广播电视新闻中的主持人跟踪系统①

汪 洋[1], 甘 涛[1], 向 军[2]

[1](电子科技大学 电子工程学院, 成都 611731)
[2](湖南人民广播电台 技术部, 长沙 410007)

摘 要: 针对广播电视新闻节目中的主持人跟踪问题, 提出了一种将说话人分割聚类和说话人确认有效结合的算法, 并根据该算法设计了一套主持人跟踪系统. 该系统首先利用音频活动检测算法去除新闻音频资料中的静音段, 然后说话人分割聚类算法将多说话人语音段分成若干单一话者语段, 最后通过基于 GMM-UBM 的说话人确认算法辨认每段单一话者语段的话者身份是否为目标主持人. 此外, 分析了 T-Norm 对系统性能的影响. 以中央电视台《新闻联播》为评测数据集, 实验结果表明, 该算法取得了良好的效果, 跟踪系统的查准率(Precision)和查全率(Recall)分别为 93.03%和 84.34%.

关键词: 广播电视新闻; 说话人分割聚类; 说话人确认; 主持人; 跟踪

Anchor Speakers Tracking in Broadcast News

WANG Yang[1], GAN Tao[1], XIANG Jun[2]

[1](School of Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)
[2](Technology Department, Hunan Broadcasting System, Changsha 410007, China)

Abstract: To address the problem of anchor speakers tracking in broadcast news, an algorithm that integrated effectively speaker segmentation and clustering and speaker verification is presented. An anchor speakers tracking system based on the proposed algorithm is also designed. The system firstly uses audio activity detection algorithm in order to remove silence segments. Secondly, a speaker segmentation and clustering process converts multi-speaker speech waveform into several speaker homogenous segments. Finally, speaker verification approach based on GMM-UBM is dedicated to decide whether a segment belongs to one targeted anchor speaker. Furthermore, the impact of T-norm on system performance is also analyzed. Experiments on CCTV Mandarin Broadcast News demonstrate the effectiveness of the proposed algorithm. The tracking system achieves precision and recall at 93.03% and 84.34% respectively.

Key words: broadcast news; speaker segmentation and clustering; speaker verification; anchor speakers; tracking

给定一段录音资料和一个目标说话人, 说话人跟踪指的是在录音资料中含有目标说话人的前提下, 定位该话者的说话时间段. 其中, 录音资料中说话人的人数和身份均未知. 主持人作为新闻中重要的说话人, 对其进行跟踪在新闻故事单元分割、新闻语音文档检索及新闻音视频信息分析中均有着十分重要的作用. 法国的 AFCP 等机构于 2003-2005 年间组织了针对法语广播电视新闻 RT(Rich Transcription)的 ESTER 评测活动, 其中就包括新闻中的说话人跟踪评测任务[1]. Dan Istrate 等人[2]在本次评测活动中提出的说话人跟踪算法与本文类似, 该算法先利用声学场景分割技术将广播新闻分成语音、电话语音、含有音乐背景的语音、音乐四个类别, 然后使用基于 EHMM 的分割聚类算法将语音段(语音、电话语音、含有背景的语音)分割成多个单一话者语段, 最后对每个语音段进行说话人识别. 但是, 该算法对新闻音频分类过细, 会导致累积

① 收稿时间:2014-02-10;收到修改稿时间:2014-03-17

误差, 且计算成本大. 文献[3-5]均使用了与本文类似的算法, 但过于依赖说话人分割聚类提供的信息(边界和聚类), 整个算法性能受说话人分割聚类的影响较大. Viet 等人[6]提出了基于 GSV-SVM 说话人识别的说话人跟踪算法, 且在 ESTER-2 测试数据集上平均 F-measure 达到了 86.7%. 但实验表明 GMM-UBM 和 GSV-SVM 说话人辨认系统性能相当, 且 GSV-SVM 系统在识别过程中需要进行 GMM-Supervector 的提取, 所需时间较长. 此外, GSV-SVM 系统的优势在于降低噪声和信道作用对识别结果的影响, 缺点是减弱了能表征说话人特性的帧向量的贡献. 对于广播电视新闻而言, 由于噪声和信道作用影响较小, 因此使用基于 GMM-UBM 的说话人识别算法更为合适. 曹洁等人[7]针对会话语音中说话人聚类初始化精度问题, 提出了一种改进的聚类初始化方法并将其应用到多说话人识别中, 有效地降低了识别的错误率.

根据上述分析可见, 目前说话人跟踪主要使用先分段再识别的方法, 许多研究人员对这两部分算法分别进行了改进, 但未对整个系统进行综合考虑. 本文分析了说话人分割聚类和说话人识别在说话人跟踪中的作用, 改进了说话人分割聚类算法并将其与说话人确认有效结合, 提出了广播电视新闻中的主持人跟踪算法并设计了一个低复杂度的说话人跟踪系统. 此外, 根据实际应用中会出现的问题, 对说话人识别结果进行后处理以提高系统的跟踪性能. 实验结果表明, 该系统能够在广播电视新闻中实现主持人的有效跟踪.

1 音频活动检测

音频活动检测主要有两方面的作用: 1)去除音频中的无用信息, 如静音段, 这样可以使训练语音数据较为纯净; 2)为说话人分割聚类提供边界信息.

本文使用了一种基于双高斯(Bi-Gaussian)模型[8]的音频活动检测算法. 该方法首先采用音频帧对数能量和 EM(Expectation Maximization)算法[9]估计出分别代表高能量帧和低能量帧分布的高斯模型的均值和方差, 然后根据最大似然比准则进行音频帧属性的判决, 最大似然比可表示为:

$$\begin{aligned} llr(x) &= \ln\frac{p(x|w_h)}{p(x|w_l)} = \ln p(x|w_h) - \ln p(x|w_l) \\ &= \frac{\sigma_h^2-\sigma_l^2}{2\sigma_h^2\sigma_l^2}x^2 + (\frac{u_h}{\sigma_h^2} - \frac{u_l}{\sigma_l^2})x + (\frac{u_l^2}{2\sigma_l^2} - \frac{u_h^2}{2\sigma_h^2}) \end{aligned} \quad (1)$$

$$+\frac{1}{2}\ln\frac{\sigma_l^2}{\sigma_h^2} \gtrless 0 \to x \in \begin{cases} w_h \\ w_l \end{cases}$$

其中, x 是输入帧对数能量, w_h 和 w_l 分别代表高能量帧和低能量帧这两类, 且 $p(x|w_h)\sim N(u_h,\sigma_h^2)$, $p(x|w_l)\sim N(u_l,\sigma_l^2)$. 如果某音频帧被判决为低能量帧, 则该帧即为静音帧并被丢弃.

由于该方法是在帧层次上进行的判决, 这样对于较短的音频段(只有几十帧)可能会出现不理想的结果. 为了避免此种现象发生, 我们对静音段的最短时间做了一定的限制, 其中训练阶段和测试阶段最小静音段时间分别设置为 0.5s 和 0.3s. 此外, 音频帧长为 20ms, 帧移为 10ms.

2 说话人分割聚类

2.1 说话人分割

说话人分割(Speaker Segmentation)[10]是指在连续的语音中找到话者跳变点, 从而将属于不同说话人的语音段分割开来. 基于距离的分割方法由于不需要说话人的任何先验知识, 且估计的参数少、计算代价低、速度快, 因而常用在说话人分割中.

本文采用了基于贝叶斯信息准则(Bayesian Information Criterion, BIC)的距离分割方法. BIC 实际上是一种最优贝叶斯模型搜索准则, 它可以用来判决哪一类模型能够最好地表示给定数据的概率分布. 该方法利用一个滑动窗以固定的步长在音频帧上滑动, 如果两个不同的高斯分布比单个高斯分布可以更好地刻画滑动窗内的数据, 则表明此滑动窗内含有话者跳变点, 否则无跳变点.

已知样本 x_i 为 d 维的特征向量. 假设两个相邻的分析窗 X 和 Y 位于时间 t_j 左右两侧, 令 $Z = X \cup Y$, 问题是检测 t_j 处是否有一个话者跳变点. 这是一个两类统计决策问题.

假设 H_0 下, t_j 处没有话者跳变点, Z 中的样本序列可以使用单一的多元高斯模型 θ_z 描述. 对数似然函数 L_0 可表示为:

$$L_0 = \sum_{i=1}^{n_X}\log p(x_i|\theta_Z) + \sum_{i=1}^{n_Y}\log p(y_i|\theta_Z) \quad (2)$$

其中, n_X 和 n_Y 分别是分析窗 X 和 Y 中的样本帧数.

假设 H_1 下, t_j 处有话者跳变点, 分析窗 X 和 Y 中的样本分别用多元高斯模型 θ_X 和 θ_Y 描述. 对数似然函数 L_1 可表示为:

$$L_1=\sum_{i=1}^{n_X}\log p(x_i\mid\theta_X)+\sum_{i=1}^{n_Y}\log p(y_i\mid\theta_Y) \tag{3}$$

H_0 和 H_1 的似然比定义为:

$$R=L_1-L_0=\frac{n_Z}{2}\log|\Sigma_Z|-\frac{n_X}{2}\log|\Sigma_X|-\frac{n_Y}{2}\log|\Sigma_Y| \tag{4}$$

其中, Σ_Z, Σ_X, Σ_Y 分别为 Z, X, Y 中样本的协方差矩阵(本文使用的是对角阵), $n_Z=n_X+n_Y$ 是 Z 中样本的帧数. 因此 H_0 和 H_1 的 BIC 的差定义为:

$$\Delta BIC=R-\frac{\lambda}{2}\left(d+\frac{d(d+1)}{2}\right)\log n_Z \tag{5}$$

其中, λ 是惩罚因子, 本文设置为 1.5. 如果 $\Delta BIC>0$, 则表明 H_1 成立, 即在 t_j 处有话者跳变点, 否则 H_0 成立, 即在 t_j 处无话者跳变点.

根据以上原理, 本文采用了多层次的说话人分割算法, 可以分为以下三个步骤:

1)检测候选话者跳变点. 在一个固定长度的较大的分析窗里用两个不断变化的数据窗查找分析窗里的每一个位置 i, 如果 $\max_i\Delta BIC(i)>0$, 则 i 是一个候选话者跳变点. 此时将分析窗的起始位置设为 i, 继续检测. 如果没有检测到话者突变点, 那么就增加分析窗长(每次增加 0.6s), 继续寻找, 直到找到为止. 由于本文针对的是单通道的广播电视语音, 话者转换的速度不是很快, 因此设置分析窗长为 3s. 这里采用的分析窗和窗增长步长均较大, 因此是候选点的一个粗估计;

2)确认候选话者跳变点. 使用一个稍小的分析窗和增长步长进行检测, 且分析窗长的中心固定在候选点上. 利用(1)中的方法进行检测, 如果分析窗内检测到跳变点, 则该跳变点作为新的跳变点; 如果没有检测到跳变点, 则丢弃该窗内的候选跳变点. 这里采用的分析窗长为 2s, 增长步长为 0.2s;

3)判决有效话者跳变点. 根据第(2)步进行的分段, 计算连续两个分段的 ΔBIC. 如果 $\Delta BIC>0$, 则保留这两个分段; 如果 $\Delta BIC<0$, 则合并这两个分段.

Moattar 等人[10]指出预分割有利于提高 BIC 分割的精度, 本文的分割正是在音频活动检测的基础上进行的, 而音频活动检测可以视为一种预分割技术. 同时本文使用了一个分析窗代替两个连续的分析窗, 有利于提高算法的鲁棒性.

2.2 说话人聚类

对于说话人分割后的语音, 进一步的工作是在说话人数目未知的情况下, 将一段语音中由同一说话人发出的语音聚合起来. 本文采用了基于自底向上[10]的层次聚类算法, 并用 BIC 准则(同 2.1)作为停止聚类的条件. 具体算法流程如下:

(1)根据说话人分割的结果, 每个语音段单独作为一个聚类;

(2)计算聚类集中的聚类两两之间的 ΔBIC, 得到距离矩阵;

(3)找出距离矩阵中最小的 ΔBIC;

(4)如果(3)中的 $\Delta BIC<0$, 则将其对应的两个聚类合并为新的聚类, 更新聚类集, 跳转到步骤(2);

(5)迭代(2), (3), (4)步直到所有的 $\Delta BIC>0$, 此时聚类停止, 得到最终的聚类集, 并将属于同一聚类且时间上相邻的两个音频段合并.

该算法中惩罚系数 λ 设置为 1.5. 说话人分割聚类所用的特征参数均为 16 阶 MFCC 加上对数能量及其一阶差分, 即特征参数共计 34 维($d=34$). 帧长为 20ms, 帧移为 10ms, 预加重系数为 0.975, 窗函数为汉明窗(Hamming Window), 每帧语音使用的 FFT 大小为 512, 截止频率为 300Hz~8000Hz.

3 基于GMM-UBM的说话人确认

说话人确认(Speaker Verification)针对单个用户, 判断测试语音是否来自所声明的用户身份. 实际应用中, 训练语音数据往往很少(几十秒), 因而很难覆盖所有的声学现象, 因此本文采用基于 GMM-UBM 的说话人确认算法以克服该问题. 一个完整的说话人确认系统往往包括两个阶段: 训练阶段和测试阶段, 如图 1 所示.

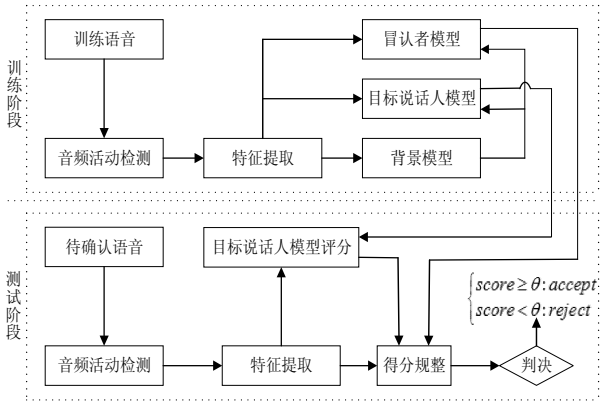

图 1　说话人确认系统框图

3.1 特征提取

与 2 中所用的特征参数相同, 并且使用了方差均值归一化, 使每个语音文件提取的 MFCC 特征的均值为 0, 方差为 1, 该方法可以提高说话人识别的鲁棒性.

3.2 背景模型训练

全局背景模型(Universal Background Model, UBM)本质上是一个大型的 GMM 模型, 它是使用所有待识别说话人的训练语音通过 EM/ML 算法训练得到的, 反映了所有待识别说话人的特征分布特征. 本文所使用的背景模型阶数为 512, 协方差矩阵为对角阵.

3.3 目标说话人模型训练

与传统的借助 EM 算法训练 GMM 不同, 每个说话人的 GMM 模型是从 UBM 模型中贝叶斯自适应得到的, 从而大大减少了训练时间和需要的数据量. 本文仅对均值采用了 EM/MAP 自适应, 目标说话人模型阶数为 512, 协方差矩阵为对角阵.

3.4 目标说话人模型评分

由于目标说话人模型是从 UBM 模型自适应得到的, 所以每个说话人模型可以共享 UBM 模型的高斯分量, 而对于一个特征向量而言, 仅有几个高斯分量对概率值贡献较大, 为此计算语音段在目标说话人模型上的概率时, 可首先从 UBM 中选取前 10 个最佳的高斯分量, 然后利用目标说话人模型中相对应的 10 个高斯分量, 计算语音段中的每一帧在目标说话人模型上的概率. 该方法可以在保持识别率不下降的前提下, 大大提高说话人模型分数计算效率.

给定一段待测试语音, 经过预处理和特征提取后, 得到特征序列 $Y=\{\vec{y}_1,\vec{y}_2,\cdots,\vec{y}_T\}$, 那么某一帧特征向量 $\vec{y}_i\ (1\le i\le T)$在目标说话人模型 x 上的对数似然概率为:

$$LLR(\vec{y}_i)=\log p(\vec{y}_i\mid X)-\log p(\vec{y}_i\mid W) \tag{6}$$

其中: W表示背景模型. 特征序列 Y的最终得分为每帧对数似然概率的数学平均数, 可以表示为:

$$LLR(Y)=\frac{1}{T}\sum_{i=1}^{T}LLR(\vec{y}_i) \tag{7}$$

因此, 就可以得到待测试语音在目标说话人模型上的得分.

3.5 冒认者模型

冒人者模型是一种说话人模型, 采用除目标说话人外的其他话者语音数据训练得到, 通常用于得分规整中.

3.6 得分规整

本文使用了 T-Norm 分数规则方法[11]. T-Norm 的基本思想是对每条测试语音计算其对冒认者模型集合得分, 得到对应于不同测试语音的冒充者模型集合得分分布, 来消除由于测试语音环境不同对打分分布的影响. 假设对冒认者模型集的得分分布服从高斯分布, 计算其分布均值 μ_{imp} 和方差 σ_{imp}, 则规整后的测试语音得分可以表示为:

$$S'=\frac{S-\mu_{imp}}{\sigma_{imp}} \tag{8}$$

其中, S、S' 分别表示分数规整前后的得分.

3.7 判决

对每条测试语音进行得分规整后, 便可根据给定的阈值判断测试语音说话人是否为目标说话人. 如果测试语音得分大于阈值, 则判决该语音段说话人是目标说话人, 反之则不是. 该阈值与说话人性别无关, 可以在实验数据集上估计得到.

4 广播电视新闻中的主持人跟踪系统

本文提出的主持人跟踪系统实际上是音频活动检测、说话人分割聚类系统和说话人确认系统三者的有效结合, 如图 2 所示.

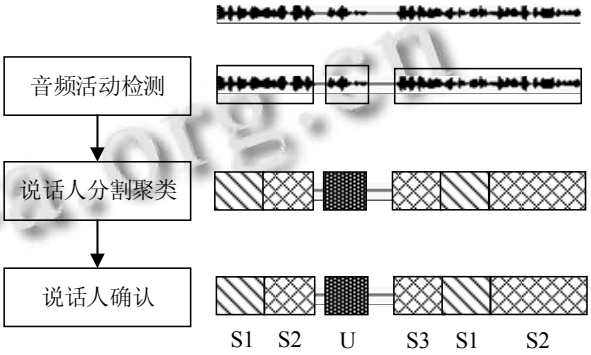

图 2 主持人跟踪系统流程图

如图 2 所示, 测试新闻音频经过去除静音以及分割聚类后, 便可利用说话人确认算法和提前训练的主持人模型为分割聚类后的每一个音频段独立地进行说话人身份的判断. 通常情况下, 在计算分割聚类后每个音频段的最终得分时, 主要有两种方法: 1)基于聚类(Cluster-based), 指的是充分利用说话人分割聚类提供的边界和聚类信息, 先对分割聚类后的每个音频段打分, 然后计算属于同一话者的所有音频段得分的平均

分并将其作为这些音频段的最终得分; 2)基于分段(Segmentation-based), 指的是只利用说话人分割聚类提供的边界信息, 独立地对分割聚类后的每个音频段打分并将其作为每个音频段的最终得分. 与说话人确认相比, 说话人分割聚类错误率较高. 从图 2 中可以看出, 相同聚类的语音段并不一定是同一个说话人所说, 例如 S3 和 Unknown 在说话人分割聚类阶段为同一个聚类, 但在说话人确认阶段才被正确识别出. 因此本文采用了基于分段的方法.

实际应用中, 还可能存在一些问题, 本系统提出了相应的解决措施.

1) 分割聚类后会出现若干较小的音频片段(0.05s~0.3s), 实验证明, 当待测试的音频段时长在 0.5s 以下时, 本文使用的说话人确认系统性能会下降. 一般情况下, 较小的语音片段不属于主持人. 因此, 当音频段低于 0.3s 时, 则丢弃该段, 不会对其进行说话人确认;

2)主持人的某一段语音可能被分割为相邻的若干不连续的语音段, 因此会对检测之后的语音段进行拼接处理. 如果后一个音频段的开始时间和前一个音频段的结束时间相差低于 0.5s, 那么将这两个音频段合并为一个新的音频段, 其开始时间和结束时间对应前一段音频的开始时间和后一段音频的结束时间.

5 实验与分析

5.1 实验数据

为了评测提出的主持人跟踪算法及系统性能, 本文利用 FFMEG 提取了来自中央电视台《新闻联播》、《焦点访谈》、《晚间新闻》、《朝闻天下》四个新闻节目中的音频文件, 均为 wav 格式, 其中的音频数据格式为单声道, 24kHz 的采样率, 16 位精度, mu-law 压缩.

实验中 UBM 的训练数据来自 2012 年 8 月 20 日至 2012 年 9 月 30 日的《新闻联播》音频资料, 手工去除了音频中的所有音乐、噪声片段(可能存在残留), 最终保留了较为纯净的说话人语音文件, 时长共计 19 小时 6 分钟. 系统使用了一个 UBM, 且与性别无关. 为了保证冒认者语音与目标说话人的无关性, T-Norm 数据集来自《焦点访谈》、《晚间新闻》、《朝闻天下》. 本系统使用了两个性别有关的男女冒认者集, 男冒认者集训练数据时长共计 2 小时 26 分钟, 女冒认者集训练数据共计 1 小时 40 分钟, 冒认者集与目标主持人性别保持一致. 目标说话人是《新闻联播》全部 10 位主持人(男、女各 5 人), 每个主持人模型的训练数据时长为 2 分钟, 均来自 2012 年 10 月 1 日至 2012 年 10 月 15 日的《新闻联播》. 测试音频为 2012 年 10 月 16 日至 2012 年 10 月 31 日的《新闻联播》, 共计 16 段音频文件合计 8 小时, 每段时长为 30 分钟且包含男女主持人各一名.

5.2 评测指标

说话人分割聚类方法性能的评估以最终的切分错误率(Diarization Error Rate, DER)作为衡量标准, DER 主要包括三部分: 漏检(Missed Speech, Mi)、虚警(False Alarm Speech, Fa)、说话人分类错误(Speaker Match Error, Spk).

基于 GMM-UBM 的说话人确认算法性能评估手段包括检测错误折中(Detection Error Tradeoff, DET)曲线、等错误率(Equal Error Rate, EER)、最小检测代价函数(Minimum Detection Cost Function, minDCF).

主持人跟踪系统性能评测指标包括查全率(Recall)、查准率(Precision)、F-measure.

$$\text{查准率}=\frac{\text{正确检出的目标说话人语段时间}}{\text{检出的说话人语段总时间}}\times 100\%$$

$$\text{查全率}=\frac{\text{正确检出的目标说话人语段时间}}{\text{目标说话人语段的实际总时间}}\times 100\%$$

$$F-Measure=\frac{2\times R\times P}{R+P}$$

5.3 实验结果与分析

5.3.1 说话人分割聚类

说话人分割聚类算法在测试音频上的实验结果如表 1 所示.

表 1　说话人分割聚类实验结果(%)

Mi	Fa	Spk	DER
0.5	0.1	17.1	17.7

从表 1 中可以看出, 漏检和虚警都很小, 这是因为在分割聚类前端引入了音频活动检测算法. 说话人分类错误(把本属于同一个的语音标记为其他话者所说)相对较大, 这也是跟踪系统中采用基于分段的方法计算音频段最终得分的主要原因.

5.3.2 基于 GMM-UBM 的说话人确认

说话人确认实验中, 男性(Male)主持人和女(Female)主持人分开测试, 且给出了 T-Norm 对确认结果的影响. DET 曲线如图 3 所示, EER、minDCF 数据

如表 2 所示.

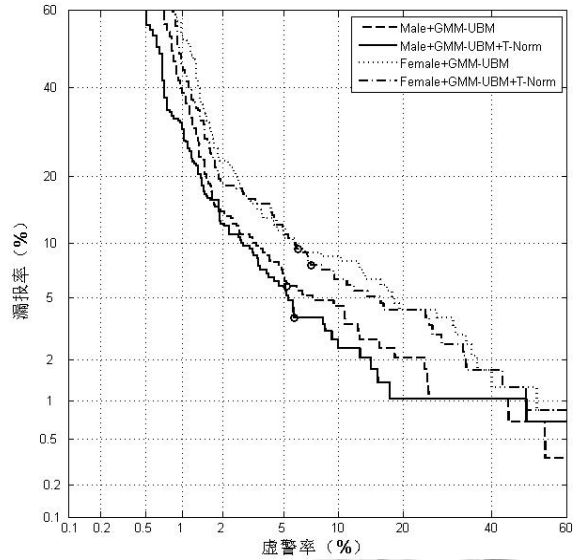

图 3 基于 GMM-UBM 的说话人确认系统的 DET 曲线

表 2 说话人确认系统 EER、minDCF 结果(%)

测试类型	EER	minDCF
Male	5.86	5.52
Male+T-Norm	5.19	4.76
Female	10.17	8.52
Female+T-Norm	8.91	8.24

从图 3 和表 2 可以看出: (1)男性主持人的识别错误率要低于女性主持人; (2)T-Norm 在男女性主持人跟踪时均可以提高系统性能: 男性主持人得分经过 T-Norm 后, EER 相对提高 12.9%, minDCF 相对提高 15.9%; 女性主持人得分经过 T-Norm 后, EER 相对提高 14.1%, minDCF 相对提高 3.4%; (3)T-Norm 对系统性能的提高不是很大, 主要是由于本系统是在广播电视新闻音频上进行测试, 而新闻音频中的信道失配现象很少. 总的来说, 说话人确认系统准确率较高, 也证明了说话人分割聚类的有效性.

5.3.3 主持人跟踪系统

经过对说话人确认系统检测出的目标主持人语音片段的处理之后, 得到了最终的跟踪结果. 系统的评测结果如表 3 所示.

表 3 主持人跟踪系统评测结果

查全率	查准率	F-Measure
93.03	84.34	88.47

如表 3 所示, 主持人跟踪系统经过合并相邻段等处理后, 整个系统跟踪性能较好, 基本可以满足广播电视新闻中的主持人查找、索引等需求.

6 结语

本文设计了一个完整的广播新闻节目中主持人跟踪系统, 它主要由三个部分组成: 音频活动检测、说话人分割聚类、基于 GMM-UBM 的说话人确认. 文章对这三部分均进行了理论分析并给出了相应的实验结果, 在此基础上, 系统实现了三者的有效结合. 通过测试发现, 系统在广播电视新闻上表现出了较好的跟踪性能. 此外, 文章还分析了分数规整对于系统性能的影响及其适用范围. 然而, 整个系统性能在很大程度上依赖于说话人分割聚类的准确率, 同时当前分割聚类的精度仍然有待于进一步提高. 因此, 如何对音频进行准确的分割聚类及其与说话人识别更为有效的结合将是下一步研究工作中需要考虑的问题.

参考文献

1 Galliano S, Gravier G, Bimbot F. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. Interspeech. 2009. 2583–2586.
2 Istrate D, Scheffer N, Fredouille C, Bonastre JF. Broadcast news speaker tracking for ESTER 2005 Campaign. European Conference on Speech Communication Technology. 2005. 2445–2448
3 Docio-Fernandez L, Garcia-Mateo C. Speaker segmentation, detection and tracking in multi-speaker long audio recordings. Third COST275 Workshop on Biometrics on the Internet. 2005. 97–100.
4 Zibert J, Vesnicer B, Mihelic F. A system for speaker detection and tracking in audio broadcast news. Informatica (Slovenia), 2008, 32(1): 51–61.
5 Huijbregts M, Leeuwen DAV. Large-scale speaker diarization for long recordings and small collections. IEEE Trans. on Audio, Speech and Language Processing, 2012, 20(2): 404–413.
6 Le VB, Barras C, Ferras M. On the use of GSV-SVM for Speaker Diarization and Tracking. Odyssey 2010-The Speaker and Language Recognition Workshop. 2010. 146–150.
7 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.计算机应用研究,2012,29(2):590–593.
8 Sahidullah M, Goutam S. Comparison of Speech Activity Detection Techniques for Speaker Recognition. Proc. of CoRR. 2012. 4–11.
9 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via EM algorithm. Journal of the Royal Statistical Society. 1977, 6(39): 1–38.
10 Moattar MH, Homayounpour MM. A review on speaker diarization systems and approaches. Speech Communication, 2012, 54(10): 1065–1103.
11 Auckenthaler R, Carey M, Lloy-Thomas H. Score normalization for text-independent speaker verification. Digital Signal Processing, 2000, 10(1-3): 42–54.