基于 MFCC 和运动强度聚类初始化的多说话人识别*

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: 针对常用基于音频特征的多说话人聚类初始化方法精度不高这一问题,提出了一种基于视频信号的新方法。该方法通过运用每一时间帧视频信号的运动强度特征对聚类初始化阶段的初始话者类进行选择,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,在整个会议集上该方法相比其他方法有了很大改善,较之线性初始化系统的错误识别率平均降低了 19.436%,较之改进的线性初始化系统的错误识别率平均降低了 16.618%。

关键词: 多说话人识别; 聚类初始化; 运动强度特征; 运动强度初始化
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 1001-3695(2012)09-3295-04
doi:10.3969/j.issn.1001-3695.2012.09.025

Multi-speaker recognition based on MFCC and motion intensity clustering initialization

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 conventional initialization methods performed on audio feature of multiple speakers clustering with poor accuracy, this paper proposed a new method visual-based feature. The method used motion intensity feature with each time-frame of visual information to find initial speaker cluster during the process of clustering initialization, and promoted the purity of speaker initial cluster effectively. Finally, applied this method to Gaussian mixture model (GMM) multi-speaker recognition system. And the experimental results show that, across the entire meeting set, this proposed new method achieved consistent improvements over other methods, and compared to linear initialization it makes the error recognition of system been reduced by 19.436% on average; 16.618% to the improved linear initialization.

Key words: multi-speaker recognition; clustering initialization; motion intensity feature; motion intensities initialization

0 引言

说话人识别是语音识别中的一个重要课题,而会议室环境下的多说话人识别仍然是目前研究中的热点及重点[1]。在该环境下参会人员围绕一个课题自由交谈,无须考虑说话人切换的时间长度,交谈过程中允许存在语音重叠,且说话人可在会议室内自由移动(甚至离开对应摄像头的检测范围)。据以往研究可知[1~4],对于这种环境下的多说话人识别研究多采用基于音频特征的方法来完成,因为音频特征在说话人识别研究中起主导作用。目前,只有少数的研究者将视频特征应用于说话人识别研究中,其研究结果证明了视频特征无疑对主导的音频特征识别研究起到了很好的辅助作用[5]。

多说话人识别系统的训练阶段分为两部分:a)将预处理后的音频流分成语音和非语音部分(称为语音/非语音检测);b)将由同一个说话人产生的语音段进行归类(称为聚类)。而对聚类这一部分任务最常用的技术为至下而上的层级凝聚聚类法。该方法中初始话者类根据其相似度进行迭代合并,直至满足一定的停止条件。显然,在迭代过程中每个话者类在层级增大,一旦初始类有错误,在后续迭代中不会被修正反而被累积,因此初始类的选择十分重要。

常用的初始化方法为线性初始化[1],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。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,而这必然会导致得到不正确的说话人模型,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。文献[6]提出了一种改进的方法,相比之下有一定的改进但仍不是很理想。为了达到提高初始类纯度的目的,一些研究者提出了各种各样的方法,如 Anguera 等人[7]提出敌友初始化方法;Koh 等人[8,9]利用空间信息,即通过麦克风阵列的声达时间延迟(time delays of arrival,TDoA)分布情况来选择初始话者类并用于后续迭代分割聚类中。可以看到,上述提及的诸多方法都是基于音频特征的,虽然近年来有人引入视频信号特征将其用于说话人识别研究[5],但主要用于聚类阶段[10],目前还未出现将视频特征用于聚类初始化的研究。其他领域的一些研究表明,人在交谈的过程中其语音的产生会更多地伴随着手势或其他姿体运动等,

收稿日期: 2012-01-04; **修回日期**: 2012-02-21 **基金项目**: 甘肃省自然科学基金资助项目(1014ZSB064);甘肃省财政厅资助项目(0914ZTB148)

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、目标跟踪;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).

而视频特征中的运动强度特征正是以此为出发点来判断当下谁是最有可能的说话人,从而找到初始话者类。

因此,鉴于视频信号的有效性及良好的辅助性,本文考虑在说话人聚类初始化中引入视频信号的运动强度特征对初始话者类进行有效选择,从而进一步提高初始类纯度,最后在 GMM 的多说话人识别系统中进行验证。

1 研究原理

类似于多说话人识别系统[6],与之相比本次实验设计的主要差异为:在训练阶段的说话人聚类初始化过程中利用视频信号的运动强度特征进行初始话者类的选择,即采用运动强度特征初始化方法,然后将得到的初始类用于后续的语音迭代聚类过程。其实验原理如图 1 所示。

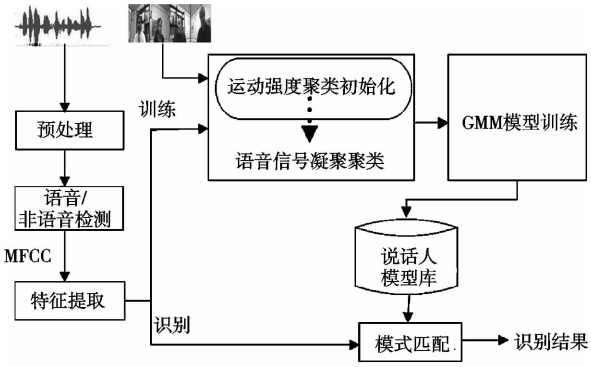

图1 实验原理

1.1 训练阶段

1.1.1 音频特征参数

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。其中预加重的目的是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。本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 的计算式为

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

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 视频特征提取

从四个近距离摄像头中分别提取每个人的运动强度特征,并将其作为视频特征参数。该特征通过计算给定的时间帧内视频序列对应的连续灰度图像的像素差,并通过均值滤波进行平均化而得到[11]。使用均值滤波是为了降低背景噪声,以及当说话人出现较大幅度动作时产生的各种影响。

运动强度特征 $\mathrm{fmov}(t)$ 的计算表达式为

$$\mathrm{fmov}(t)=\frac{\sum_{i=1}^{t-1}\left|I(x,y,i+1)-I(x,y,i)\right|}{\sum_{i=1}^{t-1}I(x,y,i)} \tag{2}$$

其中:$I(x,y,i)\ (i=1,2,\cdots,k)$ 是从检测现场拍摄到的画面中提取的一组视频图像;$\sum_{i=1}^{t-1}I(x,y,i)$ 为时间帧 t 内视频序列对应灰度图像总的像素值。

在特征提取的过程中是对整个完整的帧图像进行分析。这样做一方面可以避免人脸检测计算的复杂性,相比之下更具鲁棒性和有效性,同时还可以捕捉到说话人手势的动作等;另一方面也考虑到人在说话的过程中会伴随有更多的姿体动作这一事实。并且与文献[10]中将运动强度特征和音频 MFCC 特征提取后单独建模最后在聚类阶段进行结合相比,本次实验将运动强度特征用于聚类初始化,应用范畴不同,且模型相对简单得多。

1.1.3 运动强度特征初始化

在说话人聚类初始化阶段,依据人在交谈中语音的产生伴随着更多的手势及身体动作这一事实,本文采用视频的运动强度特征对初始的话者类进行选择,即在每一个时间帧将运动强度最大的摄像头对应的说话人作为初始话者类。在说话人聚类的初始化过程中,当参会人员暂时不在与之对应的摄像头拍摄范围内(移动至屏风或讲解板前)时将检测不到该说话人的运动,此时解决办法是引入一个新的初始话者类。当然也可通过检测出现在屏风或讲解板前的说话人的运动强度来解决这一问题,但在这种情况下对摄像头的分辨率还不足以检测出话者的脸部运动。

这样每一个时间帧产生了一个初始类,最后将得到的所有初始话者类进行聚类,得到相应的说话人模型。

当采用线性初始化时是将等长的语音数据平均分配给每一个初始话类,此时每个类用于后续建模的高斯分量的初始值是相同的。而应用运动强度特征来选择初始话者类时,语音数据的分布不再平均甚至变化很大,因此有必要依据一个准则来确定每个话者类模型的初始复杂度即高斯分量数。Luque 等人[9]采用每个高斯分量语音帧的数目作为准则。本文采用贝叶斯信息准则(Bayesian information criterion,BIC)进行判定,其表达式为

$$\mathrm{jinit}=\arg\max\left[\log L(X_i|M_j)-\lambda\frac{1}{2}j\log(N_i)\right] \tag{3}$$

其中:X_i 是话者类 i 的语音数据;M_j 是话者类 i 的语音数据 X_i 对应的高斯模型且其分量数为 j;N_i 为语音数据 X_i 的帧数;λ 为惩罚因子(实验中取值为 10)。

1.2 识别部分

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似然度,根据最大后验概率准则将概率最大的模型对应的说话人作为识别结果。

2 实验数据及结果分析

2.1 实验环境与数据

本次实验用于模型训练以及识别的预料数据来自于 AMI 语料库中视听会议的一部分[1]。

用于实验的语料数据是在一个安置有同步记录设备(麦克风、摄像头)的会议室事先录制而成,会议环境如图 2 所示。参会人员一般为四人,围绕某个主题,各自分配角色,自由交

谈。利用麦克风阵列获得实验的音频信号;利用四个近距离摄像头捕捉参会人员的视频运动行为特征;利用远距离摄像头捕捉头部姿势(该特征不用于特征提取)。会议场景分为动态和静态会议。其中静态会议中参会人员整个过程中不离开各自座位;动态会议则不限制,人员可在室内自由移动。实验中选取了七场平均长度为 20 min 的动静态会议进行仿真实验,会议编号如表 1 所示(七场会议主题一致,参加人员相同)。

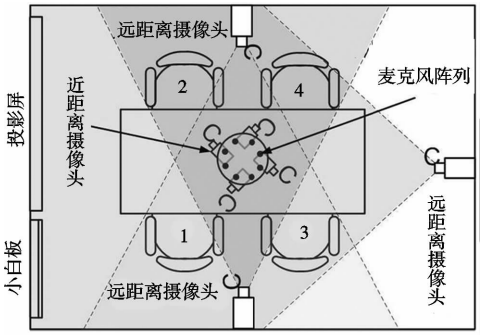

图2　实验环境

表 1　实验使用的视听会议编号

会议编号
IS1001b
IS1001c
IS1003d
IS1006b
IS1006d
IS1008a
IS1008c

说话人识别系统的性能根据系统的识别错误率(recognition error rate,RER)进行评测[1]。RER 定义为语音/非语音错误(语音信号被检测为非语音信号,或非语音信号被检测为语音信号)和说话人错误率(正确检测出的语音信号分配给了错误说话人的帧数所占比例)之和(RER = SpNsp + Spkrerr)。文中采用的语音/非语音检测器在各个实验中的平均错误检测率是一样的,因此系统的识别性能主要与说话人错误率相关。

2.2　运动强度聚类初始化以及其他初始化方法对应 ACP 的比较

在初始话者类形成后,为了检测所提初始化方法的性能,通过计算平均类纯度(average cluster purity,ACP)[12,13]来完成。ACP 的概念由 Ajmera 等人于 2002 年提出,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 的计算式为

$$\mathrm{ACP} = \frac{1}{N}\sum_{i=1}^{N_c} p_i n_i \tag{4}$$

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

$$p_i = \sum_{j=1}^{N_s} (n_{ij}/n_i)^2 \tag{5}$$

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 选中的七场会议首先采用运动强度聚类初始化方法得到每个会议的 ACP,如表 2 所示。为了进一步说明所提新方法的性能,将其与线性初始化以及改进的线性初始化方法进行比较,得出结果如图 3、4 所示。

从图 3 可以看出,运动强度初始化和改进的线性初始化方法明显优于传统的线性初始化方法,其中采用运动强度聚类初始化得到的平均类纯度最高。线性初始化是将语音数据平均分配给每个初始类,这种等长的平均分配极有可能导致“不纯净”的初始类,故平均类纯度并不高;而改性的线性初始化是在线性初始化的基础上引入 BIC 对可能“不纯净”的初始类进行再分割,显然可以在一定程度上提高话者类的纯度,但该方法不能很好地判断哪些初始类“不纯净”;文中所提的运动强度初始化方法则是将每一时间帧内对应摄像头中运动强度最大的人作为一个初始类,这样很好地避免了线性初始化方法容易产生包含有两个说话人的“不纯净”初始类的出现,且可视的视频信号也可作为一个参考。同时还可以看到运动强度初始化对动态会议的影响较大(如会议 IS1006b 等),这也进一步说明了在动态会议中视频信号可以提供更有效的辅助作用。

表 2　运动强度聚类初始化对应的 ACP

会议编号	运动强度聚类初始化
IS1001a	0.30
IS1001c	0.32
IS1003b	0.26
IS1003d	0.25
IS1008a	0.51
IS1008d	0.52
IS1006b	0.50

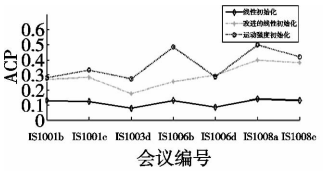

图3　不同初始化方法对应的ACP

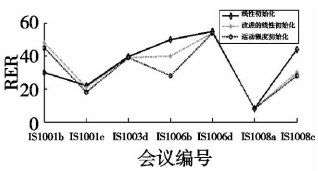

图4　不同初始化方法对应的错误识别率

从图 4 可以看出,对于大多数会议场景,采用相应初始化方法得到的 ACP 越高,其对应的系统错误识别率也就相对越低,这也说明了聚类初始化纯度的高低直接影响系统的识别性能,由此可见初始类纯度的重要性。虽然在有些会议场景中即使上阶段得到的 ACP 有一定差距,但系统的错误识别率却相差甚小(如会议 IS1008a),这是由于尽管聚类初始化对系统性能影响较大,但同样也受聚类过程的影响,因为运动强度初始化会引入一些新初始类(当检测不到说话人时),且该方法使得模型训练时初始模型复杂度有差异,因此也会受到影响。总之,新方法相对于整个会议集其系统错误识别率是最高的。

2.3　不同初始化方法在整个会议集上的平均错误识别率

为进一步说明不同初始化方法对所选七场会议(包括静态会议和动态会议)的平均系统错误识别率的影响,得出如表 3 所示的结果。

表 3　不同初始化方法在整个会议集上的平均错误识别率(RER)

会议集	初始化方法		
	线性初始化	改进的线性初始化	运动强度初始化
所有会议	35.5	34.3	28.6
[静态,动态]	[15,43.7]	[18,42.2]	[13,38.8]

从表 3 可以看出,对于整个会议集,运动强度聚类初始化方法对应的平均系统错误识别率最低,较之线性初始化方法平均 RER 降低了 19.436%,较之改进的线性初始化方法平均 RER 降低了 16.618%。同时还可以看到对于静态会议和动态会议有同样的效果。

3 结束语

本文设计了一个基于 MFCC 和运动强度特征聚类初始化的多说话人识别系统。为了进一步提高初始话者类的纯度,提出了一种基于视频信号的运动强度初始化的新方法。该方法将视频信号用于说话人聚类初始化,较常用的线性初始化及改进的线性初始化方法有较大改进。然而该方法同时也引入了一些新初始类(当检测不到说话人时),并且使得模型训练时初始模型复杂度各有差异,这将是下一步研究工作需要考虑的问题。

参考文献:

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

[2] 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. 2008: 2197-2200.

[3] HUNG H, HUANG Yan, FRIEDLAND, *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.

[4] NOULAS A K, ENGLEBINNE G, KROSE B J A. Multi-modal speaker diarisation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34(1): 79-93.

[5] HUNG H, JAYAGOPI D, YEO C, *et al*. Using audio and video features to classify the most dominant person in a group meeting[C]//Proc of the 15th International Conference on Multimedia. New York: ACM Press, 2007: 835-838.

[6] 曹洁,余丽珍. 改进的说话人聚类初始化和 GMM 的多说话人识别[J]. 计算机应用研究, 2012, 29(2): 590-593.

[7] ANGUERA X, WOOTERS C, HERNANDO J. Friends and enemies: a novel Initialization for speaker diarization[C]//Proc of the 9th International Conference on Spoken Language. 2006: 689-692.

[8] KOH E C, SUN Han-wu, NWE T L, *et al*. Speaker diarization using direction of arrival estimate and acoustic feature information[C]//Proc of Multimodal Technologies for Perception of Humans. Berlin: Springer-Verlag 2007: 484-496.

[9] LUQUE J, SEGURA C, HERNANDO. Clustering initialization based on spatial information for speaker diarization of meetings[C]//Proc of the 9th Annual Conference of the International Speech Communication Association. 2008: 383-386.

[10] GARAU G, BA S, BOURLARD H, *et al*. Investigating the use of visual focus of attention for audio-visual speaker diarisation[C]//Proc of the 17th ACM International Conference on Multimedia. New York: ACM Press, 2009: 681-684.

[11] ZOBL M, WALLHOFF F, RIGOLL G. Action recognition in meeting scenarios using global motion features[C]//Proc of the 4th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. 2003: 32-36.

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

[13] 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.