DOI:10.16208/j.issn1000-7024.2013.12.011

基于录制环境检测的数字音频取证研究

何少岩，陈蕉容，陈舜儿

（暨南大学 信息科学技术学院 电子工程系，广东 广州 510632）

摘 要：为解决数字音频伪造和篡改的检测问题，针对数字音频取证中录制环境检测，提出一种基于梅尔倒谱系数（MFCC）结合小波包分析的特征提取算法。算法用以提取音频的频域统计特性，结合时域特征构造特征集合，运用基于期望最大化（EM）的机器训练方法对音频录制地点进行分类和判断，实现数字音频录制环境的取证。实验结果表明，该算法能够较好的区分不同环境下的音频特性，纯净分类（无其他环境下的音频混入聚类组）最高可达 98%。

关键词：数字音频取证；环境监测；梅尔倒谱系数；期望最大化；小波包

中图法分类号：TN912.3 文献标识号：A 文章编号：1000-7024（2013）12-4142-04

Digital audio forensics based on recording environment detection

HE Shao-yan, CHEN Jiao-rong, CHEN Shun-er

(Department of Electronic Engineering, College of Information Science and Technology, Jinan University, Guangzhou 510632, China)

Abstract: To solve the detection problem of the forged or tampered digital audio, a feature extraction algorithm based on MEL cepstrum coefficients (MFCC) combined with wavelet packet analysis is presented according to digital audio forensic recording environment detection. Frequency domain statistical features are extracted by the algorithm mentioned above combined with the time domain features to structure a feature set. And then a machine training method based on expectation maximization (EM) algorithm is applied to detect the recording environment of the audio. A series of experimental analyses and tests show that the algorithm can distinguish the audio characteristics in different environments, and the best result is 98% with no audio recorded in other environments.

Key words: digital audio forensics; environment detection; MFCC; EM; wavelet packet

0 引 言

为了解决虚假音频广泛传播和使用在法律取证、商业版权、社会安全等方面引起的诸多问题，数字音频取证的篡改分析技术应运而生。音频取证作为一个新兴的、刚刚开辟的研究领域，在国内外的研究均处于起步阶段，深入研究的空间很大。而音频录制环境的检测方面，由于其自然环境多样性和复杂性等因素导致鲜有人涉及研究[1]。音频环境的检测可以一定程度上判断出该音频的原始性和真实性，能够为司法取证、犯罪侦查等提供重要依据，因此成为了数字音频取证技术的重要研究方向[2,3]。国外学者 Christian Kraetzer 等采取传统 MFCC 分析方法进行特征提取，应用贝叶斯分类器进行分类对音频录制环境和设备的检测进行了首次实验[4]，开辟了数字取证研究中基于音频环境和设备检测的研究领域。该实验结果显示对音频录制设备检测的准确率最高为 75.99%，对音频录制环境的准确率最高为 41.51%。随后 Robert Buchholz 等[5]人又专门针对音频录制设备进行检测，该实验利用傅里叶系数进行麦克风的分类，分类效果明显提高，准确率达 93.5%。然而，此实验思路下对音频录制环境的检测准确率一直未有提高。国内学者主要着重于音频分类领域的研究，使用傅里叶、小波等数学工具将音频文件分为语音、音乐、环境音等类型，在音频录制环境辨别与检测领域的研究并未涉及。本文借鉴前人研究思路，利用已有的音频分析工具（梅尔倒谱系数分析和小波包分析等）提取音频的频域统计特性，该频域统计特性和音频的 6 种时域特征构造特征集

收稿日期：2013-03-22；修订日期：2013-05-26

基金项目：国家自然科学基金项目（61070165）

作者简介：何少岩（1988-），女，河北石家庄人，硕士研究生，CCF 会员，研究方向为通信与信息系统；陈蕉容（1988-），男，广东汕头人，硕士研究生，研究方向为通信与信息系统；陈舜儿（1964-），女，广东佛山人，博士，副教授，研究方向为通信与信息系统。

E-mail: tchser@jnu.edu.cn

合，借助基于期望最大化的机器训练方法对音频录制地点进行分类和判断，从而实现数字音频录制环境的取证。实验结果表明，本文提出的特征提取方式和分类方法合理有效，能够对大部分音频录制环境进行正确的判断和分类，性能较好。

1　提取特征和分类方法

1.1　基于小波包变换的 MFCC 特征提取

音频信号通常采用 MFCC 进行分析处理，其本质是适应语音特性的滤波器组，是基于同态处理的去卷积倒谱改进算法。传统的 MFCC 处理方法是将信号进行傅里叶变换或短时傅里叶变换后，再经一系列处理，得到信号在不同谱带的功率变化速度。时域信号 S 的 MFCC 算法流程如图 1 所示[6]。

为了更加充分地计算音频动态特性，本文中的算法还引入信号的一阶差分梅尔倒谱系数（FMFCC）。该系数更好地消除了音频每帧之间的相关性，能够提高音频特征的辨识率[6]。FMFCC 计算如下

$$FMFCC = \frac{1}{\sqrt{\sum_{i=-j}^{j} i^2}} \sum_{i=-j}^{j} i \times sf_{mel}(n+i)$$

其中，sf_{mel}（$n+i$）表示第 n+i 帧的倒谱系数。通常 j=2，用以求第 n 帧的前两帧和后两帧倒谱系数的线性组合，即一阶差分倒谱系数。同理，继续迭代可求得二阶 FMFCC。MFCC 传统算法中的傅里叶变换将信号进行等间隔的频带划分。一旦分析窗口大小确定，FFT 分析就不能随着信号的变化而随时调整时频分辨率。而多分辨率分析由于尺度变化的局限性，也会导致其在高频段频率分辨率较差，在低频段时间分辨率较差。

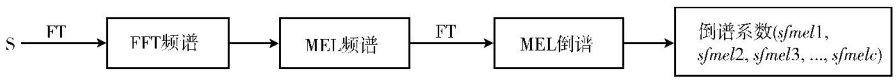

图 1　MFCC 算法流程

小波包变换思想较好地弥补了上述缺陷。小波包变换延续多分辨率分析方法，并且将时频空间进行非均匀划分，在频率较低的地方采用较长的时间窗。这使得成分复杂的音频信号能够被不同分辨率的小波系数表示。本文将小波包变换和 MFCC 以及 FMFCC 相结合，再增加音频时域的 6 个主要特征（能量熵、短时能量，频谱滚降，频谱重心，频谱通量，零值点），用以提取音频特征，从而对数据进行分类。特征提取算法步骤如图 2 所示。

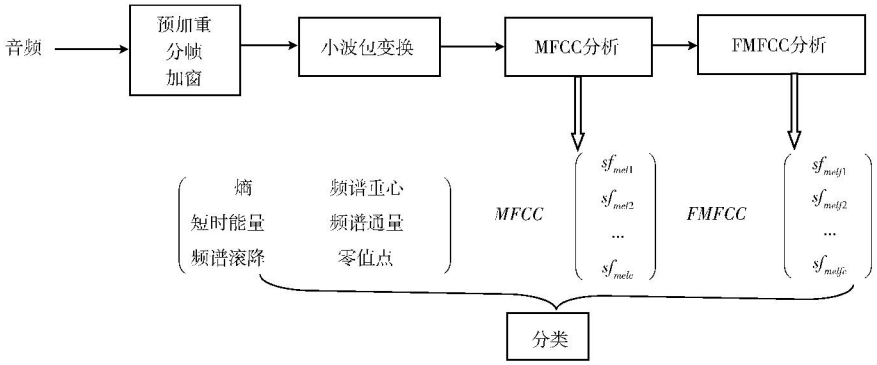

图 2　特征提取算法结构

1.2　基于期望最大化的机器训练聚类方法

期望最大化算法（EM）作为 k 均值算法的扩展，是用于估计未知参数的迭代优化过程。EM 根据隶属概率的权重把数据归到最为相似的类别中[7]。首先，EM 对整体数据集进行初始估计；再反复根据参数向量产生的混合密度对每个数据重新估计；被估计的数据最后用来更新参数估计。EM 过程中每个数据产生一个概率值，概率值反映了该数据属于某定类别集合的可能性大小。

EM 算法流程具体描述如下：

期望步骤：每个迭代过程中，EM 根据当前估计值为数据寻找一个最佳下界，用期望表示；再用如下概率将数据 x_i 归类到类别 C_k 中[8]

$$P(x_i \in C_k) = p(C_k \mid x_i) = \frac{p(C_k)p(x_i \mid C_k)}{p(x_i)}$$

其中，$p(x_i \in C_k) = N(m_k, E_k(x_i))$ 服从均值为 m_k、期望为 E_k 的高斯分布。该步骤计算了每类别中对象 x_i 的隶属概率。

最大化步骤：为了使数据集相似性更大化，EM 利用期望步骤得到的概率需要重新估计分布，给出未知变量的期望估计[7]

$$m_k = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i P(x_i \in C_k)}{\sum_j P(x_i \in C_j)}$$

其中，m_k 表示最终期望估计值。经实验验证，EM 算法容

易实现，对于某些特性的优化函数，收敛性较快。

2 性能测试方案

2.1 实验设备

实验采用的硬件设备为：得胜 PCM5550 麦克风、客所思录音外置声卡、hp 笔记本电脑。为了采集更加细微的环境噪声数据，麦克风和声卡均需要特殊的处理和配置，以增加敏感度，适应实验需求。实验采用的软件设备为：Audacity 1.3.5、Matlab 2010b 和 WEKA3.7.0，用以实现音频录制、特征提取分析和根据提取特征的分类。音频录制参数为单声道、工程采样率 44.1kHz，32-bit float。

2.2 数据采集

为了验证上述特征提取和分类算法的有效性，音频数据在六个不同的环境下进行采集[4,9]。录制地点如下：(i) 实验室，(ii) 图书馆，(iii) 自习室，(iv) 食堂，(v) 楼道，(vi) 湖边。为了反映某个环境整体的噪声特性，在一个环境下音频的采集工作将分为 10 个时间点均匀录制，时间范围是早八点至晚六点，每个时间点连续录制 5 段音频，每段音频 30s。

2.3 特征提取和分析

采用 Matlab 2010b 提取录制音频的 30 个特征数据，包括 6 种时域特征（能量熵、短时能量，频谱滚降，频谱重心，频谱通量，零值点），12 个 MFCC 特征（sfmel1，sfmel2，…，sfmel12）和 12 个 FMFCC 特征（sfmelf1，sfmelf2，…，sfmelf12）。特征数据不需要预处理，采用 WEKA3.7.0 分类工具直接进行聚类分析。分类工具采用 EM 算法[10,11]，聚类模式选用训练模型。

表 1 使用 MFCC 结合小波包的特征提取算法和 EM 分类器的分类结果

	Cluster0 实验室	Cluster1 图书馆	Cluster2 自习室	Cluster3 食堂	Cluster4 楼道	Cluster5 湖边
R0 实验室	**58%**	20%	—	—	—	22%
R1 图书馆	—	**92%**	—	2%	—	6%
R2 自习室	—	30%	**68%**	—	—	2%
R3 食堂	—	—	—	**100%**	—	—
R4 楼道	—	—	—	2%	**98%**	—
R5 湖边	—	10%	—	—	—	**90%**

3 测试结果与分析

由表 1 对角线方向数据可知，本文算法分类的正确率最高可达 100%（食堂），最差的正确率为 58%（实验室）。该系统实验的正确检测期望为 84%。

观察表 1 纵向数据可知，在 6 个类别中 Cluster1（图书馆）和 Cluster5（湖边）分类情况最为复杂。Cluster1 中，图书馆的分类率为 92%，但同时又有一定数目的其他环境下的录音也被分到了该类中，包括 20% 的实验室录音、30% 自习室录音和 10% 的湖边录音。Cluster5 也有相似的结果。这两组虽然自身的分类正确率均在 90% 以上，但混入了相当数量的其他类别的录音，说明图书馆和湖边这两个环境下的音频特征存在与其他环境下音频特征的相似之处，或者其他环境下某些时段的音频特征与图书馆和自习室的音频特征相似。而 Cluster0、Cluster2 和 Cluster4 没有混入其他环境下的音频。虽然 Cluster3 的分类正确率为 100%，但该组仍混入了其他环境下的音频。相比而言，Cluster4（楼道）分类正确达 98%，说明楼道的音频特征较为明显，综合辨识率较好。

表 2 仅使用 MFCC 特征提取算法和 EM 分类器的分类结果

	Cluster0 实验室	Cluster1 图书馆	Cluster2 自习室	Cluster3 食堂	Cluster4 楼道	Cluster5 湖边
R0 实验室	**84%**	—	—	16%	—	—
R1 图书馆	44%	**50%**	—	6%	—	—
R2 自习室	—	—	**88%**	12%	—	—
R3 食堂	—	—	—	**72%**	26%	2%
R4 楼道	—	—	—	8%	**92%**	—
R5 湖边	4%	50%	—	—	—	**46%**

横向观察表 1 数据，可知每行的百分数相加均为 100%，但不同行数据的离散程度相差较大。R0、R1 和 R2 数据都分布了 3 列，表明实验室、图书馆和自习室的音频特征明显度较低，或者该环境下不同时段的音频特征变化较大，易被误认为其他环境下的音频。显而易见，食堂的音频只集中在 1 列，没有被误判到其他环境。

与上述对比，表 2 列出了使用 MFCC 傅里叶变换进行特征提取的分类结果（其他条件相同）。观察表 2 对角线方向发现该算法未采用小波包变换，辨识准确率较低。但该算法在实验室和自习室两种环境下的分类效果优于小波包提取算法（如图 3 所示），仍然具有研究意义。

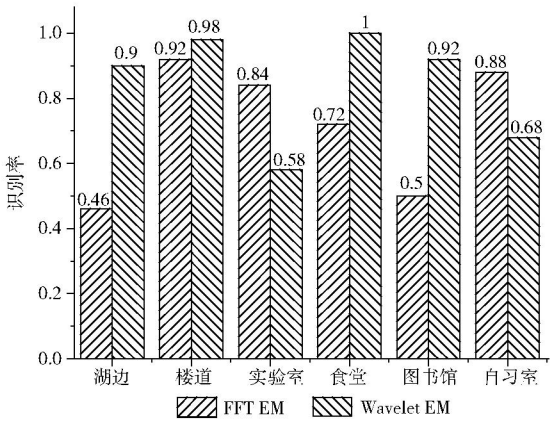

图 3 使用傅里叶和小波提取算法的分类结果对比

此外，本文还采用了基于质心的 k 均值分类方法进行结果分类，结果如表 3 所示。可以看出，无论采用傅里叶还是小波包进行特征提取，k 均值的分类准确率低于期望最大化算法的准确率。然而，无论采用 k 均值还是期望最大化分类算法，使用小波包进行特征后分类的结果总是优于傅里叶算法的。表 3 展示了各个算法下结果的期望、方差以及最大值，用来评价对比各算法的优劣。

表 3　EM 分类算法和 k 均值算法分类结果对比

	FFT EM	Wavelet EM	FFT K-mean	Wavelet K-mean
湖边	46%	90%	70%	100%
楼道	92%	98%	100%	96%
实验室	84%	56%	74%	68%
食堂	72%	100%	12%	86%
图书馆	50%	92%	52%	66%
自习室	88%	68%	98%	70%
期望	72%	84%	67%	81%
方差	0.04	0.03	0.11	0.02
最大值	0.92	1.00	1.00	1.00

4　结束语

本文针对数字音频盲取证技术中的环境检测进行了分析和测试，采用小波包和梅尔倒谱系数分析等数学工具提取特征，结合期望最大化算法进行机器训练聚类。实验结论如下：其一，对音频录制环境的分类准确率得到了大幅提高；其二，基于期望最大化的分类器比贝叶斯分类器更适合背景环境检测分类；其三，在 k 平均值分类器下小波包特征提取算法仍然占优。

本文提出的算法有较大的优越性，但同时有需要提升改进的地方：首先，音频大多仅具有有限时间长度的纯背景噪声，故在极短噪声采样下的音频环境监测成为了实验的一大挑战；其次，算法核心为小波包变换，故小波函数的选取对分类检测结果有着不可估量的作用，继续寻找合适的小波函数也是需要进行的又一工作。因此本文的后续研究将继续深入和提高，期待形成行之有效的检测系统，从而达到实用的效果。

参考文献：

[1] Ghulam Muhammad, Khaled Alghathbar. Environment recognition from audio using mprg-7 features [C] // IEEE Embedded and Multimedia Computing International Conference, 2009: 1-6.

[2] Malik H, Farid H. Audio forensics from acoustic reverberation [C] // IEEE International Conference on Acoustics Speech and Signal Processing, 2010: 1710-1713.

[3] Ikram S, Malik H. Digital audio forensics using background noise [C] // IEEE International Conference on Multimedia and Expo, 2010: 106-110.

[4] Kraetzer C, Oermann A, Dittmann J. A digital audio forensics: A first practical evaluation on microphone and environment classification [C] // the 9th workshop on Multimedia & Security, 2007: 63-74.

[5] Bucholz R, Kraetzerr C, Dittmann J. Microphone classification using fourier coefficients [C] // 11th International Workshop, Darmstadt, 2009: 236-246.

[6] Kraetzer C, Dittmann J. Mel-cepstrum based steganalysis for voIP-steganography [C] //Security, Steganography and Watermarking of Multimedia Contents IX, 2007: 6505.

[7] Ngai Ewt, Hu Yong, Wong Yh. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature [J]. Decision Support Systems, 2011, 50 (3): 559-569.

[8] Uri Nodelman, Christian R, Daphne Koller. Expectation maximization and complex duration distributions for continuous time bayesian networks [C] //the Twenty-First Conference on Uncertainty in Artificial Intelligence, 2012: 421-430.

[9] Hong Zhao, Hafiz Malik. Audio forensics using acoustic environment [C] //Statistical Signal Processing Workshop, 2012: 373-376.

[10] ZHANG Xueyuan, HE Qianhua, LI Yanxiong, et al. An inverted index based audio retrieval method [J]. Journal of Electronics Information Technology, 2012, 34 (11): 2561-2567 (in Chinese). [张雪源，贺前华，李艳雄，等. 一种基于倒排索引的音频检索方法 [J]. 电子与信息学报，2012，34 (11)：2561-2567.]

[11] Godiy Daniela. One-class support vector machines for personalized tag-based resource classification in social bookmarking systems [J]. Concurrency and Computation-Practice & Experience, 2012, 24 (17): 2193-2206.