改进的说话人聚类初始化和 GMM 的多说话人识别 *

曹 洁[a], 余丽珍[b]
(兰州理工大学 a. 计算机与通信学院; b. 电气工程与信息工程学院, 兰州 730050)

摘 要: 针对多说话人聚类线性初始化方法精度较差的问题,提出了一种改进的聚类初始化方法。该方法引入 BIC 对由线性初始化产生的初始类进行检测分割,有效提升了说话人初始类纯度。最后将该方法应用到高斯混合模型(GMM)多说话人识别系统。实验结果表明,所提方法使说话人平均类纯度(ACP)提高了 48.51%,系统的错误识别率平均降低 12.09%。

关键词: 多说话人识别; 改进的聚类初始化; 高斯混合模型; 平均类纯度
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0590-04
doi:10.3969/j.issn.1001-3695.2012.02.050

Improved speaker clustering initialization and GMM multi-speaker recognition

CAO Jie[a], YU Li-zhen[b]
(*a. College of Computer & Communication, b. College of Electrical & Information Engineering, Lanzhou University of Technology, Lanzhou* 730050, *China*)

Abstract: Aiming at the problem of the linear initialization method of multiple speaker clustering with poor accuracy, this paper proposed an improved method of clustering initialization. The method by introducing BIC to detect and segment for initial cluster produced by the linear initialization, and promoted the purity of speaker initial cluster effectively. Finally, applied the method to Gaussian mixture model (GMM) multi-speaker recognition system. And the experimental results show that this proposed method makes the average cluster purity (ACP) have been increased by 48.51%, and the error recognition of system have been reduced by 12.09% on average.

Key words: multi-speaker recognition; improved clustering initialization; Gaussian mixture model; average cluster purity

0 引言

说话人识别是目前语音识别中的一个热点课题,它可以简单地定义为:以说话人的语音作为输入,然后将待测语音与训练得到的说话人模型库中的模型进行模式匹配,最终识别出说话人的身份[1]。近年来,随着 IT 技术和音频检索技术的发展,各类音频文档的获取途径越来越丰富,包括电话语音、广播语音以及会议语音等。同时,由于此类文档数据量的爆炸式增长,使得对其进行文档管理的难度越来越大。其中,对会议语音的检索难度尤为突出,因为会议文档中包含有多个信道、多个说话人。为了解决这一难题,多说话人识别技术应运而生。多说话人识别的关键就是要从一段给定的目标语音中找到说话人的改变点以及找出每个人分别说了哪些话,即回答“谁在说话”以及“在什么时候说话”这两个问题。多人识别任务大大复杂于普通的单说话人识别的原因在于单人识别只含有单一说话人的纯净语音,而多人则首先需要从多人混合语音段中提取出单人的纯净语音,此过程称为多说话人分割聚类。分割聚类处理后提取的单人语音的“纯度”极大地影响了说话人识别的整体正确率。

目前对说话人的分割聚类多采用自底向上聚类法[2](也称为凝聚聚类)。在该聚类算法过程中由于初始话者类后续被迭代分割聚类,如果初始类的选择不够精确,在后续迭代中将会造成错误累积,从而影响正确的说话人模型训练,最终影响系统的识别性能。因此聚类算法的初始化显得尤为重要。文献[3]中采用线性初始化,即将语音数据平均分成 K 个初始话者类。该方法简单易行,但会造成某个话者类中含有多个说话人语音的情况,从而导致说话人模型训练不准确。文献[2]中提到基于 MFCC 的 K-means 初始化,但该方法较之线性初始化并无改进。

继说话人的分割聚类完成后需要为每一个说话人建立与之对应的模型。目前常用的说话人模型训练方法有隐马尔可夫模型(HMM)、支持向量机模型(SVM)、矢量量化模型(VQ)、人工神经网络模型(ANN)以及高斯混合模型(GMM)[47]。其中性能较好且较为流行的经典模型是 GMM。此外,常用的说话人特征有线性预测系数(LPC)、线性预测倒谱系数(LPCC)以及 Mel 倒谱系数(MFCC)。其中主流的说话人特征为 MFCC。

本文在常用的说话人聚类线性初始化方法基础上采用贝叶斯信息准则(BIC)对其进行改进,有效提升了初始话者类的纯度。最后将该方法应用到 GMM 模型的多说话人识别系统中。

1 多说话人识别原理

本次实验设计的多说话人识别系统主要由两个阶段组成:

收稿日期: 2011-07-24; **修回日期**: 2011-08-30 **基金项目**: 甘肃省财政厅资助项目(0914ZTB148);甘肃省自然科学基金资助项目(1014ZSB064)

作者简介: 曹洁(1966-),女,安徽宿州人,教授,博导,硕士,主要研究方向为信息融合理论与应用;余丽珍(1985-),女,硕士研究生,主要研究方向为智能信息处理(286262112@qq.com).

a)训练阶段,包括预处理、特征提取、说话人聚类、模型训练;b)识别阶段,包括预处理、特征提取、模式匹配。其原理框图如图 1 所示。

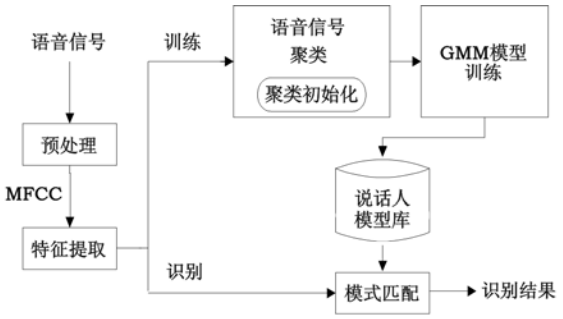

图1　多说话人识别原理框图

1.1　训练阶段

1.1.1　特征提取

在特征提取前需要对训练的语音信号进行预处理,主要包括对输入的语音信号进行预加重、加窗等操作。其中预加重的目的即通过对语音信号中高频信号的加强,使语音信号的频谱便于统一的分析与处理。实验中选择预加重系数为 0.97。加窗的目的是考虑到语音信号的短时平稳特性,采用加窗操作将语音分成一系列的语音帧。实验中采用的是汉明窗,帧长为 30 ms,帧移为 10 ms。

语音的 MFCC 描述了人的感知特性,故更符合人耳听觉特性。实验中将其作为说话人的特征参数。

MFCC 的计算公式如下:

$$\mathrm{MFCC}(t,i)=\sqrt{\frac{2}{N}}\sum_{j=1}^{N}\lg[E_{\mathrm{met}}(t,j)]\cos\left[i(j-0.5)\frac{\pi}{N}\right] \quad (1)$$

其中:$E_{\mathrm{met}}(t,j)$为 t 时刻第 j 个滤波器输出的能量;N 为三角滤波器个数;$\{\mathrm{MFCC}(t,i)\}_{i=1,2,\cdots,p}$为 t 时刻对应的 MFCC 参数,P 为阶数,实验中 P 确定为 19。

1.1.2　改进的说话人聚类初始化

提取的 MFCC 经语音/非语音检测器后[2],去掉非语音部分,只保留语音部分,至此得到的语音部分中包含有多个说话人的语音。要想为每个说话人训练与之对应的模型,首先必须从混有多人的语音部分中提取出属于每个人的“纯净”语音,该过程称之为多说话人的语音分割聚类。完成这一任务典型的方法是凝聚聚类法。该方法是一个迭代分割聚类的过程,且首先必须对其进行初始化,即得出初始话者类以备后续迭代之用。在此过程中,迭代操作一直进行直到满足停止条件。如果初始的话者类存在错误,后续过程中将无法得到修正,甚至会造成错误累积。因此初始类的选择对整个聚类过程显得尤为重要,这也进一步说明了聚类初始化的重要性和必要性。

常用的初始化方法为线性初始化,该方法将语音部分平均分成 K 类,将其作为 K 个初始的话者类(K 一般大于实际说话人数,通常取值为 16)。

而这种等长的平均分配不可避免地会带来一些“不纯净”的初始类,即一个初始类中包含有多个说话人的语音数据,一旦由于这种“不纯净”语音数据造成的错误形成,在后续迭代过程中不但不会被消除,反而将会被累积。因此不仅会得到不正确的说话人模型,而且还会漏掉另一些说话人的语音数据,从而导致误警率和漏警率的增大。可见初始类的纯净度会影响整个聚类过程的精确度。当然,也可以通过增大 K 值,即产生更多的初始类,在一定程度上可以避免“不纯净”初始类的形成。但考虑到聚类过程本身的复杂性,过多的初始类会进一步增大该过程的时间复杂度及空间复杂度。

故可以考虑提高初始类的纯度,为此本实验在线性初始化的基础上采用 BIC 对其进行改进,将检测出的“不纯净”的初始类进行再分割,从而得到更为纯净的初始话者类。改进线性初始化的算法描述如图 2 所示。

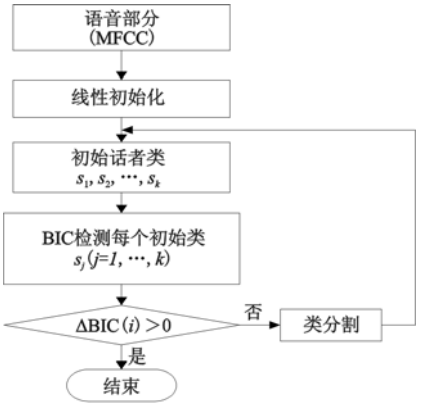

图2　改进的线性初始化

其主要步骤如下:

a)对语音部分(MFCC)进行线性初始化,产生初始话者类 $s_1, s_2, \cdots, s_k$;

b)对每一个初始话者类 $s_j(j=1,2,\cdots,k)$进行 BIC 检测;

c)若 $\Delta\mathrm{BIC}(i)<0$,在 i 时刻将 s_j 分割,否则保留原始类。

其中对话者改变点的检测描述如下:

假设每一个初始类的语音数据 X 满足多元高斯分布,对是否发生说话人改变作假设检验如下:

$$H_0:(x_1,\cdots,x_N)N(\mu,\Sigma) \quad (2)$$

$$H_1:(x_1,\cdots,x_i)N_1(\mu_1,\Sigma_1);(x_{i+1},\cdots,x_N)N_2(\mu_2,\Sigma_2) \quad (3)$$

其中:H_0 表示该语音数据属于同一个说话人,H_1 表示该语音数据属于不同说话人。若在时刻 i 出现了说话人改变点,则 H_0 和 H_1 的极大似然比定义为

$$R_i=N\log|\Sigma|-N_1\log|\Sigma_1|-N_2\log|\Sigma_2| \quad (4)$$

其中:Σ、Σ_1、Σ_2分别表示$\{x_1,\cdots,x_N\}$,$\{x_1,\cdots,x_i\}$,$\{x_{i+1},\cdots,x_N\}$的协方差矩阵,μ、μ_1、μ_2 为对应的均值。若从 H_0、H_1 两个模型选其一来考虑,它们之间的 BIC 差值可表示为

$$\Delta\mathrm{BIC}(i)=-R(i)+\lambda P \quad (5)$$

其中:$P=\frac{1}{2}\left(d+\frac{1}{2}d(d+1)\right)\log N$, d 是样本空间的维数,λ 是惩罚因子。如果 $\Delta\mathrm{BIC}(i)<0$,则表明 H_1 假设成立,即在 i 时刻说话人发生改变,故在该时刻将对应的初始类进行分割;否则 H_0 假设成立,即在 i 时刻说话人未改变,保留原初始话者类。

初始话者类形成后,下一步的工作是要判断哪些语音段是同一个说话人所说的,即进行说话人聚类。说话人聚类的任务是把对话语音中属于同一说话人的所有语音段聚集成一个集合,一个聚类应该只包含一个说话人的语音,同一个说话人的语音应该只包含在一个聚类中。本实验采用凝聚聚类算法,该算法先把每一个语音段都单独作为一个类,然后进行迭代合并,每次迭代把两个最符合合并条件的类合并成为一个新的类,直到算法终止。

1.1.3　说话人的 GMM 模型训练

本文采用 GMM 模型对聚类后的话者类进行建模,得到每一个说话人的模型。GMM 的主要步骤和思想是在说话人训练的时候,从已知说话人的语音中提取出特征矢量,由多个特征

矢量估计出一组参数集,使得训练语音的概率密度最大。

每一个说话人的语音特征在特征空间中都形成了特定的分布,可以利用这一分布来描述说话人的个性。高斯混合模型用多个高斯分布的线性组合来近似描述说话人的特征分布。GMM 本质上是一种多维概率密度函数,每一说话人的概率密度函数的形式是相同的,所不同的只是函数中的参数[8,9]。一个具有 M 个混合成分的 D 维 GMM,可以用 M 个高斯成员的加权和来表示,即

$$P(x_n \mid \lambda) = \sum_{i=1}^{M} \omega_i b_i(x_n) \tag{6}$$

其中:x_n 是一个 D 维观测矢量;λ 为某一特定说话人模型;$\omega_i (i = 1,2,\cdots,M)$ 为混合权值,相当于每个高斯成员出现的概率,且 $\sum_{i=1}^{M} \omega_i = 1$,这样就可以保证混合密度能代表一个真正的概率密度函数;$b_i(x_n)$ 为 D 维高斯函数,即

$$b_i(x_n) = \frac{1}{2\pi^{D/2} |\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2}(x_n - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (x_n - \mu_i) \right\} \tag{7}$$

其中:μ_i 为状态平均矢量,Σ_i 为协方差矩阵,D 为特征矢量维数。这里共有 M 个高斯分布函数,每个函数用 $b_i(x_n)$ 表示,$i = 1,2,\cdots,M$,其参数为 μ_i 和 Σ_i。每个函数经 ω_i 加权后,取和得到 x_n 的概率分布。

整个高斯混合模型便可以由各均值矢量、协方差矩阵及混合分量的权值来描述,因此将得到一个 GMM 模型参数 λ 表示为如下三元式:

$$\lambda = \{\omega_i, \mu_i, \Sigma_i\} \quad i = 1,2,\cdots,M \tag{8}$$

协方差矩阵 Σ_i 可以用满矩阵,但用满矩阵套用公式运算量将是非常大的。因此在实现中常常将其简化成对角矩阵,从而给运算带来很大的方便。另外,GMM 模型中的参数估计采用最大似然估计,即 EM 算法。对 GMM 进行训练需用初始模型进行初始化。EM 算法能够找到一个概率的局部最大值,但是 GMM 模型概率方程有若干个局部最大值,不同的初始化可能导致不同的局部最大值。为了考察不同的模型初始化方法对说话人识别性能的影响,说话人模型用不同的初始化进行训练,然后进行实验。

1.2　识别阶段

在识别阶段,待识别的语音信号经预处理及特征提取之后,计算该特征向量与训练阶段得到的每一个说话人模型的后验概率,根据最大后验概率准则将概率最大的那一组模型所代表的说话人作为识别结果。具体实现如下:

在训练阶段为每个说话人建立了一个 GMM 模型,记其对应的 GMM 模型分别为 $\lambda_1, \lambda_2, \cdots, \lambda_N$。假设待测语音的观察特征矢量序列为 $X = \{x_1, x_2, \cdots, x_T\}$,则识别结果由最大后验概率准则给出,即

$$\hat{s} = \arg\max_{1 \leqslant k \leqslant N} p(\lambda_k \mid X) = \arg\max_{1 \leqslant k \leqslant N} \frac{p(X \mid \lambda_k) P(\lambda_k)}{p(X)} \tag{9}$$

假设每个说话人出现的先验概率相等,即 $p(\lambda_k) = 1/N$,此外由于 $p(X)$ 对每个说话人都是相同的,故式(9)可简化为

$$\hat{s} = \arg\max_{1 \leqslant k \leqslant N} p(X \mid \lambda_k) \tag{10}$$

这样最大后验概率准则就转换为最大似然准则。为了简化计算采用了对数似然函数:

$$L(X \mid \lambda_k) = \log p(X \mid \lambda_k) \quad k = 1,2,\cdots,N \tag{11}$$

则说话人识别的计算结果可表示为

$$\hat{s} = \arg\max_{1 \leqslant k \leqslant N} \sum_{t=1}^{T} \log p(x_t \mid \lambda_k) \tag{12}$$

识别过程就按以上过程对每一帧的特征参数得到的对数进行累加,得分最高的说话人为最后识别结果。

2　实验结果与分析实验环境和数据

2.1　实验环境与数据

本次实验用于训练模型以及识别的语料数据来自于 AMI 语料库中的视听会议[10]。

该语料库中的会议环境如图 3 所示。在这个小型会议室中间的桌子中央固定了一个圆盘,圆盘上装有八个麦克风阵列,用来捕捉参会人的语音信号。圆盘四周装有四台近距离摄像机,分别对准四个参会人员以捕捉他们的视频信号(主要是脸部特征)。另外,会议室两侧及后面的墙上各装有一部远距离摄像机(用来捕获视频信息,但不用于特征提取)。会议室前面有一块白板和一个幻灯片投影屏幕,供参会人讲解之用。此外,对于每一位参会人,实验人员还要求他们佩戴耳机麦克风或翻领话筒等近距离音频捕捉设备。

实验中选取了七场平均长度为 20 min 的会议进行仿真实验,会议编号如表 1 所示(七场会议主题一致,参加人员相同)。

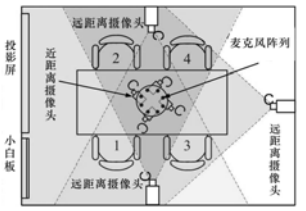

图3　实验环境

表1　实验使用的视听会议编号

会议编号
IS1001a
IS1001c
IS1003b
IS1003d
IS1008a
IS1008d
IS1006b

2.2　不同聚类初始化方法对平均类纯度的影响

平均类纯度(average cluster purity, ACP)的概念由 Ajmera 等人于 2002 年首次提出[3,11],它是用来测度一个话者类中仅包含一个说话人语音的精确度的程度。ACP 的计算公式如下:

$$\mathrm{ACP} = 1/N \sum_{i=1}^{N_c} p_i n_i \tag{13}$$

其中:N 为总帧数;N_c 为总话者类数;n_i 为话者类 i 的总帧数;p_i 为话者 i 的纯度,即

$$p_i = \sum_{j=1}^{N_s} (n_{ij}/n_i)^2 \tag{14}$$

其中:N_s 为说话人总数;n_{ij} 为话者类 i 中包含说话人 j 的语音帧数;n_i 为话者类 i 的总帧数。实验中分别对表 1 选中的七个会议采用不同的聚类初始化方法,得到每个会议的 ACP 如表 2 所示。

表 2　不同初始化对应的 ACP

会议名称	聚类初始化	
	线性初始化	改进的线性初始化
IS1001a	0.12	0.18
IS1001c	0.21	0.27
IS1003b	0.12	0.17
IS1003d	0.09	0.13
IS1008a	0.16	0.25
IS1008d	0.13	0.24
IS1006b	0.18	0.26

从表 2 可以看出,对于每场会议中采用改进后的线性初始化方法得到的 ACP 有了明显提高。这是因为通常采用的线性初始化是将语音数据平均分成 K 个话者类,而这种平均分配很难保证所分配的每一个话者类中只包含一个说话人的语音(一般情况下,只考虑最多包含两个说话人语音的情况),一旦这种错误的话者类在初始化阶段形成,在后续的聚类过程中是

无法消除的,并且会造成错误累积,这也进一步证明了初始类纯度的重要性。因此,当引入改进的线性初始化时,由于对每个初始话者类进行 BIC 检测,并且将"不纯净"进行分割,从而得到相对较纯净的初始话者类,表 2 中改进的线性初始化对应的 ACP 平均提高了 48.51%。

2.3　GMM 不同初始化方法对系统错误识别率的影响

GMM 进行说话人识别训练时,必须确定 GMM 模型的高斯混合分量个数。高斯混合分量个数越多,对说话人特征矢量空间分布密度越逼近,从而提高说话人系统的鲁棒性(robustness)。但是高斯混合分量个数太多,一方面加大系统的计算量和占用更多的系统资源,另一方面在有限时长的训练数据情况下使得模型训练不够充分,从而使得系统性能降低。一般认为模型的高斯混合分量个数在 32 以上系统的性能趋于稳定(本实验高斯分量数确定为 32)[1]。高斯混合分量个数确定后,下一步初始化工作就是给式(8)分别赋初值。常用的参数初始化方法有两种。

实验中选取了上述七个会议中的两场会议进行 GMM 参数初始化的实验,其中采用了两种初始化方法,并且每种方法对应了不同的聚类初始化,得到如表 3 所示的结果。

表 3　不同 GMM 初始化对应的系统错误识别率

GMM 参数初始化	聚类初始化	错误识别率/%
K-means	线性初始化	18.7
	改进的线性初始化	16.3
随机	线性初始化	18.5
	改进的线性初始化	16.4

表 3 中可以看出,如果聚类初始化中采用改进的线性初始化,在 GMM 建模阶段不管采用哪种参数初始化方法,系统的错误识别要低于常用的线性初始化方法。并且从表中还可以看出,对 GMM 进行参数初始化时不论采用 K-means 初始化还是随机初始化,在相同的聚类初始化下,系统的错误识别率相差不大。这是因为 GMM 的不同初始化方法可能聚集到似然函数的不同局部最大值上,但从说话人识别性能的角度来看,最终模型之间的差异不大。为了使似然函数收敛,上述两种初始化方法也需要大概相同的 EM 迭代次数,大约 10 次迭代就足以满足参数收敛的要求。故对每种初始化方法训练速度也基本相同。这些结果说明对训练高斯混合模型精心挑选初始模型是没有必要的。

2.4　训练语音、测试语音长度对 GMM 识别率的影响

实验在 32 阶 GMM 模型基础上测试了不同训练语音和测试语音长度对系统错误识别率的影响。其中每个说话人分别 60、90、120、150 s 的训练语音,用 1 min 的测试语音,测试语音长度分别为 1、5、10 s。实验结果如表 4 所示。

表 4　不同训练、测试语音长度对 GMM 识别率的影响　　s

训练语音长度	测试语音长度		
	1	5	10
60	70.12	72.14	73.52
90	74.05	78.73	79.01
120	80.21	86.89	87.03
150	81.23	86.91	87.12

从表 4 中可以看出,随着训练语音的增大,识别的性能有较为明显的提高,尤其对较短段长的测试语音,其识别率有更大的提高。当训练语音小于 120 s 时,所有段长的测试语音识别性能有明显幅度的提高;训练语音超过 120 s 后,所有段长的测试语音识别性能也有所提高,但提高幅度较小。虽然增加训练语音长度会提高识别性能,但训练语音的增加同时会加大聚类阶段的运算量及运行时间。基于以上考虑,实验将训练语音长度确定为 120 s。

从表 4 中还可以看出,当测试语音长度分别为 1、5、10 s 时,随着测试数据的增加,识别的性能提高;并且当测试语音段长由 1 s 增加到 5 s 时,识别率有明显提高;当测试语音段长由 5 s 增加到 10 s 时,识别率也有所提高,但提高幅度不大,因此测试语音长度至少在 5 s 以上才能保证较好的识别性能。

3　结束语

本文设计了一个基于改进的说话人聚类初始化和 GMM 的多说话人识别系统。文中提出的改进的线性初始化方法使得对应的 ACP 平均提高了 48.51%,但同时可以看到在整个数据集中采用改进后的方法得到的 ACP 仍然有所偏低,从而使得整个系统的识别率并不高,可见 ACP 有待进一步提高。故在今后的工作中,将考虑引入其他的特征参数如视频特征,来进一步对说话人聚类初始化进行改进,期望得到纯度更高的初始类。

参考文献:

[1] 邓菁. 电话信道下多说话人识别研究[D]. 北京:清华大学,2007.

[2] WOOTERS C,HUIJBREGTS M. The ICSI RT07s speaker diarization system[J]. Multimodal Technologies for Perception of Humans,2008,4625:509-519.

[3] GARAU G,BOURLARD H. Using audio and visual cues for speaker diarisation initialization[C]//Proc of International Conference on Acoustics,Speech and Signal Processing. [S.l.]:IEEE Signal Processing Society,2010:4942-4945.

[4] HUNG H,HUANG Yan, FRIEDLAND G, *et al*. Estimating the dominant person in multi-party conversations using speaker diarization strategies[C]//Proc of International Conference on Acoustics,Speech and Signal Processing. [S.l.]: IEEE Press,2008:2197-2200.

[5] 赵晖,顾亚强,唐朝京. 基于乘积 HMM 的双模态语音识别方法[J]. 计算机工程,2010,36(8):7-9.

[6] FRIEDLAND G,HUNG H,YEO C. Multi-modal speaker diarization of real-world meetings using compressed-domain video features[C]//Proc of International Conference on Audio,Speech and Signal Processing. [S.l.]:IEEE Press,2009:4069-4072.

[7] HUNG H,FRIEDLAND G. Towards audio-visual on-line diarization of participants in group meetings[C]//Proc of Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications. Marseille: European Conference on Computer Vision,2008:1-12.

[8] HUNG H,HUANG Yan, FRIEDLAND G, *et al*. Estimating dominance in multi-party meetings using speaker diarization[J]. IEEE Trans on Audio, Speech and Language Processing,2010,19(4):847-860.

[9] NOULAS A, ENGLEBIENNE G, KROSE B. Multi-modal speaker diarisation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2011,34(1):79-93.

[10] GARAU G,DIELMANN A,BOURLARD H. Audio-visual synchronisation for speaker diarisation[C]//Proc of International Conference on Speech and Language Processing. Makuhari, Chiba:[s. n.],2010:2654-2657.

[11] PARDO J,XNGUERA X,WOOTERS C. Speaker diarization for multiple-distant-microphone meetings using several sources of information[J]. IEEE Trans on Computers,2007,56(9):1212-1224.