基于特征均值距离的短语音段说话人聚类算法

李艳雄* 吴 永 贺前华
(华南理工大学电子与信息学院 广州 510640)

摘 要：该文提出一种基于特征均值距离的短语音段说话人聚类算法。首先，定义特征均值距离用来在特征层而不是模型层刻画两个类之间的相似度；然后，迭代合并特征均值距离最小的两个类，直到任意两类之间的特征均值距离的最小值大于一个自适应门限为止。采用取自两个语音数据库的短于 3 s 的语音段进行实验测试，结果表明：与基于 AHC+BIC 的算法相比，F 度量值平均提高了 5%，运算速度约为以前算法的 4.68 倍。

关键词：语音信号处理；说话人聚类；特征均值距离；短语音段
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2012)06-1404-04
DOI: 10.3724/SP.J.1146.2011.01139

Feature Mean Distance Based Speaker Clustering for Short Speech Segments

Li Yan-xiong Wu Yong He Qian-hua
(*School of Electronic and Information Engineering*, *South China University of Technology*, *Guangzhou* 510640, *China*)

Abstract: An algorithm of speaker clustering is proposed based on Feature Mean Distance (FMD) for short speech segments. First, a distance measure, i.e. FMD, is introduced to represent the similarities between two clusters on the level of feature instead of the level of model. Then, two clusters with the minimum of FMDs are iteratively merged until the minimum of FMDs is larger than an adaptive threshold. Experimental results show average 5% improvements in F measure are obtained in comparison with the AHC+BIC based algorithm. In addition, the proposed algorithm is 4.68 times faster than the AHC+BIC based algorithm.

Key words: Speech signal processing; Speaker clustering; Feature Mean Distance (FMD); Short speech segments

1 引言

随着多媒体技术的发展，语音文档记录(例如会议语音，谈话语音，电视和广播电台语音)正在迅猛增加[1]。如何有效地组织、浏览和检索这些语音文档已经成为信号处理领域研究者急需解决的一个问题[2,3]。说话人聚类是减轻海量语音文档管理任务的有效工具之一。说话人聚类是将一个语音文档记录中由同一个人发出的所有语音段聚成一类并分配一个标签标识它们[4]。

目前主流的说话人聚类算法是以贝叶斯信息准则(Bayesian Information Criterion, BIC)为收敛准则的凝聚分层聚类(Agglomerative Hierarchical Clustering, AHC)算法，即基于 AHC+BIC 的说话人聚类算法[4−6]。对于一个给定的模型，输入数据的边界对数似然度估计值就是所谓的 BIC 值。BIC 只在大数据量(即较长的语音段)的情况下才有效[7]。另外，采用 BIC 进行说话人聚类时，需要根据不同的输入数据不断地调整 BIC 惩罚系数，要调整得到一个合适的系数非常困难，而该系数对基于 AHC+BIC 算法性能的影响非常大。已有研究表明：采用基于 AHC+BIC 的算法进行说话人聚类时，短语音段(短于 3 s)是产生聚类误差的一个重要因素[8]。也就是说，基于 AHC+BIC 的说话人聚类算法对于短语音段的聚类不是有效的。然而，在会议或谈话语音自动摘要等应用中，能检测出那些只说了几个字(短语音段)的说话人是非常重要的，因为那些短语音段也许正是会话过程中所做出的关键性结论。由于短语音段频繁出现在谈话语音中，特别是多人参与的会议语音中[8]，因此为了克服目前说话人聚类算法在短语音段说话人聚类时所存在的问题，本文提出一种基于特征均值距离的说话人聚类算法。

2 算法介绍

首先，给出特征均值距离(Feature Mean Distance, FMD)的定义，并用它表示任意两个类之

2011-11-03 收到，2012-02-24 改回
国家自然科学基金(61101160, 60972132)，中央高校基本科研业务费专项基金(2011ZM0029)和广东省自然科学基金博士启动项目(10451064101004651)资助课题
*通信作者：李艳雄 eeyxli@scut.edu.cn

间的相似度；接着，给出针对短语音段的说话人聚类算法流程。

2.1 特征均值距离

假设 $\boldsymbol{C}_i$ 和 $\boldsymbol{C}_j$ ($i,j=1,\cdots,N_c$)表示两个类，其中 N_c 表示类的个数。第 n 个语音段 $\boldsymbol{S}_n$ ($n=1,\cdots,N$) 的特征矩阵表示为 $\boldsymbol{O}_n=\{O_n^{m,l},m=1,\cdots,M_n,l=1,\cdots,L\}$ ，其中 N ， M_n 和 L 分别表示语音段个数，特征矩阵 $\boldsymbol{O}_n$ 的帧数和维数。 N_i 表示第 i 个类 $\boldsymbol{C}_i$ 所包含的语音段个数，且

$$N=\sum_{i=1}^{N_c}N_i \tag{1}$$

$\boldsymbol{C}_i$ 由 N_i 个均值矢量 $\boldsymbol{U}_n$ 组成，即 $\boldsymbol{C}_i=[\boldsymbol{U}_1,\cdots,\boldsymbol{U}_n,\cdots,\boldsymbol{U}_{N_i}]^{\mathrm{T}}$, T 表示矩阵的转置。分别对特征矩阵 $\boldsymbol{O}_n$ 的每一列求均值得到均值矢量 $\boldsymbol{U}_n=\{U_n^l,l=1,2,\cdots,L\}$ ，其中

$$U_n^l=\frac{1}{M_n}\sum_{m=1}^{M_n}O_n^{m,l} \tag{2}$$

采用欧式距离(Euclidean distance) $d(\boldsymbol{U}_n,\boldsymbol{U}_n^{'})$ 刻画任意两个均值矢量 $\boldsymbol{U}_n$ (取自类 $\boldsymbol{C}_i$)与 $\boldsymbol{U}_n^{'}$ (取自类 $\boldsymbol{C}_j$)之间的相似度，即

$$d(\boldsymbol{U}_n,\boldsymbol{U}_n^{'})=\sqrt{\sum_{l=1}^{L}(U_n^l-\boldsymbol{U}_n^{'l})^2} \tag{3}$$

从而得到任意两个类 $\boldsymbol{C}_i$ 与 $\boldsymbol{C}_j$ 之间的距离矩阵 $\boldsymbol{d}(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 如下：

$$\boldsymbol{d}(\boldsymbol{C}_i,\boldsymbol{C}_j)=\begin{bmatrix} d(\boldsymbol{U}_1,\boldsymbol{U}_1^{'}), & \cdots, & d(\boldsymbol{U}_1,\boldsymbol{U}_{N_j}^{'}) \\ \vdots & \ddots & \vdots \\ d(\boldsymbol{U}_{N_i},\boldsymbol{U}_1^{'}), & \cdots, & d(\boldsymbol{U}_{N_i},\boldsymbol{U}_{N_j}^{'}) \end{bmatrix}_{N_i\times N_j} \tag{4}$$

依次计算距离矩阵 $\boldsymbol{d}(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 每一行元素之和，从而得到一个列矢量 $\boldsymbol{d}=[d_1,\cdots,d_{N_i}]^{\mathrm{T}}$ ，再计算该列矢量各元素之和。因此，任意两类之间的特征均值距离 $D(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 定义如下：

$$D(\boldsymbol{C}_i,\boldsymbol{C}_j)=\frac{\xi}{2}\cdot\sum_{i=1}^{N_i}\left(\sum_{j=1}^{N_j}d(\boldsymbol{U}_i,\boldsymbol{U}_j^{'})\right),\quad \xi=\lg\left(N_i+N_j\right) \tag{5}$$

$D(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 越小，表示这两类越相似。 ξ 是一个权重系数。由于 ξ 的存在，两个大类(N_i 和 N_j 较大，即语音段个数或均值矢量个数较多的两个类)之间的特征均值距离将倾向于大于两个小类之间的特征均值距离。也就是说，由于 ξ 的存在，在迭代合并的过程中，两个大类将倾向于不被合并(因为它们之间的距离较大)，而两个小类将倾向于被合并。

从式(5)可知，特征均值距离是在特征层而不是模型层刻画具有不同语音段个数的两个类之间的相似度(距离)。采用该距离测度表示任意两个类之间的相似度时，不需要对包含不同语音段个数的类(即不同大小的类)做规整处理，也不需要假设每个类中各个语音段的特征的概率分布。因此，它能够有效刻画具有不同语音段个数的两个类之间的相似度，特别是短语音段所组成的类。

2.2 说话人聚类

基于特征均值距离的说话人聚类算法流程如下：

第 1 步　将 N 个待聚类的语音段分成 N_c 类。初始的任意一类 $\boldsymbol{C}_i$ 只包含一个语音段(即 $N_c=N$)，并表示为： $\boldsymbol{C}_i=[\boldsymbol{U}_n]$ ， $n=i$ 。

第 2 步　根据式(5)计算任意两个类($\boldsymbol{C}_i$ 与 $\boldsymbol{C}_j$)之间的特征均值距离 $D(\boldsymbol{C}_i,\boldsymbol{C}_j)$ ，得到 $(1/2)\times N_c\times(N_c-1)$ 个 $D(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 。提取出最小的 $D(\boldsymbol{C}_i,\boldsymbol{C}_j)$ ，即 $D_{\min}(\boldsymbol{C}_i,\boldsymbol{C}_j)$ ，并将它与一个门限 d_t 进行比较，其中

$$d_t=\min\left(D(\boldsymbol{C}_i,\boldsymbol{C}_{\overline{ij}}),D(\boldsymbol{C}_j,\boldsymbol{C}_{\overline{ij}})\right) \tag{6}$$

$\min(\cdot,\cdot)$ 表示求取两个元素中的最小值。 $\boldsymbol{C}_{\overline{ij}}$ 由 $N_{\overline{ij}}$ 个均值矢量 $\boldsymbol{U}_n$ 组成，但不包括第 i 类和第 j 类中的所有均值矢量。也就是， $\boldsymbol{C}_{\overline{ij}}=[\boldsymbol{U}_1,\cdots,\boldsymbol{U}_n,\cdots,\boldsymbol{U}_{N_{\overline{ij}}}]^{\mathrm{T}}$ ， $(\boldsymbol{U}_n\notin\boldsymbol{C}_i,\boldsymbol{U}_n\notin\boldsymbol{C}_j$ ， $N_{\overline{ij}}=N-N_i-N_j)$ 。根据式(6)可知， d_t 表示待合并的两个类与其他类(即背景类)之间的相似度，且随着迭代次数的增加， d_t 会自适应地变化(因为各类所包含的均值矢量会被更新)。如果两个待合并类之间的距离小于它们与背景类之间的距离，即 $D_{\min}(\boldsymbol{C}_i,\boldsymbol{C}_j)<d_t$ ，则跳到第 3 步；否则跳到第 4 步。

第 3 步　假设特征均值距离最小的那两个类为： $\boldsymbol{C}_i=[\boldsymbol{U}_1,\cdots,\boldsymbol{U}_{N_i}]^{\mathrm{T}}$ 和 $\boldsymbol{C}_j=[\boldsymbol{U}_1,\cdots,\boldsymbol{U}_{N_j}]^{\mathrm{T}}$ ，将它们合并为： $\boldsymbol{C}_i=[\boldsymbol{U}_1,\cdots,\boldsymbol{U}_{N_{ij}}]^{\mathrm{T}}$ ，其中 $N_{ij}=N_i+N_j$ 。然后类别数减 1，即 $N_c=N_c-1$ ，且删除第 j 类，跳到第 2 步继续迭代。

第 4 步　 N_c 作为最后的类别数;那些语音段 $\boldsymbol{S}_n$ 被判为同一个说话人，如果与它们对应的均值矢量 $\boldsymbol{U}_n$ 被聚类包含在同一个类 $\boldsymbol{C}_i$ 中。

2.3 算法计算复杂度

基于 AHC+BIC 的说话人聚类算法流程[4]与本文第 2.2 节所描述的流程基本相同，主要区别在于两个类之间的距离计算式及收敛条件。假设两种算法所采用的特征相同，则它们的计算量的差别主要在于两个类之间的距离计算。以两个语音样本之间的距离计算为例，比较两种距离计算的复杂度。

假设两个语音段 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 的特征矩阵分别为 $\boldsymbol{O}_1=\{O_1^{m,l},\ m=1,\cdots,M_1,\ l=1,\cdots,L\}$ 和 $\boldsymbol{O}_2=\{O_2^{m,l},\ m=1,\cdots,M_2,\ l=1,\cdots,L\}$ ，其中 M_1 ， M_2 分别表示两个特征矩阵的行数(帧数)， L 表示特征矩阵的列数(维数)。根据式(2)计算两个特征矩阵的均值矢量所

需计算量：$(M_1-1)+(M_2-1)$ 次 L 维矢量加法，$1+1$ 次 L 维矢量除法；根据式(3)计算两个特征均值之间的距离所需计算量：1 次 L 维矢量减法，1 次 L 维矢量乘法，$L-1$ 次标量加法，1 次标量开方。因此，计算两个语音样本之间的特征均值距离所需的总计算量：$(M_1+M_2-1)\times L-1$ 次标量加法，$2\times L$ 次标量除法，L 次标量减法，L 次标量乘法，1 次标量开方。

计算两个语音样本之间的 BIC 距离的公式为[4]

$$\begin{aligned}\Delta\mathrm{BIC}=&M_1\times\ln\left(\left|\det\left(\mathrm{cov}\left(\boldsymbol{O}_1\right)\right)\right|\right)\\&+M_2\times\ln\left(\left|\det\left(\mathrm{cov}\left(\boldsymbol{O}_2\right)\right)\right|\right)\\&+\lambda\times\ln\left(M_3\right)\times\left(L+0.5L\times(L+1)\right)\\&-M_3\times\ln\left(\left|\det\left(\mathrm{cov}\left(\boldsymbol{O}_3\right)\right)\right|\right)\end{aligned}\tag{7}$$

其中 $\mathrm{cov}(\cdot)$ 表示计算特征矩阵的协方差矩阵；$\det(\cdot)$ 表示计算方阵的行列式值；$\ln(\cdot)$ 表示自然对数；$\boldsymbol{O}_3$ 由特征矩阵 $\boldsymbol{O}_1$ 和 $\boldsymbol{O}_2$ 合并而成，即 $\boldsymbol{O}_3=[\boldsymbol{O}_1;\boldsymbol{O}_2]$，$M_3=M_1+M_2$ 是特征矩阵 $\boldsymbol{O}_3$ 的行数(帧数)。ΔBIC 的主要计算量在于协方差矩阵 $\mathrm{cov}(\boldsymbol{O}_n)$ 和行列式值 $\det(\cdot)$ 的计算，其中

$$\left.\begin{aligned}\mathrm{cov}\left(\boldsymbol{O}_n\right)=&\begin{bmatrix}\mathrm{cov}\left(O_n^{m,1},O_n^{m,1}\right),\mathrm{cov}\left(O_n^{m,1},O_n^{m,2}\right),\cdots,\mathrm{cov}\left(O_n^{m,1},O_n^{m,L}\right)\\\mathrm{cov}\left(O_n^{m,2},O_n^{m,1}\right),\mathrm{cov}\left(O_n^{m,2},O_n^{m,2}\right),\cdots,\mathrm{cov}\left(O_n^{m,2},O_n^{m,L}\right)\\\vdots\qquad\qquad\vdots\qquad\qquad\ddots\qquad\qquad\vdots\\\mathrm{cov}\left(O_n^{m,L},O_n^{m,1}\right),\mathrm{cov}\left(O_n^{m,L},O_n^{m,2}\right),\cdots,\mathrm{cov}\left(O_n^{m,L},O_n^{m,L}\right)\end{bmatrix}_{L\times L}\\\mathrm{cov}\left(O_n^{m,l+1},O_n^{m,l}\right)=&\mathrm{mean}\left[\left(O_n^{m,l+1}-\mathrm{mean}\left(O_n^{m,l+1}\right)\right)\times\left(O_n^{m,l}-\mathrm{mean}\left(O_n^{m,l}\right)\right)\right]\end{aligned}\right\}\tag{8}$$

其中 $m=1,2,\cdots,M_n$。从式(8)可知，$\mathrm{cov}(\boldsymbol{O}_n)$ 是一个 L 阶对称方阵，需要计算 $0.5\times L\times(L+1)$ 个不同的元素 $\mathrm{cov}(O_n^{m,l+1},O_n^{m,l})$，其中，需要计算：$L$ 次均值 $\mathrm{mean}(O_n^{m,l})$ 和 $0.5\times L\times(L+1)$ 次均值 $\mathrm{mean}[(O_n^{m,l+1}-\mathrm{mean}(O_n^{m,l+1}))\times(O_n^{m,l}-\mathrm{mean}(O_n^{m,l}))]$。计算均值 $\mathrm{mean}(O_n^{m,l})$ 需要计算：M_n-1 次标量加法，1 次标量除法；计算 $\mathrm{mean}[(O_n^{m,l+1}-\mathrm{mean}(O_n^{m,l+1}))\times(O_n^{m,l}-\mathrm{mean}(O_n^{m,l}))]$ 需要：$2\times M_n$ 次标量减法，M_n 次标量乘法，M_n-1 次标量加法，1 次标量除法。因此，计算 $\mathrm{cov}(\boldsymbol{O}_n)$ 需要：$0.5L\times(L+3)\times(M_n-1)$ 次标量加法，$M_n\times L\times(L+1)$ 次标量减法，$0.5M_n\times L\times(L+1)$ 次标量乘法，$L\times(0.5L+2)$ 次标量除法。计算 3 个协方差矩阵($\mathrm{cov}(\boldsymbol{O}_n)$，$n=1,2,3$)总共需要：$0.5L\times(L+3)\times(M_1+M_2+M_3-3)$ 次标量加法，$(M_1+M_2+M_3)\times L\times(L+1)$ 次标量减法，$0.5(M_1+M_2+M_3)\times L\times(L+1)$ 次标量乘法，$L\times(1.5L+6)$ 次标量除法。

计算行列式值 $\det(\cdot)$：

$$\begin{aligned}\det(\cdot)&=\begin{vmatrix}a_{11},&a_{12},&\cdots,&a_{1L}\\a_{21},&a_{22},&\cdots,&a_{2L}\\\vdots&\vdots&\ddots&\vdots\\a_{L1},&a_{L2},&\cdots,&a_{LL}\end{vmatrix}\\&=\sum_{p_1p_2\cdots p_L}(-1)^{\tau\left(p_1p_2\cdots p_L\right)}a_{1p_1}a_{2p_2}\cdots a_{Lp_L}\end{aligned}\tag{9}$$

其中 $p_1p_2\cdots p_L$ 是 $1,2,\cdots,L$ 的全排列，$\tau(p_1p_2\cdots p_L)$ 是 $p_1p_2\cdots p_L$ 的逆序数，Σ 是对所有 $1,2,\cdots,L$ 的全排列求和(共有 $L!$ 项)。根据式(9)，计算 L 阶行列式 $\det(\cdot)$ 需要：$L\times(L!)$ 次标量乘法，$L!$ 次标量加法。

综上所述，计算式(7)中的 ΔBIC 需要计算 3 个协方差矩阵($\mathrm{cov}(\boldsymbol{O}_n)$，$n=1,2,3$)和 3 个 L 阶行列式值，至少需要：$0.5L\times(L+3)\times(M_1+M_2+M_3-3)+3(L!)$ 次标量加法，$(M_1+M_2+M_3)\times L\times(L+1)$ 次标量减法，$0.5(M_1+M_2+M_3)\times L\times(L+1)+3L\times(L!)$ 次标量乘法，$L\times(1.5L+6)$ 次标量除法。因此，计算两个语音样本之间的 BIC 距离 ΔBIC 所需计算量明显大于计算特征均值距离所需计算量。

3 实验及结果分析

3.1 实验数据及设置

实验数据取自汉语普通话自然口语对话语料库(Chinese Annotated Dialogue and Conversation Corpus, CADCC)和863四大方言普通话语音语料库 RASC863(863 annotated 4 regional accent speech corpus)。上述语音数据的采样频率为16 kHz, 16 bit量化，并保存为单声道WAV格式。第1组实验数据取自CADCC，包括 $N=3600$ 个语音段，由3种不同时长(1 s, 2 s和3 s)的语音段组成，共有6个说话人，每个说话人在每种时长情况下包含200个语音段，共 $N=6\times3\times200=3600$ 个语音段。第2组实验数据取自RASC863，包括 $N=1800$ 个语音段，由3种不同时长(1 s, 2 s和3 s)的语音段组成，共有6个说话人，每个说话人在每种时长情况下包含100个语音段，共 $N=6\times3\times100=1800$ 个语音段。

实验平台：Intel(R) Core(TM) i3 CPU 2.53 GHz, 2 GB, RAM，Windows XP操作系统，编程软件MATLAB。对每个语音段分帧并提取特征，帧长为32 ms，帧移为16 ms，特征维数 $L=29$。特征

包括短时能量(1维)、过零率(1维)、基频(1维)、梅尔频率倒谱系数(13维)及其一阶差分系数(13维)。上述特征的提取方法参见文献[9]。

n_{ik} 表示在第 i 类中由第 k 个说话人发出的所有语音段个数； N_s 表示说话人总个数； N_c 表示类的总个数； N 表示语音段总个数； $n_{\bullet k}$ 表示由第 k 个人所发出的语音段总个数。 $n_{i\bullet}$ 表示第 i 类所包含的语音段总个数。第 i 类的纯度， $\pi_{i\bullet}$ ，定义如下：

$$\pi_{i\bullet} = \sum_{k=1}^{N_s} \frac{n_{ik}^2}{n_{i\bullet}^2} \tag{10}$$

平均类纯度 ACP (Average Clustering Purity)[4]定义如下：

$$\mathrm{ACP} = \frac{1}{N} \sum_{i=1}^{N_c} \pi_{i\bullet} n_{i\bullet} \tag{11}$$

第 k 个说话人的纯度 $\pi_{\bullet k}$ ，定义如下：

$$\pi_{\bullet k} = \sum_{i=1}^{N_c} \frac{n_{ik}^2}{n_{\bullet k}^2} \tag{12}$$

平均说话人纯度 ASP (Average Speaker Purity)[4]定义如下：

$$\mathrm{ASP} = \frac{1}{N} \sum_{k=1}^{N_s} \pi_{\bullet k} n_{\bullet k} \tag{13}$$

最后，采用 F 度量值作为算法整体性能评价指标，定义如下：

$$F = \frac{2 \times \mathrm{ACP} \times \mathrm{ASP}}{\mathrm{ACP} + \mathrm{ASP}} \tag{14}$$

3.2 实验结果

基于 AHC+BIC 的说话人聚类算法所采用的特征与本文算法所采用的相同，BIC 惩罚系数实验设置最优值为 2.0。在相同实验条件下，两种算法对第 1 组实验数据进行说话人聚类，结果如表 1 所示；对第 2 组实验数据进行说话人聚类，结果如表 2 所示。由表 1 可知：对 1 s, 2 s 和 3 s 的语音段分别进行聚类时，本文算法平均获得的 ASP, ACP 和 F 分别为 86.7%, 79.9%和 83.2%，耗时 4772 s，与基于 AHC+BIC 算法相比，F 值平均提高了 5.4%，运算速度是后者的 4.67 倍。由表 2 可知：对 1 s, 2 s 和 3 s 的语音段分别进行聚类时，本文算法平均获得的 ASP, ACP 和 F 分别为 84.2%, 79.7%和 81.9%，耗时 2375 s，与基于 AHC+BIC 算法相比，F 值平均提高了 4.6%，运算速度约为后者的 4.69 倍。

表 1 各种算法对第 1 组实验数据的聚类结果

	本文算法			基于 AHC+BIC 算法		
时长(s)	ASP	ACP	F	ASP	ACP	F
1	0.781	0.715	0.747	0.535	0.979	0.692
2	0.897	0.821	0.857	0.678	0.982	0.802
3	0.922	0.862	0.891	0.716	0.988	0.830
均值	0.867	0.799	0.832	0.643	0.983	0.778
耗时		4772			22271	

表 2 各种算法对第 2 组实验数据的聚类结果

	本文算法			基于 AHC+BIC 算法		
时长(s)	ASP	ACP	F	ASP	ACP	F
1	0.770	0.715	0.742	0.552	0.931	0.693
2	0.861	0.813	0.836	0.681	0.961	0.797
3	0.895	0.864	0.879	0.723	0.956	0.823
均值	0.842	0.797	0.819	0.652	0.949	0.773
耗时		2375			11135	

4 结论

本文提出一种基于特征均值距离的短语音段说话人聚类算法。该算法在计算两个类的相似度时不需要对类长度及语音段长度做规整处理，也不需要假设类的特征的概率分布，门限参数 d_t 能够自适应调整。与基于 AHC+BIC 算法相比，F 值平均提高了 5%，运算速度约为后者的 4.68 倍，表明本文算法对短语音段说话人聚类更加有效。

参 考 文 献

[1] Ostendorf M, Favre B, Grishman R, *et al*.. Speech segmentation and spoken document processing[J]. *IEEE Signal Processing Magazine*, 2008, 25(3): 59–69.

[2] Bouamrane M M and Luz S. Meeting browsing state-of-the-art review[J]. *Multimedia Systems*, 2007, 12(4–5): 439–457.

[3] Tur G, Stolcke A, Voss L, *et al*.. The CALO meeting assistant system[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, 18(6): 1601–1611.

[4] Margarita K, Vassiliki M, and Constantine K. Speaker segmentation and clustering[J]. *Signal Processing*, 2008, 88(5): 1091–1124.

[5] Xavier A and Jean-Francois B. Fast speaker diarization based on binary keys[C]. International Conference on Acoustics, Speech and Signal Processing, IEEE, Prague, 2011: 4428–4431.

[6] Imseng D and Friedland G. Tuning-robust initialization methods for speaker diarization[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, 18(8): 2028–2037.

[7] Valente F, Motlicek P, and Vijayasenan D. Variational Bayesian speaker diarization of meeting recordings[C]. International Conference on Acoustics, Speech and Signal Processing, IEEE, Dallas, 2010: 4954–4957.

[8] Han K J, Kim S, and Narayanan S S. Robust speaker clustering strategies to data source variation for improved speaker diarization[C]. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, Kyoto, 2007: 262–267.

[9] Li Y X and He Q H. Detecting laughter in spontaneous speech by constructing laughter bouts[J]. *International Journal of Speech Technology*, 2011, 14(3): 211–225.

李艳雄： 男，1980 年生，博士，讲师，研究方向为信号处理及模式识别.

吴　永： 男，1969 年生，博士生，副教授，研究方向为信号处理及嵌入式系统设计.

贺前华： 男，1965 年生，博士，教授，从事语音识别、身份认证、多媒体信息处理及嵌入式系统设计方面的研究.