，称之为 BIC，其定义如下：

$$\mathrm{BIC} = \log P(X \mid \hat{\theta}, m) - \frac{1}{2}\lambda L \log N \quad (2)$$

式中，L 为模型的自由参数个数，N 为样本数，$\hat{\theta}$ 是模型参数集合的最大后验估计，λ 为模型复杂度的惩罚系数.

1.2　说话人分割

传统的说话人分割算法有 BIC 距离准则，交叉似然比(CLR)，KL 距离等，本文使用基于变窗长的 BIC 距离判决进行说话人变化点检测.

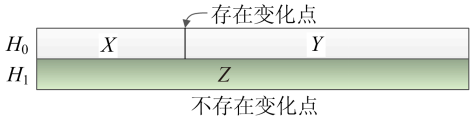

图 2　一个窗内的变化点检测图解

Fig. 2　Diagram of change point detection in a window

对给出的一段语音中是否存在变化点这一假设进行检验：

H_0:说话人在一个窗内的 t 时刻发生了跳变;

H_1:说话人在一个窗内都没有跳变;

则,

$$\Delta\mathrm{BIC} = \mathrm{BIC}(H_0) - \mathrm{BIC}(H_1)$$
$$= Nz\log|\boldsymbol{\Sigma}z| - Nx\log|\boldsymbol{\Sigma}x| - Ny\log|\boldsymbol{\Sigma}y| - \Gamma \quad (3)$$

式中,$\Gamma = \frac{1}{2}\lambda(k + \frac{k(k+1)}{2})\log N$ 为模型惩罚项,$\boldsymbol{\Sigma}$ 为协方差矩阵,N 为每个小段的样本数,即语音帧数,k 为特征参数的维度.

在一个滑动窗中,每 100 帧作为一个假想变化点时,计算窗两端的 ΔBIC 值,若 ΔBIC>0,则认为此处存在变化点,然后将起始窗移动到此位置,继续检测下一个变化点. 否则增大窗长,窗的起点不变,继续检测可能存在的变化点.

1.3 说话人聚类

传统的聚类方法有已知类别个数的 K-means 聚类以及自底向上的层次聚类(HAC)[6],HAC 是目前很多分离系统中使用较多的方法,其聚类如图 3 所示.

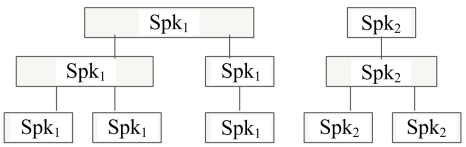

图 3 层次聚类图示

Fig. 3 Diagram of hierarchical agglomerative clustering

通过构建距离矩阵,不断计算两个片段的 BIC 距离,搜索两个 BIC 距离最小的片段进行合并,每次合并时算法的复杂度为:

$$O(n) = \frac{(n-k+1)(n-k)}{2} \quad (4)$$

式中,n 为 BIC 分割后的初始段个数,k 为层次聚类的迭代次数. 由式(4)可以看出,随着初始段 n 的增多,层次聚类算法的复杂度快速增长. 一种优化的策略是在每次合并相似片段时,只更新合并后的段与其他片段之间的距离,在减少运算的同时,不影响实验结果.

1.4 Viterbi 重分割

理想情况下,BIC 分割后的每个片段应当只包含一个说话人. 实际上,分割后的小片段会含有其他说话人的语音,因为基于 BIC 距离的说话人分割是一种完全无先验信息的模型检验,所以目前的说话人分离系统几乎都加入了基于 Viterbi 解码的重估过程.

通过将层次聚类得到的聚类结果作为包含说话人信息的先验知识,构建多状态的隐马尔科夫模型,每个 HMM 的状态实际上是一个 GMM 模型,代表一个说话人. 将经过 BIC 分割后的说话人语音段以帧级别的形式进行解码,获取最优变化序列,即新的说话人分割点边界,然后用这些新的数据重训 GMM 模型,迭代更新 HMM 模型参数,经过若干次迭代后,即可得到更加精确的说话人转换边界.

2 高斯概率线性判别分析(G_PLDA)

2.1 总变化因子(I-vector)

近年来,基于 TV 空间(total variability space)的总变化因子[7]技术在声纹识别中使系统性能获得显著提升. 文献[8]将该技术应用于说话人分离,通过提取每个分割段上的 I-vector,PCA 变换后,再进行 K-means 聚类,获得了不错的分离效果. 引入总变化因子实质上包涵了因子分析的思想,它在 GMM 均值超矢量空间上,将说话人部分和信道部分放在一起建模,不再区分说话人空间和信道空间的差异. 通过该技术,经 BIC 分割后的时长不同的语音段被映射成了固定维度的矢量,与此同时尽可能地保留了说话人信息. 这样在低维空间中,我们可以更方便地对固定长度的矢量进行处理.

给定一段语音,与说话人及信道相关的 GMM 均值超向量 $\boldsymbol{M}$ 由下式表示:

$$\boldsymbol{M}_{s,h} = \boldsymbol{m}_u + \boldsymbol{T}\boldsymbol{\omega}_{s,h} \quad (5)$$

式中,$\boldsymbol{m}$ 为通用背景模型 UBM 的均值超矢量,$\boldsymbol{T}$ 是一个低秩矩阵,代表总变化空间,$\boldsymbol{\omega}$ 因子是与说话人及信道相关的总变化因子,也就是低维矢量 I-vector,服从均值为 0,方差为 I 的高斯分布.

TV 空间的训练以及 I-vector 的提取过程参考文献[9],通过直接计算两个 I-vector 矢量的余弦相似度,可以很方便地在(−1,1)之间划分相似度门限,当余弦距离大于某个设定的阈值时,认为这两个语音段是属于同一个人发出,进而进行聚类. 两个 BIC 分割段提取出的 I-vector 余弦距离相似度计算公式如下:

$$S(\omega\mathrm{Spk}A,\omega\mathrm{Spk}B)=\frac{\langle\omega\mathrm{Spk}A,\omega\mathrm{Spk}B\rangle}{\|\omega\mathrm{Spk}A\|\cdot\|\omega\mathrm{Spk}B\|} \tag{6}$$

TV 系统将所有可变因子视为一个总的变化因子，简化了建模过程，但是由于该模型不区分说话人和信道差异，ω 因子中会包括除说话人外的其他扰动因子，这些扰动因子会影响其在 T 空间的分布，为了有效抑制扰动因子的影响，可以在 I-vector 层面上进行 PCA、LDA＋WCCN 或 PLDA[10] 处理，以减弱扰动因子的影响.

2.2　G_PLDA 模型

忽略 I-vector 的提取机制，PLDA 把它看成是由一种生成式模型产生的声学特征. 在 PLDA 的框架下，I-vector 的产生过程可以用一个隐藏变量来描述，不同的隐藏变量数目，不同的先验假设构成了不同的 PLDA 模型.

最常用的简化后的 PLDA 模型如下：

$$D_{s,r}=\mu+Vy_s+\varepsilon_{s,r} \tag{7}$$

假设对于说话人 s 的第 r 句语音，提取出的 I-vector 为 $D_{s,r}$，μ 表示所有说话人的 I-vector 的均值，对应的说话人因子为 y_s，残差项为 ε. 假设隐含因子的先验分布符合学生分布（t 分布）：

$$y_s\sim N(0,v_{y,s}^{-1}I);$$

$$\varepsilon_{s,r}\sim N(0,v_{\varepsilon,r}^{-1}\Sigma).$$

此时的 PLDA 模型称为 HT_PLDA[11]，重写公式(7)得

$$p(D_{s,r})\sim N(\mu,\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{\Sigma}).$$

相应的条件概率公式为：

$$p(D_{s,r}\mid \boldsymbol{y}_s,\boldsymbol{v}_{y,s},\boldsymbol{v}_{\varepsilon,r})\sim N(\mu+\boldsymbol{V}y_s,\boldsymbol{v}_{\varepsilon,s}^{-1}\boldsymbol{\Sigma}).$$

式中，$\boldsymbol{v}_{y,s}$ 和 $\boldsymbol{v}_{\varepsilon,r}$ 称为学生分布的自由度，服从 Gamma 分布.

在与文本无关的话者确认系统中，由 HT_PLDA 模型得出的等误识率 EER 和 MinDCF 确实优于传统的高斯分布假设. 在实时率要求较高的电话信道下的双人分离任务中，由于该模型的计算复杂度较高以及为了保持和提取出的 I-vector 为高斯分布这一假设相一致，本文使用了 G_PLDA 对提取出的 I-vector 进行建模.

同样在模型(7)下，假设隐含因子的先验分布服从如下的高斯分布，且统计独立：

$$\varepsilon_{s,r}\sim N(0,\Sigma),\ y_s\sim N(0,I).$$

此时的 PLDA 模型称为 G_PLDA[12]，重写公式(7)得

$$p(D_{s,r})\sim N(\mu,\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{\Sigma}).$$

相应的条件概率公式为：

$$p(D_{s,r}\mid y_s)\sim N(\mu+\boldsymbol{V}y_s,\boldsymbol{\Sigma}).$$

2.3　G_PLDA 的参数训练

令模型参数 $\lambda=(\mu,\boldsymbol{V},\boldsymbol{\Sigma})$，训练集合为 D，由贝叶斯公式可得

$$p(\lambda\mid D)=\frac{p(D\mid\lambda)p(\lambda)}{p(D)} \tag{8}$$

目标是使得在所有的训练数据上的似然度最大：$\mathrm{argmax}p(\lambda\mid D)\propto p(D\mid\lambda)$，模型的参数可以由 EM 算法进行估计. 初始化模型参数 λ，在 E 步，估计出隐含因子 ys 的后验分布：

$$\boldsymbol{E}[y_s]=(J\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}+\boldsymbol{I})^{-1}\boldsymbol{V}^{\mathrm{T}}\sum_{h=1}^{J}(D_{s,h}-\mu) \tag{9}$$

$$\boldsymbol{E}[y_s y_s^{\mathrm{T}}]=(^{J}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}+\boldsymbol{I})^{-1}+\boldsymbol{E}[y_s]\boldsymbol{E}[y_s]^{\mathrm{T}} \tag{10}$$

M 步，利用上述求得的后验均值和方差更新模型参数：

$$\boldsymbol{V}=\Big(\sum_{S}\sum_{r}(D_{s,r}-\mu)\boldsymbol{E}[y_s]^{\mathrm{T}}\Big)\Big(\sum_{S}\sum_{r}\boldsymbol{E}[y_s]\boldsymbol{E}[y_s]^{\mathrm{T}}\Big)^{-1};$$

$$\boldsymbol{\Sigma}=\frac{1}{IJ}\sum_{S}\sum_{r}(D_{s,r}-\mu)(D_{s,r}-\mu)^{\mathrm{T}}-\boldsymbol{V}\boldsymbol{E}[ys](D_{s,r}-\mu)^{\mathrm{T}}.$$

式中，I 表示训练数据的所有说话人个数，J 表示每个说话人的句子总数.

为了利用 G_PLDA 对提取出的 I-vector 建模，必须将非高斯分布的 I-vector 进行非线性变换，以减弱这种非高斯的影响，常见的做法是对提取出的 I-vector 进行白化处理以及长度规整[13].

2.4　G_PLDA 的得分计算

存在两个语音段 X，Y，提取出的 I-vector 分别为 D_1、D_2，存在以下两个假设：

H_s：D_1、D_2 来自同一人，他们具有相同的隐含

因子 y_s，

H_d：D_1、D_2来自不同人，他们分别具有隐含因子 y_{s1}, y_{s2}.

对于上述假设可以通过对数似然得分进行验证：

$$\text{score} = \ln \frac{p(D_1, D_2 \mid H_s)}{p(D_1, D_2 \mid H_d)} = \ln \frac{p(D_1, D_2 \mid H_s)}{p(D_1 \mid H_d)p(D_2 \mid H_d)} \tag{11}$$

在 G_PLDA 模型下，由于边缘似然函数的分布具有高斯形式，直接从 I-vector 的分布出发，采用如下的近似可以避免较多的隐藏变量计算：

从 G_PLDA 的定义公式(7)出发，如果 D_1，D_2来自同一个人，则有：

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = \left(\begin{bmatrix} \mu \\ \mu \end{bmatrix} + \begin{bmatrix} \boldsymbol{V} & \\ & \boldsymbol{V} \end{bmatrix} \begin{bmatrix} y_s \\ y_s \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \right);$$

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}\text{tot} & \Sigma ac \\ \boldsymbol{\Sigma} ac & \boldsymbol{\Sigma}\text{tot} \end{bmatrix} \right).$$

如果 D_1，D_2 来自不同的人，则有：

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = \left(\begin{bmatrix} \mu \\ \mu \end{bmatrix} + \begin{bmatrix} \boldsymbol{V} & \\ & \boldsymbol{V} \end{bmatrix} \begin{bmatrix} y_{s1} \\ y_{s2} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \right);$$

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}\text{tot} & \\ & \boldsymbol{\Sigma}\text{tot} \end{bmatrix} \right).$$

式中，$\boldsymbol{\Sigma}\text{tot} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{\Sigma}$，$\boldsymbol{\Sigma} ac = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$，因此似然度得分公式(11)变成：

$$\text{score} = \log N \left(\begin{bmatrix} D_1 \\ D_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}\text{tot} & \boldsymbol{\Sigma} ac \\ \boldsymbol{\Sigma} ac & \boldsymbol{\Sigma}\text{tot} \end{bmatrix} \right) - \log N \left(\begin{bmatrix} D_1 \\ D_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}\text{tot} & \\ & \boldsymbol{\Sigma}\text{tot} \end{bmatrix} \right) \tag{12}$$

$$\text{score} \propto (D_1 - \mu)^{\mathrm{T}} Q (D_1 - \mu) + (D_2 - \mu)^{\mathrm{T}} Q (D_2 - \mu) + 2(D_1 - \mu)^{\mathrm{T}} P (D_2 - \mu)$$

式中，

$$Q = \boldsymbol{\Sigma}\,\text{tot}^{-1} - (\boldsymbol{\Sigma}\text{tot} - \boldsymbol{\Sigma} ac \boldsymbol{\Sigma}\,\text{tot}^{-1} \boldsymbol{\Sigma} ac)^{-1};$$

$$P = \boldsymbol{\Sigma}\,\text{tot}^{-1} \boldsymbol{\Sigma} ac\, (\boldsymbol{\Sigma}\text{tot} - \boldsymbol{\Sigma} ac \boldsymbol{\Sigma}\,\text{tot}^{-1} \boldsymbol{\Sigma} ac)^{-1}.$$

由于上述公式中的 Q、P 只与 V 和 Σ 有关，所以可以在训练完 G_PLDA 模型后就直接计算出来，这样在每次判断两个说话人的 G_PLDA 得分时，不再重新计算 Q、P 的值，大大简化了计算得分的时间.

2.5 G_PLDA 得分分布

为了利用 G_PLDA 进行长时段上的聚类分析，我们通过大量的统计结果，获得了 UBM 混合度为 256，TV 因子数为 200，G_PLDA 说话人因子数为 100 时，相同人和不同人在不同时长上的 G_PLDA 得分分布，结果如图 4 所示.

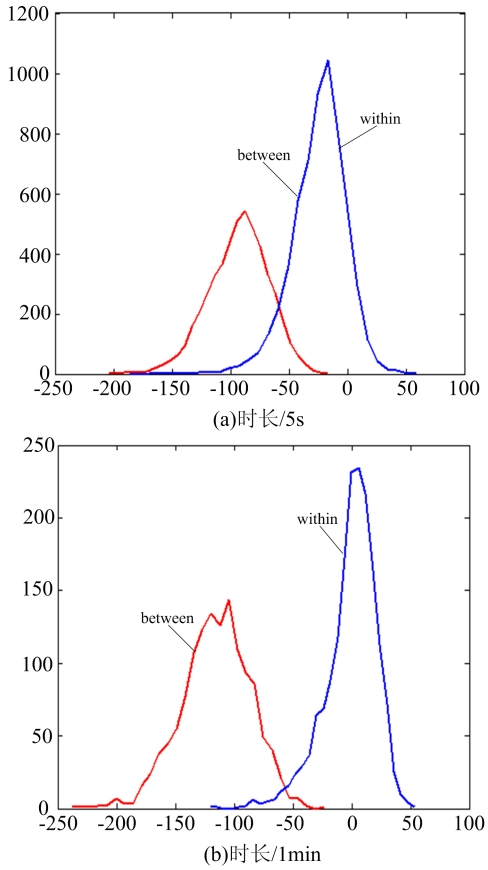

图 4 不同时长 G_PLDA 得分分布图示

Fig. 4 Diagram of G_PLDA score in different duration

从图 4 可以明显看出，当时长大于 5s 时，G_PLDA就已经显示出良好的区分性，当时长大于 1min 时，几乎能完全区分两个不同人的语音. 另外从图中我们发现，随着时长的增加，相同人和不同人的得分分布相交的位置波动很小，这就意味着在进行说话人聚类时，可以很方便地设置聚类门限，而不必考虑门限值波动带来的影响.

3 BIC 与 G_PLDA 融合系统

BIC 距离作为相似性度量的说话人聚类过程，在短时语音段上通过单高斯就有很好的描述能力，但是随着层次聚类不断进行，数据的时长也会增加，仅仅依靠单高斯不足以对相同人和不同人的数据分布进行描述，而且以传统的 BIC 距离进行相似性判断，BIC 值会随着数据长度 N 的增加而迅速变化，相应的区分门限很难划定.

基于上述问题，本文提出了一种融合短时 BIC 和

长时 G_PLDA 的说话人分离方法，即充分利用 BIC 在短时聚类的可靠性和 G_PLDA 在长时段上的优异区分性，进行多人分离测试，系统框架如图 5 所示.

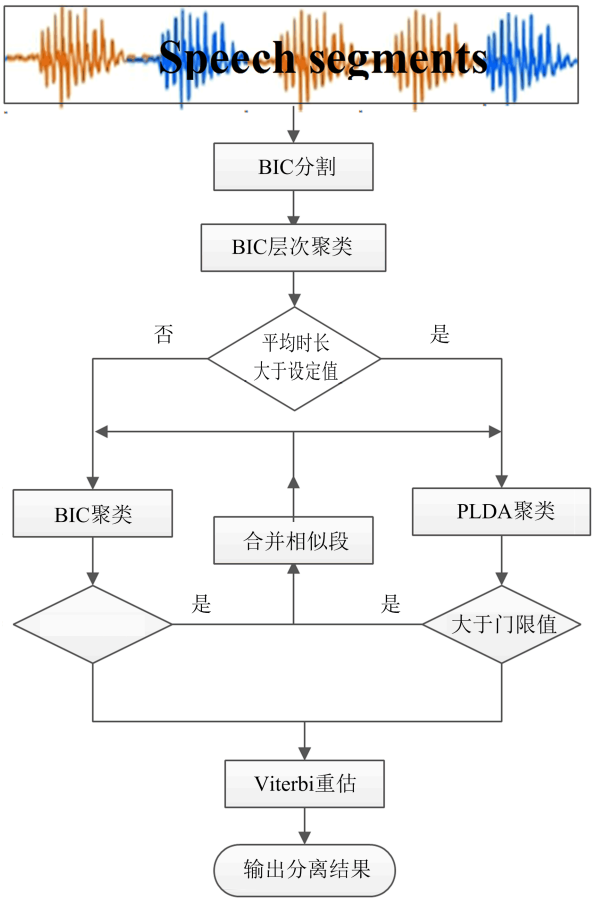

图 5　融合 BIC 和 G_PLDA 系统框图

Fig. 5　Fusion system based on BIC and G_PLDA

给定一条语音，分帧后提取 13 维 PLP 特征，通过 VAD 获取端点信息，然后在 VAD 段内利用 BIC 分割获得说话人转折点信息；在初始的分割段片段上进行 BIC 层次聚类，聚类到一定程度后，如果片段的平均时长超过设定的阈值，则利用 G_PLDA 进行长时聚类，否则继续进行 BIC 层次聚类，当不满足合并条件时，停止聚类，并输出分离结果.

4　实验结果及分析

4.1　性能评价指标

目前 NIST 举行的 RT 丰富转写评测中，说话人分离系统都是以分类错误率(DER)来作为衡量指标. 其定义如下：

$$DER = Miss + False + SpkERR \quad (13)$$

式中，Miss 指将语音段当作静音段处理，丢失的有效时长占实际有效语音段的百分比. False 指将静音段当作语音段处理，多余的错误时长占实际有效语音段的百分比. SpkErr 指将一个说话人的语音段归于另一个说话人的百分比.

本文另外还加入了一个新的指标来辅助判断分离的效果，叫大类错误率，指的是当前分离系统下，分类错误率达到 10%以上的语音个数占总语音个数的百分比.

4.2　训练集和测试集

本实验中，训练 UBM、TV 的数据来自 NIST04、NIST05、NIST06，大约 500 h 的音频数据，训练 G_PLDA 的数据来自 NIST04、NIST05、NIST06 共 577 个说话人，平均每个人约有 15 句话.

测试数据来自 NIST 08 Summed 电话信道数据集，总共有 2 212 条双人对话语音，每条时长约 5min. 为了消除静音检测对系统性能的影响，本文直接使用 NIST 提供的 ASR 识别结果转换得到的 VAD 标注信息，因此 DER 中的 Miss 和 False 都是 0%，只关心 SpkErr.

4.3　基线系统以及融合系统的分离结果

表 1 列出了基线系统、G_PLDA 单系统以及融合 BIC 和 G_PLDA 后最好的分离效果. 图 6 显示了不同模型下的融合系统分离效果.

表 1　不同系统下的分离结果

Tab. 1　Results ofspeaker diarization in different systems

测试系统	SpkErr (%)	大类错误率(%)	实时率 (倍数)
BIC 基线系统	2.34	5.7	140
G_PLDA 单系统	3.55	9.8	8
融合 BIC+G_PLDA 系统	1.54	2.9	65

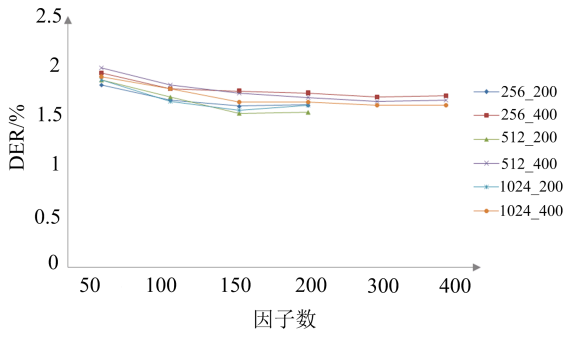

图 6　不同模型下的融合系统分离效果

Fig. 6　Results of fusion system under different model

4.4 实验结果分析

(Ⅰ)从表1的对比结果中可以明显看出,融合后的系统相对于基线BIC系统,SpkErr有34.2%的下降,大类错误率也有49.1%的下降,说明基于图5的分离策略是有效果的.另外从表1中可以看出,G_PLDA单系统用于说话人分离,性能与基线BIC系统相比有所下降,应该是由于刚开始聚类时,每个小的说话人片段时长都很短,包含的说话人信息有限,在这种短时的情况下提取I-vector,会存在比较大的误差,进而通过G_PLDA对提取出的不准确的I-vector进行得分判断,误差会越来越大.另外,由于本文采用的是自底向上的层次聚类方法,它是一种不可逆过程,一旦出现聚类错误,会一直保持到最终结果,所以才会出现表1中G_PLDA单系统效果较差的现象.

(Ⅱ)对比表1中不同系统的实时率可以发现,融合后的系统相对于基线系统,性能大概有1倍的下降,因为在BIC分组聚类后需要对每个分组中的小段提取一次I-vector,并计算G_PLDA得分,在处理离线音频的情况下,1倍实时率的下降换取34%的效果提升,还是可以接受的,而G_PLDA单系统由于从一开始就在每个BIC分割段上提取I-vector,然后每次层次聚类后合并相似段时又提取了一次I-vector,所以效率会变得十分低下,文献[6]提出的基于短时I-vector+PCA+K-means的方法虽然能取得比本文更好的分离效果,但是其实时率很难保障.

(Ⅲ)从图6可发现,不同的UBM混合度、TV因子数以及G_PLDA中说话人因子数导致的分离效果是不同的.在同等的TV因子数下,随着UBM混合度的增加,分离效果有时反而变差,同样,在UBM混合度相同的情况下,TV因子数的增加也会导致分离效果变差.按照常理,随着UBM混合度的增加以及TV因子数的增加,对数据的分布描述得会更精细,分离效果应该会有所提升,但是现在反而没有达到预期的效果.出现这种情况原因可能是由于本论文是基于电话信道下双人对话的分离,电话情况下的数据量本身就不多,如果对数据的描述精细到一定程度,反而会出现过拟合的现象,因此并不是UBM的混合度以及TV因子数越高越好.实验表明,在UBM混合度为512,TV因子数为200时,系统的分离效果能达到最佳.

(Ⅳ)对比图6的横坐标,在同等的UBM混合度和TV因子数的情况下,不同的G_PLDA说话人因子数对分离效果的影响也不同,总体来看,随着说话人因子数的不断增加,分离的效果越来越好.由于G_PLDA可以看作对声学特征的一种二次降维,在假定I-vector对数据分布的描述合理的情况下,因子数越多,包含的信息量就越丰富,可用于区分相同人和不同人的信息就越多,但是当因子数增大到一定程度后,相同说话人中的一些差异信息也会体现出来,所以出现图6所示的分类错误率略微上升的现象.

5 结论

本文尝试基于短时BIC和长时G_PLDA融合的方法,在传统的以BIC距离为聚类准则的基线系统上,加入了区分性更好的G_PLDA模型,依据大量的数据统计出似然度得分分布,很方便地得出了层次聚类时相似度判决门限.实验结果表明,这种融合策略能有效提升说话人分离效果.

实际应用中,说话人分离技术面临着各种挑战,例如,如何正确地估计说话人的数目、如何尽量摒除背景噪声的干扰以及如何有效地检测出重叠音,都关系到分离系统的整体性能好坏.虽然有文献提出谱聚类、E-HMM模型等能适当地估计出说话人数目,但是当个别说话人发音较少,想要准确地估计出说话人数目还存在一定的困难.与此同时,在复杂背景下,基于模型的重叠音位置很难准确估计,且其混合程度也无法判别,这些都将是接下来需要研究的重点.

参考文献(References)

[1] Moattar M H, Homayounpour M M. A review on speaker diarization systems and approaches[J]. Speech Communication, 2012, 54(10): 1065-1103.

[2] TranterS E, Reynolds D A. An overview of automatic speaker diarization systems[J]. IEEE Transactions on Audio, Speech, and Language Processing, , 2006, 14(5): 1557-1565.

[3] Makino S, Lee T W, Sawada H. Blind Speech Separation[M]. Berlin, Germany: Springer, 2007.

[4] Wang D L, Brown G J. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications [M]. New Jersey, USA: Wiley, 2006.

[5] Chen S S, Gopalakrishnan P S. Speaker, environment and channel change detection and clustering via the Bayesian information criterion[C]// Proceedings of the

DARPA Broadcast News Transcription and Understanding Workshop. Morgan Kaufman, 1998: 127-132.

[6] Ben M, Betser M, Bimbot F, et al. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs [C]// Proceedings of the International Conference on Spoken Language Processing. Jeju, Korea: IEEE Press, 2004: 2329-2332.

[7] Dehak N, Kenny P, Dehak R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788-798.

[8] Shum S, Dehak N, Chuangsuwanich E, et al. Exploiting Intra-Conversation Variability for Speaker Diarization [C]// Proceedings of the 11th Annual International Speech Communication Association. Florence, Italy: IEEE Press, 2011: 945-948.

[9] GlembekO, Burget L, Matějka P, et al. Simplification and optimization of i-vector extraction [C]// International Conference on Acoustics, Speech and Signal Processing. Brno, Czech: IEEE Press, 2011: 4516-4519.

[10] Prince S J D, Elder J E. Probabilistic linear discriminant analysis for inferences about identity[C]// 11th International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE Press, 2007: 1-8.

[11] Kenny P. Bayesian speaker verification with heavy-tailed priors[C]// Proceedings of the Odyssey Speeker and Language Recognition Workshop. Brno, Czech Republic: IEEE Press, 2010: 14.

[12] Kenny P, Stafylakis T, Ouellet P, et al. PLDA for speaker verification with utterances of arbitrary duration[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE Press, 2013: 7649-7653.

[13] Garcia-Romero D, Espy-Wilson Y. Analysis of I-vector length normalization in speaker recognition systems [C]// Proceedings of the 11th Annual International Speech Communication Association. Florence, Italy: IEEE Press, 2011: 249-252.