ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 1,Jan. 2017,pp. 54－61
DOI:10.16337/j.1004-9037.2017.01.006

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于变分贝叶斯改进的说话人聚类算法

李敬阳[1]　李　锐[2]　王　莉[1]　王晓笛[1]

(1.公安部物证鉴定中心,北京,100038;2.中国科学技术大学电子科学与技术系,合肥,230027)

摘　要:说话人聚类是说话人分离中的一个重要过程,然而传统的以贝叶斯信息准则作为距离测度的层次聚类方式,会出现聚类误差向上传递的情况。本文提出了一种逐级算法增强处理机制。当片段之间的最小贝叶斯信息准则距离超过设定的门限值时,或者类别个数到达一定程度时,将当前聚类结果作为初始类中心,通过变分贝叶斯迭代法重新对每个类别中的片段调优,最后再依据概率线性判别分析得分门限确定说话人个数。实验表明,本文方法在美国国家标准技术署08 summed测试集上,使得"类纯度"和"说话人纯度"比传统算法都有了一定提升,且使得说话人分离整体性能相对提升了27.6%。

关键词:说话人聚类;贝叶斯信息准则;概率线性判别分析;变分贝叶斯

中图分类号:TN912.34　　**文献标志码**:A

Improved Algorithm of Speaker Clustering Based on Variation Bayesian

Li JingYang[1],Li Rui[2],Wang Li[1],Wang Xiaodi[1]

(1. Institute of Forensic Science, Ministry of Public Security, Beijing,100038, China;2. Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, 230027, China)

Abstract: The speaker clustering is an important process of speaker diarization, yet traditional method for hierarchical agglomerative clustering (HAC) with distance measurement based on Bayesian information criterion (BIC) can lead to the clustering error propagation. To solve this problem, step by step algorithm is proposed, when the minimum BIC distance between segments exceeds a predefined threshold, or the number of the categories on hierarchical clustering reaches a certain number. The current clustering result as the initial class center, and then variational Bayesian method will be exploited to tune the speaker segments among the categories iteratively. Finally, the number of speaker is determined according to the probabilistic linear discriminant analysis (PLDA) score threshold. Experiments on national institute of standards and technology (NIST) 08 summed test set show that this method improves the "class purity" and "speaker purity" compared with conventional algorithms. Moreover, performance of speaker diarization is relatively improved by 27.6%.

Key words: speaker clustering; Bayesian information criterion; probabilistic linear discriminant analysis; variational Bayesian

---

基金项目:公安部应用创新计划(2014YYCXGAES048)资助项目。
收稿日期:2015-05-15;**修订日期**:2015-07-14

引　言

随着信息处理技术的不断提升、互联网的普及,人们获取各种音频的途径越来越广泛,然而在音频数据爆炸式增长的同时,如何合理有效地管理和存储这些海量数据是迫切需要解决的问题[1]。传统的基于文本形式的音频检索方式已经满足不了人们对海量数据的检索需求,采用人工标注的方式不但成本昂贵、效率低,而且很容易加入个人的主观色彩。于是,基于内容形式的音频检索应运而生,并成为多媒体研究领域的热点问题。然而实际上大多数的语音信号不仅仅包含文本信息,也同样包含说话人信息,这种基于内容的处理方法会使说话人的信息丢失,存在一定的缺陷。为此,文献[2]使用说话人分离技术,构造和建立说话人索引,为在更高语义层次上实现音频检索提供基础。说话人聚类是说话人分离技术中的一个重要环节,其关注的是如何将杂乱无序的说话人片段通过一种无监督的聚类方式,自动地组合在一起。理想情况下,聚类后每个类别的片段仅属于同一个说话人,而不同类别中的片段属于不同的说话人。说话人聚类在语音识别和电话会议转写中有着广泛的应用价值,通过说话人聚类技术将相同的说话人语音段聚为一类,可以为说话人自适应提供更可靠的说话人模型,最终提升语音识别和电话会议转写的准确度。

现有的说话人聚类多采用基于距离准则的层次聚类方法[3],可选的相似性度量准则包括贝叶斯信息准则、相对熵、广义似然比、归一化似然比以及信息瓶颈等。其中贝叶斯信息准则(Bayesian information criterion,BIC)距离准则最早是由 Chen 在文献[4]中将其用于说话人分割和聚类,但是随着聚类时长的增加,BIC 的单高斯模型不足以描述说话人数据的分布,而基于通用背景模型(Universal background model,UBM)和最大后验估计(Maximum a posterior,MAP)的交叉似然比,在说话人片段时长足够的情况下能够得到较好的结果[5]。伯克利大学的 Yella 等[6]利用信息瓶颈(Information bottleneck,IB)作为一种准则,在多人会议的分离中也取得了一定效果。文献[7]在已知说话人个数的情况下引入了总变化因子(I-vector)技术,通过 K 均值聚类的方式对每个说话人片段进行调优,在电话信道下的双人对话中获得了很好的效果。然而真实情况下说话人的个数是未知的,K-means 的局限性显而易见。本文针对传统的说话人聚类算法会出现聚类误差向上传递的情况,提出了改进策略,利用 VB-I-vector 的软聚类方式,对说话人片段重新调优,并结合短时 BIC 和长时 PLDA 的优异区分性实现说话人聚类。实验表明,在 Nist-08 summed 数据集上,无论是聚类的类纯度还是说话人纯度都有了一定提升。

1　BIC+PLDA 基线系统

基线采用的是基于短时 BIC 和长时概率线性制别分析(Probabilistic linear discriminant analysis,PLDA)融合的方法[8],充分利用了 BIC 的单高斯对短时说话人片段的描述能力和 PLDA[9]对长时片段的区分性优势,其流程如图 1 所示。

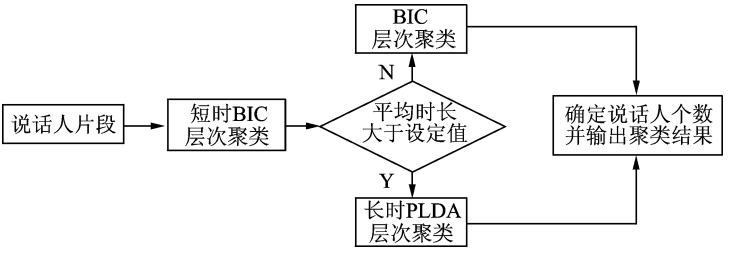

图 1　基线系统实现框图

Fig. 1　Implementation block diagram of baseline system

给定分割后的说话人片段，通过自底向上的 BIC 层次聚类方式，将可能属于同一个人的片段不断地合并在一起，同时检查合并后段的平均时长是否大于设定值，若满足条件则利用区分性更强的 PLDA 模型对每个说话人片段建模，并计算两两之间的 PLDA 得分，否则继续 BIC 层次聚类。其中说话人聚类的最终停止门限可以由大量实验数据统计出的相同人和不同人之间的 PLDA 得分分布获得。

1.1 贝叶斯信息准则

贝叶斯信息准则作为一种常见的模型选择准则，是对边缘似然函数的一种拉普拉斯近似，其定义为

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

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，为模型复杂度的惩罚系数。

给定两个说话人片段为 $\boldsymbol{X}$，$\boldsymbol{Y}$，存在下面两个假设：

(1) H_s：片段 $\boldsymbol{X}$，$\boldsymbol{Y}$ 由同一个说话人发出；(2) H_d：片段 $\boldsymbol{X}$，$\boldsymbol{Y}$ 由不同的说话人发出，则两个说话人片段之间的 BIC 距离为

$$\begin{aligned}\Delta\mathrm{BIC}&=\mathrm{BIC}(H_d)-\mathrm{BIC}(H_s)=\\&N\log|\boldsymbol{\Sigma}|-Nx\log|\boldsymbol{\Sigma}_x|-Ny\log|\boldsymbol{\Sigma}_y|-\Gamma\end{aligned} \tag{2}$$

式中：$\Gamma=\frac{1}{2}\lambda\left(k+\frac{k(k+1)}{2}\right)\log N$ 为模型惩罚项；$\boldsymbol{\Sigma}$ 为协方差矩阵；N 为每个片段的帧数；k 为特征参数的维度。从式(2)可以发现，BIC 的值会随着数据时长的增加而迅速变化，层次聚类到最后确定说话人个数的得分门限就很难划定。为此引入声纹识别中的 I-vector-PLDA 区分性模型，将帧数不定的说话人片段映射为长度固定的低维矢量，与仅仅使用 BIC 距离作为度量准则相比，极大地提升了说话人聚类时的类纯度和说话人纯度。而且由于两个说话人片段的 PLDA 得分几乎不受时长的影响，因此能够很方便地划出门限确定最终的说话人数目。

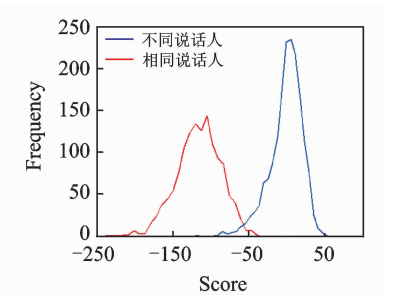

图 2 时长 1 min PLDA 得分分布

Fig. 2 Diagram of PLDA score between 1 min

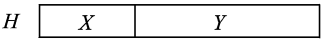

图 3 BIC 距离准则图解

Fig. 3 Diagram of Bayesian information criterion

1.2 概率线性判别分析

I-vector-PLDA 是建立在高斯混合模型(Gaussian mixture model，GMM)均值超矢量空间上的一套框架[10]。给定一段语音，与说话人及信道相关的 GMM 均值超向量 $\boldsymbol{M}$ 可表示为

$$\boldsymbol{M}=\boldsymbol{m}+\boldsymbol{T}\boldsymbol{\omega} \tag{3}$$

式中：$\boldsymbol{m}$ 为 UBM 均值超矢量；$\boldsymbol{T}$ 表示总变化空间(不再区分说话人子空间和信道子空间)，$\boldsymbol{\omega}$ 为总变化因子，也就是最终得到的低维矢量 I-vector，服从均值为 0，方差为 I 的高斯分布。总变化空间的建立使得均值超向量由维数非常小的隐变量 $\boldsymbol{\omega}$ 决定，这样对说话人建模只需要估计 $\boldsymbol{\omega}$，极大地减少了需要估计参数的个数，在注册数据不足时，能得到较为准确的说话人 GMM 模型。

在 PLDA 框架下，I-vector 被看成是由一种生成式模型产生的声学特征，其产生过程可以用一个隐藏变量来描述，而不同的隐藏变量数目和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 模型为

$$D_{s,r}=\mu+Vy_s+\varepsilon_{s,r} \tag{4}$$

式中：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 均值，与其对应的说话人因子为 y_s，残差项为 ε。若隐含因子的先验分布服从如下的高斯分布，则此时的 PLDA 模型称为高斯 PLDA[11]，即有

$$y_s\sim N(0,I)$$

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

同 BIC 距离准则类似,PLDA 采用似然比得分形式,对于两个说话人片段 $\boldsymbol{X}$,$\boldsymbol{Y}$,分别提取出 I-vector 为 D_1,D_2,存在以下两个假设:

H_s:D_1,D_2 来自同一人,它们具有相同的隐含因子 y_s;

H_d:D_1,D_2 来自不同人,它们分别具有隐含因子 y_{s1},y_{s2},则 PLDA 似然比得分公式为

$$\text{score}=\ln\frac{p(D_1,D_2\mid H_s)}{p(D_1,D_2\mid H_d)}=\ln\frac{p(D_1,D_2\mid H_s)}{p(D_1\mid H_d)p(D_2\mid H_d)} \tag{5}$$

2 变分贝叶斯调优系统

基于说话人聚类基线搭建的完整的说话人分离系统虽然在场景较为简单的双人对话中能够达到完全实用的水平,但是在复杂场景如背景音较强、对话中含笑声、重叠音以及多人参与的对话中,表现得却差强人意。其根本原因在于实际的聚类系统,初始时每个说话人片段的类纯度就不能够得到保证;每个片段大多在 1~2 s,包含的可用信息太少;基于 BIC 距离准则的层次聚类方式本质上是一种贪心算法,它并不能保证全局最优,一旦出现聚类误差,会一直地向上传递,并保持到最终结果;且 BIC 的单高斯模型在短时上更偏向于文本信息,而非说话人信息。

为了使系统的聚类效果有所提升,本文在原有系统的基础上,提出了逐级算法处理机制。首先当 BIC 层次聚类的类别数到达一个预先设定的值时,或者 BIC 距离超过设定的门限值时,通过 VB 对所有的说话人片段全局调优,再将属于一个类别的所有片段作为注册数据,提取一个 I-vector,最后再 PLDA 层次聚类,依据得分门限确定说话人个数。其框图如图 3 所示。

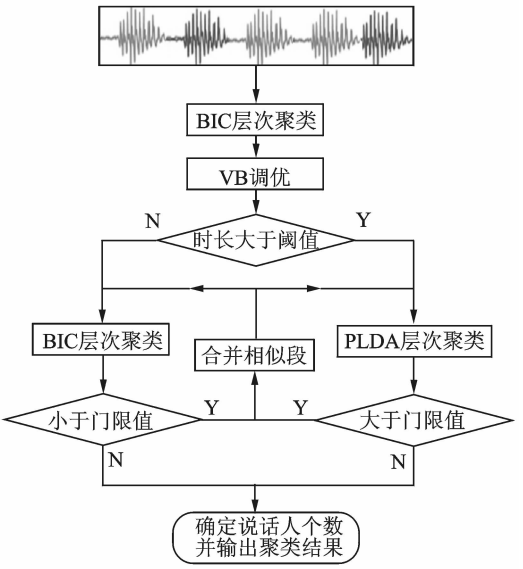

图 3 改进系统实现框图

Fig. 3 Implementation block diagram of the improved system

2.1 变分贝叶斯调优

近年来,Kenny[12],Zheng 等[13]结合变分贝叶斯、联合因子分析以及 I-vector,使得说话人分离效果获得了很大的提升。不同于 K-means 的硬判决方式,VB 是在保证最优化目标函数的前提下,通过对某个片段属于某个说话人的最大后验概率估计的一种软聚类方式。

VB[14]的思想是一种利用形式简单的分布去近似形式比较复杂难以求解的分布 D(如给定数据集 Q,求模型 θ 的后验概率 $p(\theta|x)$ 和边缘似然值 $p(x)$)。而 Q 和 P 之间的相似度可以用 K_L 距离表示,K_L 距离越小,表示 Q 和 P 越相似。在变分贝叶斯的框架下,为了得到后验 $p(\theta|x)$ 的近似表达式和 evidence 的下界 $L(Q)$,可以通过选择合适的 Q,使得 $L(Q)$ 便于计算和求极值。

2.1.1 问题描述

假设给定一条语音,被分成 M 小段 $\boldsymbol{X}=(x_1,x_2,\cdots,x_M)$,每段只包含一个人的语音,且语音中至多有 S 个说话人。因此,需要确定语音中有多少个说话人以及每个说话人对应哪些语音段。

对于每一个段 x_m,给赋予一个 $S\times1$ 维的向量 $\boldsymbol{i}_m$,其中 $i_{ms}=1(s=1,\cdots,S)$ 表示第 m 段来自说话人 s,否则 i_{ms};令 $P(i_{ms}=1)=\pi_s$(π_s 表示了说话人 s 在给定语音段中说话的先验概率),故有

$$P(\boldsymbol{i}_m)=\prod_{s=1}^{S}\pi_s^{i_{ms}} \tag{6}$$

为了估计整条语音中的说话人数目，通过最大化边缘似然值 $P(\boldsymbol{X}\mid\pi)$ 来估计出，π 中非零项的数目即为说话人的数目；为了确定每个说话人对应哪些语音段，需要计算出后验概率 $P(\boldsymbol{I}\mid\boldsymbol{X}\mid\pi)$，即有

$$P(\boldsymbol{X}\mid\pi)=\int P(\boldsymbol{X},\boldsymbol{\theta}\mid\pi)\mathrm{d}\theta$$

$$P(\boldsymbol{I}\mid\boldsymbol{X},\pi)=\frac{P(\boldsymbol{I},\boldsymbol{X}\mid\pi)}{P(\boldsymbol{X}\mid\pi)}\propto P(\boldsymbol{I},\boldsymbol{X}\mid\pi)=\int P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I}\mid\pi)\mathrm{d}Y$$

其中 $\boldsymbol{\theta}=(\boldsymbol{Y},\boldsymbol{I})$，而无论是 $P(\boldsymbol{X}\mid\pi)$ 还是 $P(\boldsymbol{I}\mid\boldsymbol{X}\mid\pi)$ 都不能精确求解，故使用变分方法来分别求解。

2.1.2 变分法

假设 π 是已知的，通过使用分布 $Q(\boldsymbol{Y},\boldsymbol{I})$ 来近似真实的后验分布。这不仅可以在 π 已知的情况下算得每段话所对应的说话人，而且还能计算得到 $P(\boldsymbol{X}\mid\pi)$，从而用来估算 π。在这里假设

$$Q(\boldsymbol{Y},\boldsymbol{I})=Q(\boldsymbol{Y})Q(\boldsymbol{I}) \tag{7}$$

定义辅助函数 $L(Q\mid\pi)$ 为

$$L(Q\mid\pi)=\int Q(\theta)\ln\frac{P(\boldsymbol{X},\boldsymbol{\theta}\mid\pi)}{Q(\boldsymbol{\theta})}\mathrm{d}\theta$$

对于任何 Q，有

$$L(Q\mid\pi)\leqslant\ln P(\boldsymbol{X}\mid\pi)$$

根据变分贝叶斯公式，$Q(Y)$ 和 $Q(I)$ 的更新公式分别为

$$\ln Q(\boldsymbol{Y})=E_I[\ln P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I}\mid\pi)]+\mathrm{const}$$

$$\ln Q(\boldsymbol{I})=E_Y[\ln P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I}\mid\pi)]+\mathrm{const}$$

由于两个更新公式相关联，需要交替更新 $Q(\boldsymbol{Y})$ 和 $Q(\boldsymbol{I})$，每次的更新保证了 $L(Q|\pi)$ 的增长。即

$$L(Q\mid\pi)=\sum_{m=1}^{M}\sum_{s=1}^{S}q_{ms}\left[E_{y_s}[\ln P(x_m\mid y_s)]+\ln\pi_s\right]+\frac{1}{2}\left[R_t S-\sum_{s=1}^{S}(-\ln|\boldsymbol{\Lambda}_s|+\mathrm{tr}(\boldsymbol{\Lambda}_s^{-1}+\boldsymbol{a}_s\boldsymbol{a}_s^{\mathrm{T}}))\right]-$$

$$\sum_{m=1}^{M}\sum_{s=1}^{S}q_{ms}\ln q_{ms} \tag{8}$$

其中

$$\boldsymbol{\Lambda}_s=\boldsymbol{I}+\boldsymbol{V}^{\mathrm{T}}\left(\sum_{m=1}^{M}q_{ms}N_m\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{V}$$

$$\boldsymbol{a}_s=\boldsymbol{\Lambda}_s^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\sum_{m=1}^{M}q_{ms}F_m\right) \tag{9}$$

式中：R_t 为 TV 的因子数；S 为说话人个数；$\boldsymbol{V}$ 为全差异空间变换矩阵；N_m 和 F_m 分别为零阶和一阶统计量；q_{ms} 表示第 m 句话属于第 s 个人的后验概率；不同于一般意义上的 I-vector，此处的 $\boldsymbol{a}_s$ 称为 VB-I-vector；每个说话人片段的后验概率为

$$\ln q_{ms}=\ln\pi_s P(x_m\mid y_s=\boldsymbol{a}_s)-\frac{1}{2}\mathrm{tr}(V^{*}N_m\boldsymbol{\Sigma}^{-1}V\boldsymbol{\Lambda}_s^{-1}) \tag{10}$$

2.2 VB 调优算法流程

已知有 M 个说话人片段，记为 $\boldsymbol{X}=(x_1,x_2,\cdots,x_M)$，一步 BIC 层次聚类后获得最大说话人个数 $S_{\max}$。

(1)For $m=1$ to M

For $s=1$ to $S_{\max}$

计算第 m 个片段和第 S 个人之间的距离 $\mathrm{dis}(m,s)$

End for

End for

(2)初始化 $q_{\mathrm{ms}}=\mathrm{Norm}(\mathrm{dis}(m,s))_{0-1}$

(3)For $i = 1$ to 最大迭代次数

更新 $\boldsymbol{a}_s$,$\boldsymbol{\Lambda}_s$

更新 q_{ms}

计算 $L(Q|\pi)$

If $(L(Q|\pi)-L(Q|\pi)<\epsilon$ then stop

End for

(4)取最大的 q_{ms},并将 m 归为第 s 个说话人;

End when

3 实验结果及分析

3.1 性能评价指标

说话人聚类性能的好坏可以用"平均说话人纯度(Average speaker purity,ASP)"和"平均类纯度(Average cluster purity,ACP)"作为衡量指标,ACP 反映的是同一类别的数据是否都是来自同一说话人,ASP 反映的是同一说话人的语音被分为了多少类别。

设 S 为说话人的总数,C 为聚类后得到的类数,N 为语音的总帧数,n_i 为第 i 类的总帧数,n_j 为第 j 个说话人语音的总帧数,n_{ij} 为第 i 类中包含说话人 j 的语音帧数。则 ACP 和 ASP 的定义为

$$\mathrm{ASP}=\frac{1}{N}\sum_{j=1}^{S}\sum_{i=1}^{C}\frac{n_{ij}^2}{n_j} \qquad \mathrm{ACP}=\frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{S}\frac{n_{ij}^2}{n_i}$$

为了评价不同系统之间的说话人聚类算法整体效果,通常加入一个总体评价指标 K 值[15],定义为

$$K=\sqrt{\mathrm{ACP}*\mathrm{ASP}} \tag{11}$$

为了更明显地看出不同的聚类方式对整体的说话人分离的效果影响,本文在 NIST 提供的 VAD 标注基础上,使用变窗长的 BIC 距离准则进行说话人变换点检测,并以说话人分离错误率(Diarization error rate,DER)作为辅助的标准,即有

$$\mathrm{DER}=\mathrm{Miss}+\mathrm{False}+\mathrm{SpkErr}$$

式中:Miss 指语音段当做静音段处理,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,由于使用的是标签 VAD,所以此处的 Miss 和 False 都为 0;SpkErr 是指将一个说话人的语音段归于另一个说话人的百分比,此值越小表示说话人分离的效果越好。

3.2 训练集、测试集及参数配置

本实验中,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 5 min。训练 UBM,TV 的数据来自 nist04,05,06 大约 500 个小时的音频数据,训练 PLDA 的数据来自 nist04,05,06 共 577 个说话人,平均每个人约有 15 句话。其中模型 UBM 采用 512 高斯数,而 T 空间的因子数取 200,PLDA 的说话人因子数取 150。

3.3 基线系统以及改进系统的聚类结果

表 1 列出了基线方法以及本文的 VB 调优的聚类方法结果,表 2 列出了在 VB 框架下,对每个片段

表 1 不同系统下的性能对比

Tab. 1 Performance comparison in different systems

测试系统	ACP/%	ASP/%	K-value	DER/%
基线系统	88.92	88.37	0.886	1.52
VB 调优	90.41	90.40	0.904	1.10

表 2 VB 不同初始化方式对性能影响

Tab. 2 Influence between different initialization methods on VB

VB 初始化	ACP/%	ASP/%	K-value	DER/%
随机初始化	89.87	89.14	0.895	1.32
余弦值规整	90.41	90.40	0.904	1.10

后验概率以不同方式初始化的结果。

3.4 实验结果分析

(1)从表 1 的对比结果中可以看出,在说话人聚类层面上改进后的系统相比于基线系统,平均类纯度和平均说话人纯度分别提升了 1.68%和 2.30%,而 K-value 值从 0.886 提升到了 0.904,在其他情况完全相同的条件下,改进后的系统使得最终的说话人分类错误率 DER 也相对下降了 27.6%。

(2)表 2 给出了不同的 VB 初始化方式对聚类和分离效果的影响,从中可以发现,无论是随机初始化每个片段的后验概率,还是一步 BIC 层次聚类后,将每个说话人片段与类中心求 Cos 距离作为后验概率的初值(0-1 规整后),其效果相比与基线 BIC+PLDA 都要好。而后者相当于人为地加强或抑制了某个片段属于某个说话人的可能性,所以效果上会有所增加。

(3)从表格中可以看出无论是基线还 VB 改进后的系统,ACP 和 ASP 都不是太高,通过对类纯度较低的音频分析后发现,这些对话双方音色都比较接近,即使在标签 VAD 下,其转折点检测依旧存在误差,很容易产生漏警,进而直接影响聚类结果。

4 结束语

本文针对基线中的 BIC+PLDA 说话人聚类方法,在层次聚类时会出现误差向上传递的情况,提出了逐级算法增强处理机制,依据短时片段上提取出的 VB-I-vector,保证在最优化目标函数的情况下,通过最大后验估计方法对每个短时片段调优。实验结果表明,这种 VB 调优策略对于聚类效果有了一定的提升,且对于整个说话人分离系统也有了很大的效果提升。但是值得注意的是,VB 调优是一种迭代过程,涉及复杂的后验均值和方差的计算,因此就计算的实时率来说,相对于基线系统还是会慢很多。本文提出的方法虽然在电话信道上有很大的改善,但是在实际应用中还面临着各种各样的复杂场景,如背景音较强、对话中含笑声和重叠音等,这些都会影响聚类效果,进而影响说话人分离系统的性能。另外,如何精确地确定说话人实际数目将是未来工作的一个重点。

参考文献:

[1] Nguyen T H, Chng E S, Li H. Speaker diarization: An emerging research[M]//Speech and Audio Processing for Coding, Enhancement and Recognition. New York:Springer,2015: 229-277.

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

[3] 马勇, 鲍长春. 说话人分割聚类研究进展[J]. 信号处理, 2013, 29(9): 1190-1199.
Ma Yong, Bao Changchun. Advances in speaker segmentation and clustering[J]. Signal Processing, 2013, 29(9): 1190-1199.

[4] Chen S. Speaker, environment and channel change detection and clustering via the Bayesian information criterion [C]// Proc DARPA Broadcast News Transcription and Understanding Workshop. Morgan Kaufman:[s. n.], 2000:127-132.

[5] 凌锦雯, 陆伟, 刘青松, 等. 利用 EHMM 和 CLR 的说话人分割聚类算法[J]. 小型微型计算机系统, 2012, 33(6): 1389-1392.
Ling Jinwen, Lu Wei, Liu Qingsong, et al. Speaker diarization using EHMM and CLR[J]. Journal of Chinese Computer Systems, 2012, 33(6): 1389-1392.

[6] Yella S H, Valente F. Information bottleneck features for HMM/GMM speaker diarization of meetings recordings[C]//In-

terspeech. Florence,Italy:Conference of the International Speech Communication Assocition, 2011: 953-956.
[7] Shum S, Dehak N, Chuangsuwanich E, et al. Exploiting intra-conversation variability for speaker diarization[C]// Confeence of the International peech ommunication Association. Florence,Italy:DBLP,2011:945-948.
[8] 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(4):286-293.
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(4):286-293.
[9] Prince S J D, Elder J H. Probabilistic linear discrimant analysis for inferences about identity[C]// ICCV 2007, IEEE 11th International Conference on Computer Vision. Glasgow, Scotland:IEEE, 2007: 1-8.
[10] Rajan P, Kinnunen T, Hautamäki V. Effect of multi condition training on I-vector PLDA configurations for speaker recognition[C]//Interspeech. Lyon,Franle:[s. n.],2013: 3694-3697.
[11] Kenny P, Stafylakis T, Ouellet P, et al. PLDA for speaker verification with utterances of arbitrary duration[C]// IEEE International Conference on Acoustics, Speech and Signal. [S. l.]:IEEE, 2013: 7649-7653.
[12] Kenny P. Bayesian analysis of speaker diarization with eigenvoice priors[R]. CRIM, Montreal, Technical Report, 2008.
[13] Zheng R, Zhang C, Zhang S, et al. Variational bayes based I-vector for speaker diarization of telephone conversations[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy:IEEE, 2014: 91-95.
[14] Tzikas D G, Likas A C, Galatsanos N P. The variational approximation for Bayesian inference[J]. IEEE Signal Processing Magazine,2008,25(6):131-146.
[15] Ajmera J, Bourlard H, Lapidot I, et al. Unknown-multiple speaker clustering using HMM[C]//International Conference on Spoken Language Processing. Colorado,USA:DBLP,2002:573-576.

作者简介:

李敬阳(1964-),男,研究员,研究方向:司法语音音频检验,E-mail:lijingyang@cifs.gov.cn。

李锐(1991-),男,硕士,研究方向:语音信号处理,E-mail:lirui005@mail.ustc.edu.cn。

王莉(1969-),女,副研究员,研究方向:司法语音音频检验,E-mail:wangli@cifs.gov.cn。

王晓笛(1981-),女,副研究员,研究方向:司法语音音频检验,E-mail:wangxiaodi@cifs.gov.cn。