ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 5,Sep. 2018,pp. 779－792
DOI:10.16337/j.1004-9037.2018.05.002

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

音频取证中录音设备识别研究进展

包永强　梁瑞宇　王青云

(南京工程学院通信工程学院,南京,211167)

摘　要:从音频信号中提取录音设备特征是司法比较研究和音频取证的前沿课题。由于录音设备识别技术受到环境、语义、说话人等因素干扰,需要攻克的难题较多,国内外的研究还处于起步阶段。为此回顾了录音设备研究的发展情况、基本理论和组成结构,特别对组成结构中非话音段检测、特征参数、识别模型和数据库建设的研究现状进行了介绍和分析。最后,进一步分析了录音设备识别存在的不足,并展望未来的研究发展方向,指出加快构建现有各品牌各型号的录音设备、各场合、各类人群的数据库建设与深度学习在录音设备中的应用是下一阶段研究的重点。

关键词:音频取证;录音设备识别;特征提取;数据库建设

中图分类号:TP391　　**文献标志码**:A

Advance in Recorder Recognition for Audio Forensics

Bao Yongqiang, Liang Ruiyu, Wang Qingyun

(School of Communication Engineering,Nanjing Institute of Technology,Nanjing, 211167, China)

Abstract: Extracting the characteristics of recording devices from audio signals is a frontier in judicial comparative research and audio forensics. As a research hotspot of audio forensics, recorder recognition technology is disturbed by environment, semantics, speaker and other factors. This paper introduces the development, basic theory and structure of recording equipment research. Especially, the research status of non-speech segment detection, feature parameters, recognition model and database construction is introduced and analyzed. Finally, the shortcomings and prospects of recorder identification are discussed. It is considered that the next stage should focus on how to speed up the database construction and the application of deep learning in recorder recognition.

Key words: audio forensics; recorder recognition; feature extraction; database construction

引　　言

根据2018年2月国家工信部统计数据显示[1],全国移动电话用户为14.2亿户,移动电话普及率达1.024部/人,手机成为人们生活中必备物品,从京东、天猫等网站在售手机资料显示,目前国内市场手机品牌在200个左右,各大品牌手机类型累计在1 000款左右,而且市场上的手机基本都具有录音、录

基金项目:国家重点研发计划(2017YFB0903504)资助项目;国家自然科学基金(61871213)资助项目;江苏省教育科学"十三五"规划(B-a/2016/01/44)重点资助项目。
收稿日期:2018-05-18;**修订日期**:2018-07-20

像等功能。由于手机、录音笔等具有录音功能的电子产品快速发展和普及,2017 年《中华人民共和国民事诉讼法》[2]第六章明确把视听材料作为法庭认可的 8 大证据之一。音频证据由于具有互动性、录制便捷,呈现直观等原因在法庭证据中所占比例越来越大。但随着音频信号处理技术的飞速发展,音频编辑工具如 Cool Edit、Adobe Audition、Samplitude、Cakewalk、Steinberg Nuendo 等音频编辑软件层出不穷而且功能越来越强大,证据提供方可以方便地通过这些软件进行音频的裁减、修改和伪造,形成有利于自己的音频证据,从而对法庭证据的鉴定带来很大的困难。《最高人民法院关于民事诉讼证据的若干规定》明确规定合法性、真实性和关联性是音频资料成为证据的 3 大条件,存在疑点的视听材料不能作为法庭证据。1958 年美国麦柯文案中专家采用了音频证据认定的 7 个准则[3]。这 7 个准则主要还是围绕音频材料的合法性和真实性展开的。音频证据的合法性和真实性认定从音频信号处理角度来看,实际上就属于音频取证的范畴。

音频取证[4]主要鉴定音频信号是何时(录音时间识别)何地(录音地点识别)由何设备(录音设备识别)所录制,是否经过篡改(音频篡改检测),还包括录音内容、说话人识别和录音格式识别等。2007 年,Kraetzer 等[5]第 1 次提出了录音设备识别的模型,采用内核为 K-均值算法的怀卡托智能分析环境(Waikato environment for knowledge analysis, WEKA)机器学习软件和 Bayes 分类算法对 4 种麦克风在 10 个不同的房间的录音进行识别,取得了超过 75%的识别率。2009 年,Buchholz 等[6]提出了一种麦克风识别模型,该模型利用 9 个有/无声判决门限进行寂音段判决,然后采用寂音段的 2 048 个傅里叶变换参数作为特征参数,采用线性 Logistic 回归模型作为分类算法,对 7 个麦克风进行分类,取得了 93.5%的正确识别率。近十年来,手机音频在全部录音设备中所占比重越来越大,2014 年,Aggarwal 等[7]研究了诺基亚、三星、黑莓、索尼、Zen5 个手机厂商的 26 个型号的手机,建立了音频数据库,提出了手机音频识别的模型,该模型采用 24 维 Mel 频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)混合参数作为特征参数,以 SMO-SVM 为分类模型,5 种品牌识别率达 90%,对诺基亚厂商各型号的平均识别率达 72%。

在音频信号特征矢量中,蕴涵了话音内容特征、说话人个性特征、录音环境特征和录音设备特征等多类特征。这些特征在时域和频域相互重叠,一般的分类方法很难在时频域上进行有效的线性分离。从目前的研究现状来看,录音设备识别研究还处于起步阶段,尚未有对各型号手机或录音设备的传递特性的研究,在录音设备数据库构建、录音设备特征与其他特征分离和提取、识别模型设计等方面还未形成有效的解决手段,还有很大的研究空间。

1 录音设备识别理论

录音设备识别是从音频信号中提取表征录音设备的特征参量,采用模式识别的方法进行训练和识别,从而做出该音频信号是哪种录音设备所录制的判断。Aggarwal 等[7]认为,如果把录音设备看作一个对音频信号的滤波器,则寻找表征录音设备的特征参量实际上就是从音频信号中提取录音设备的传递函数。录音设备主要包括手机、录音笔、麦克风等。由于手机已成为人们生活中随身必备的物品,手机识别成为录音设备识别中的热点。

图 1 给出了录音设备识别的基本模型,与传统的语音识别、说话人识别模型一样,录音设备模型主要包括非话音段检测、特征提取、模式识别与数据库构建等部分。其中非话音段检测是检测该段音频是属于话音段还是非话音段。值得说明的是录音设备识别模型中的特征提取一般是从音频信号中的噪声段或噪声谱中提取录音设备特征,其中蕴含了录音设备特征与其他特征的分离。

2 非话音段检测

音频信号从时域上看,可分为话音段和非话音段两大部分,录音设备特征贯穿于整个音频信号的话音段和非话音段,一般而言,话音信号功率在整个音频信号中占比较大,如果不预先进行处理,易影响到

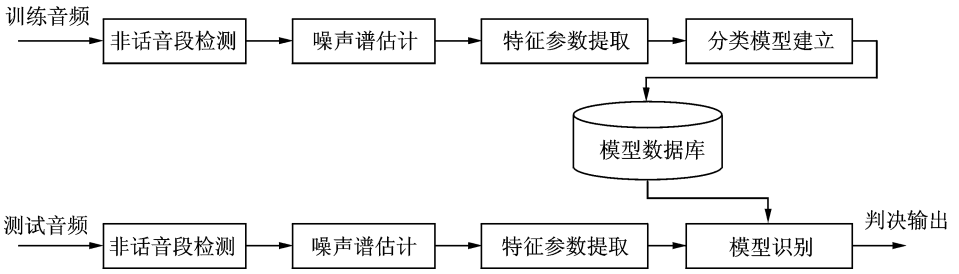

图 1 录音设备识别常用模型

Fig. 1 The models for recorder recognition

录音设备识别性能。另一方面,话音信号频谱与录音设备特征在频域上重叠,很难进行区分,常用的方法是在时域将话音段进行抛弃,只考虑非话音段。非话音段检测在语音信号处理领域又叫端点检测(Voice activity detection, VAD)。对于话音段中的噪声等成分,通常的处理方法也是以非话音段为基准进行自适应估计。因此,在录音设备识别中,一般只对非话音帧进行处理,提取特征参数。

非话音段检测一般包括分帧、预滤波、有无声特征提取、端点判决和后处理 5 个阶段,如图 2 所示。由于语音信号为短时稳态过程,一般认为是 20～30 ms,分帧是将音频信号分割成多个音频段,根据音频信号采样率不同,一般以 256,512,1 024 点为一帧,相邻帧之间有重叠,可以有 25%,50%,75%这 3 种重叠率。预滤波主要是采用高通滤波器滤除低频噪声。

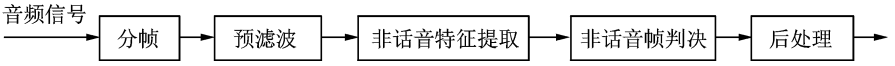

图 2 非话音段检测基本流程

Fig. 2 The basic process of endpoint detection

非话音特征提取主要是寻找一种能够最大区分话音与非话音的特征或判决准则,目前主要可分为以下几个类别:

(1)时域特征:1975 年,贝尔实验室的 Rabiner 等[8]提出了基于短时能量和过零率的双门限检测方法,能够有效应用于低噪声环境下的端点检测,20 世纪 90 年代以来,Taboada,Marzinzik 等[9-10]分别在此基础上进行了改进;2003 年,Yang 等[11]采用分形维来区分语音与噪声,性能优于短时能量和过零率方法。时域方法具有计算量小,便于实时处理等优点,但对非平稳噪声性能较差,正确识别率较低。

(2)变换域特征:1980 年,Kobayashi 等[12]从音高升降曲线角度出发,提出了基于频谱特征的词边界检测方法;1999 年,Vahatalo 等[13]采用 9 个 IIR 滤波器组,提出了一种基于频谱和周期检测的方法,用于 GSM 语音编码;2002 年,Lin 等[14]提出了一种基于最小 Mel 频带参数和增强时频参数用于估计动态噪声,正确识别率可达 75%,成功应用于 VAD 检测;2006 年,Evangelopoulos 等[15]基于滤波器组,提出了 Teager-Kaiser 能量算子应用于 VAD 检测。变换域算法具有计算量较小,能够部分识别非平稳噪声,正确识别率一般。

(3)基于距离和测度方法:1993 年,Haigh 等[16]提出了一种基于倒谱距离的 VAD 检测方法,成功应用于低信噪比环境;2015 年,赵新燕等[17]针对强噪声环境,通过引入倒谱距离乘数和门限增量系数,提出了一种基于自适应倒谱距离的端点检测方法,在－5 dB 噪声环境下取得了大于 80%的正确识别率。基于距离和测度的方法在较低噪声环境下具有较强的鲁棒性、计算量较小、正确识别率较高、便于实时处理、对清音效果一般等特点。

(4)信息论方法:1995 年,Mcclellan 等[18]将谱熵有效应用于线性激励预测编码的端点检测中,取得

了较好效果；2008 年，Lorber 等基于谱熵，提出了 Mel 域上两级维纳滤波器[19]用于 VAD 检测，并采用了 NOISEX-92 数据库进行了验证；李艳等[20]将子带谱熵与短时平均幅度结合在一起，取得了较好的鲁棒性能。Liu 等[21]基于预测增益和预测值提出了一种自适应增量调制结构用于 VAD 检测。基于信息论的方法具有能有效应用于低噪声平稳环境、大部分方法计算量较小、正确识别率尚可、对清音效果较差等特点。

(5)基于神经网络的方法：2000 年，Hussain 等[22]将多层感知器和自适应线性神经网络应用于 VAD 检测，性能优于传统 VAD 检测方法；2002 年，Ghisellicrippa 等[23]将两层反馈神经网络结合拟牛顿误差最小化方法应用于有无声检测，误识率降至 3%～5%，取得了比时域和变换域方法更好的性能。基于神经网络的方法具有性能好、能够有效应用于低噪声平稳环境、参数选择合适能有效针对非平稳环境、对清音效果较好、计算量较大、实时性较差等特点。

(6)统计模型和分类算法：2000 年，Couvreur 等[24]提出了一种非参数估计隐马尔科夫模型(Hidden Markov model, HMM)模型，成功应用于 VAD 检测；2003 年，Gazor 等[25]提出了一种两态 HMM 模型用于 VAD 检测，该模型假定语音和噪声分别为 Laplacian 与 Gaussian 模型，取得了较好的性能；2018 年，Selvakumari 等[26]将 SVM 分类方法引入 VAD 检测，该方法结合了 Bayes 算法。基于统计模型和分类算法的方法具有性能好、能够有效应用于低噪声平稳环境[27]、模型选择合适能有效针对非平稳环境、对清音效果较好、计算量较大、实时性较差、不同噪声环境需要不同模型等[28]特点。

除了以上各类方法外，还有基于混沌的方法、基于谱减的方法等。与说话人识别一样，录音设备识别对有无声判断在计算量、实时性方面要求不一定很高，但对非平稳噪声环境、清音判别、正确识别率等方面具有较高要求。可以根据需要选择统计模型和分类算法、基于神经网络的方法和变换域方法等，或者多种方法相结合。

3 录音设备特征参数

音频设备特征提取是从给定音频信号中提取录音设备特征信息。从现有文献来看，录音设备特征参数一般针对非话音段提取傅里叶系数直方图[29]、功率谱[30]、MFCC 倒谱参数(Mel-frequency cepstrum coefficient, MFCC)、感知线性预测参数(Perceptual linear prediction, PLP)、随机谱特性(Random spectral features, RSFs)、Bark 倒谱参数(Bark-frequency cepstrum coefficient, BFCC)和线性预测编码参数(Linear predictive coding, LPC)等频域参数。

3.1 基于傅里叶系数的方法

在语音信号领域中，傅里叶系数一般作为语音信号特征参数，很少作为录音设备特征参数，实际上傅里叶系数不仅包含语音信号特征、说话人特征、噪声、环境特征，还包括录音设备等特征。2009 年，Buchholz 等[29]提取无声段的傅里叶系数直方图作为特征参数，对 7 个麦克风进行分类，正确识别率达 93.5%。该论文是目前为止录音设备识别比较有代表性的一篇学术论文。

Buchholz 等针对给定的 8 段音频信号分别由电动式、电容式和压电式的 3 类 7 个型号的麦克风在 12 个不同的房间(楼梯间、小办公室、大办公室、演讲厅等)进行录制，形成 672 个音频信号(每个音频信号采样频率 44.1 kHz, 16 bit 量化，每段音频时长约 30 s，未压缩脉冲编码调制(Pulse code modulation, PCM)格式编码，图 3 为特征提取和分类模型。

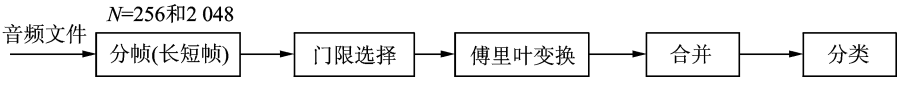

图 3 傅里叶系数提取方法

Fig. 3 Fourier coefficient extraction method

对于输入的音频信号,按照帧(窗)长分别为256和2 048进行分帧,帧与帧之间不重叠,对于帧长$n=256$,其门限对应为$t\in\{0.01, 0.025, 0.05, 0.1, 0.2, 0.225, 0.25, 0.5, 1\}$;对于帧长$n=2\,048$,其门限对应为$t\in\{0.01, 0.025, 0.05, 0.1, 0.25, 0.35, 0.4, 0.5, 1\}$。对于通过门限的信号进行傅里叶变换并求其直方图,作为WEKA学习模型的特征参数。WEKA是一种机器学习算法,该论文选择了其中的朴素Bayes、序列最小化优化支持向量机(Sequential minimal optimization-support vector machine,SMO-SVM)、Logistic回归、J48决策树、K最近邻($K=1,2$)分类算法等分类算法。从正确识别率看,可以获得以下结论:

(1)各分类算法最佳门限t不尽相同,多集中在0.25~0.4之间。

(2)当采用长帧时($n=2\,048$)按照分类算法的正确识别率从高到低排序:Logistic回归 SMO-SVM K等于1最近邻算法>K等于2最近邻算法 J48决策树朴素Bayes。Logistic回归识别率最高,可达93.5%;SMO-SVM,可达90.6%;朴素Bayes性能最差,达46.9%。

(3)当采用短帧时($n=256$)按照分类算法的正确识别率从高到低排序:Logistic回归> K等于1最近邻算法> K等于2最近邻算法 SMO-SVM>J48决策树> 朴素Bayes。Logistic回归识别率最高,可达90.6%;$K=1$最近邻算法其次,可达87.1%;朴素Bayes性能最差,达40.5%。SMO-SVM受帧长的影响很大,短帧时与J48决策树性能相当,其他几种分类算法受帧长影响不大。

3.2 非话音段功率谱估计

功率谱估计可分经典谱估计和现代谱估计两大类,最简单的经典谱估计是周期法,可采用信号频谱与其共轭乘积获得,也可通过计算信号自相关函数的傅里叶变换获得。但周期法具有分辨率较低的缺点,人们在此基础上提出了加权周期法和Welch法,改善了谱曲线的平滑度和分辨率;现代谱估计立足于提升平滑度和分辨率,可分为参数模型(自回归(Auto regressive, AR)模型、滑动平均(Moving average,MA)模型、自回归滑动平均(Auto regressive moving average,ARMA)模型等)和非参数模型(最小方差法、多信号分类(Multiple signal classification, MUSIC)法等)。

对非话音段功率谱的估计非常重要,估计过低易导致录音设备特征被排除在外,过高则导致话音信号影响录用设备特征提取。实际中,取证音频录制一般在非平稳噪声环境下,信噪比一般较低。20世纪90年代以前,噪声功率谱提取方案主要基于VAD硬判决,且只针对平稳噪声环境[31-32],20世纪90年代以来,噪声功率谱提取开始采用软判决模式或无VAD模式[33-42],主要可分为基于噪声统计特性的方法[33-35]、基于噪声空间结构模型的方法[36-38]、基于传播矢量的非盲方法[39-40]、基于单/多通道的语音存在概率(Speech presence probability,SPP)的方法[41-43]等。

(1)基于噪声统计特性的方法

动态语音检测VAD技术在低信噪比环境下,特别是在非平稳噪声情况下表现不佳,严重影响语音类识别系统性能,Hirsch等[33]开始研究无VAD的噪声功率谱估计,引入平滑因子,通过门限设置,将上一时刻噪声谱和当前子带谱的加权和作为噪声谱估计进行降噪,有效提高了说话人识别的性能。Martin等[34]认为话音持续时间内噪声短时能量经常在噪声平均值左右,可以通过跟踪窗内最小功率来衔接估计噪声功率谱,将功率谱平滑和最小统计特性技术有机结合在一起,对于非话音段引入平滑因子,将上一时刻噪声谱和当前信号功率谱的加权和作为噪声谱估计;对应话音段,噪声功率谱由窗内最小功率谱决定。基于噪声统计特性的方法采用软判决的方法,通过对话音段和非话音段噪声进行跟踪平滑,能够有效适用于非平稳噪声情况,可以应用于录音设备特征提取。

(2)基于噪声空间结构的方法

一般认为,实际录音环境下的噪声中有3类相干噪声占主要成分:麦克风电路产生的相干噪声、已知方向上的噪声源产生的众多相干噪声和由非相干信号产生的弥散噪声,如图4所示。基于噪声空间结构的方法主要是假设噪声是由单一的噪声源通过不同路径反射获得的具有较强相关性的合成信号,

即噪声空间特性具有一定结构，可以采用空间协方差结构来估计噪声。Yousefian 等[36]研究了相干噪声产生和降噪机理，通过定义信号与噪声的功率谱相干函数来估算噪声功率谱，提出了一种双麦克风降噪算法，性能优于普通波束形成技术。

由于基于噪声空间结构的方法主要着重研究相干噪声，不能直接用于录音设备特征提取中，但可以应用于录音设备特征与由噪声源产生的相干噪声分离中，从而提升录音设备识别率。

(3)基于语音存在概率的方法

针对非平稳噪声和低噪声环境，Martin 等[34]提出的最小统计算法来估计噪声功率谱，但该方法对异常值比较敏感，而且估计方差要大于传统估计器，Cohen 等[41]在此基础上，通过引入由语音存在概率控制的平滑因子，采用信号谱平均作为噪声谱估计，提出了一种最小控制递归平均方(Minima controlled recursive averaging，MCRA)，子带中的语音存在率是由时间窗内局部能量与最小能量之比确定，并采用平滑处理来解决语音与非语音间的波动，该方法能够快速跟踪噪声的突然变化。Souden 等[42]将其应用到多通道模式，基于高斯统计模型、时间平均和双迭代技术，所提出的噪声功率谱矩阵能有效降低 F16 战斗机噪声和嘈杂噪声。基于语音存在概率的方法从时频域上精确估计语音成分，有效降低噪声影响，可以有效应用于录音设备特征估计。

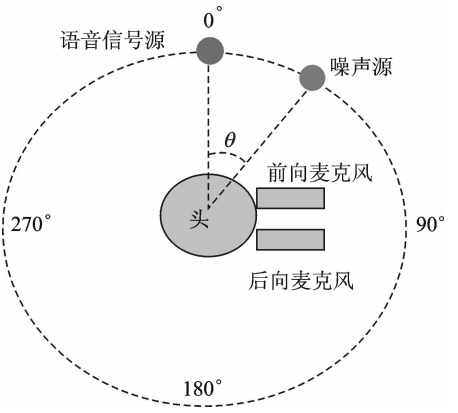

图 4 噪声空间结构模型

Fig. 4 Noise spatial structure model

3.3 MFCC 倒谱参数

作为语音信号处理领域主要的特征参数之一，模拟人耳感知特性的 MFCC 倒谱参数不仅抗噪声能力强，而且在语音识别、说话人识别、情感识别、端点检测等应用中皆具有较高的识别率[7,44-45]。2014 年，Aggarwal 等将其用于表征手机特性参数，对 5 种不同品牌的手机取得了 90%的识别率。

分析研究表明 Mel 刻度[44]在 1 000 Hz 以下与声音频率 f 呈线性关系，1 000 Hz 以上呈对数关系 MFCC 倒谱参数提取过程如图 5 所示。图 5 中 Mel 滤波器组一般采用个数为 12 或 24 的三角滤波器组，其频谱呈 50%的重叠。

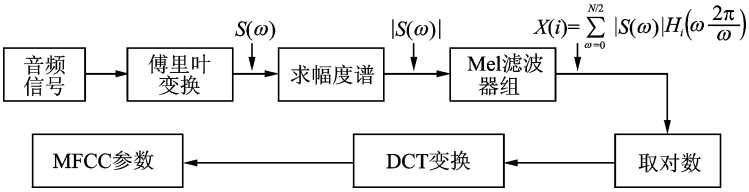

图 5 MFCC 倒谱参数提取过程

Fig. 5 MFCC cepstrum parameter extraction process

MFCC 倒谱及差分倒谱携带有丰富的个性特征，常作为语音类识别的特征参数，不仅携带语音特征、说话人特征，而且也能很好地表征录音设备特征、录音场合特征。

3.4 感知线性特征

1990 年，Hermansky 等提出了感知线性特征参数，该特征[46-49]基于人耳掩蔽效应、等响曲线等听觉特性，采用自回归全极点模型来逼近语音信号频谱，Vachhani 等[47]在此基础上又提出了 PLPCC，RASTA-PLPCC 等参数，在语音识别、说话人识别、语音分割等领域得到了广泛应用，Daly 等[49]认为 PLP 系列参数与 MFCC 性能相当。具体过程如图 6 所示。图 6 中的临界带分析是对音频信号通过 Bark 域滤

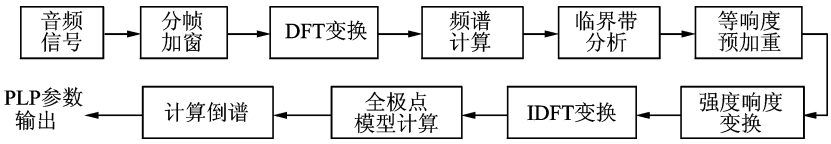

图 6 PLP 提取过程

Fig. 6 PLP extraction process

波器组(符合人耳掩蔽效应的临界带)的输出进行分析。

3.5 随机谱特性

2012 年,Panagakis 等[50]采用随机谱特性(RSFs)作为固定电话音频特征参数,共提取 325 维参数作为稀疏表示分类器的输入参数,通过 LLHDB[51]数据库测试,识别率达 95.55%,优于 23 维 MFCC 参数。RSF 计算过程如图 7:首先获得音频信号短时功率谱,并进行时间域上平均获得平均功率谱,然后为降低平均功率谱维度,通过随机投影算子获得音频信号的 RSFs 参数。

图 7 RSFs 提取过程

Fig. 7 RSFs extraction process

表 1 中共采用 3 种分类器,分别为支撑向量机(Support vector machine,SVM)、稀疏表示分类器(Sparse representation classifier,SRC)和神经网络(Neural network,NN)。从表 1 中可以看出,同等情况下,RSFs 识别率比 MFCC 要好大约 7%。

表 1 两种特征参数识别率[50]比较

Tab. 1 Comparison of recognition rates between two characteristic parameters

特征参数	维数	分类器	正确识别率/%
RSFs	325	SRC	95.55
RSFs	625	SVM	94.81
RSFs	475	NN	88.23
MFCC	23	SRC	89.79
MFCC	23	SVM	87.35
MFCC	23	NN	81.95
基于高斯超矢量的 MFCC	N/A	SVM	93.20

4 录音设备分类算法

录音设备识别中采用分类算法一般有 Bayes 分类算法[52]、决策树[53]、K-最近邻分类法[54]、SRC[55]、Logistic 回归模型[56-57]、SVM、高斯混合模型(Gaussian mixed model, GMM)[51]、NN 等,其中 SVM[58]最常用。

(1) SVM 分类算法

作为统计学习理论领域重要的学习方法,SVM 利用边界样本的分类信息,通过调整判别函数,较好地解决了小样本、非线性、高维情况下模式识别问题。在端点检测[59]、语音识别[60]、说话人识别[61]、语音情感识别[62-64]、录音设备识别[30]等领域应用广泛。研究表明,SVM 性能优劣和泛化能力受到惩罚因

子 C 和核函数 g 的影响，可以采用遗传算法(Genetic algorithm，GA)[65]（见图 8）、模拟退火算法(Simulated annealing，SA)[66]、粒子群优化算法(Particle swarm optimization，PSO)[67]等算法进行优化。

Hanilci 等[58]认为可以采用聚类方法[68]或采用广义线性判决序列核(Generalized linear discriminant sequence，GLDS)[69]将特征参量序列映射到核函数空间，来降低训练数据长度。

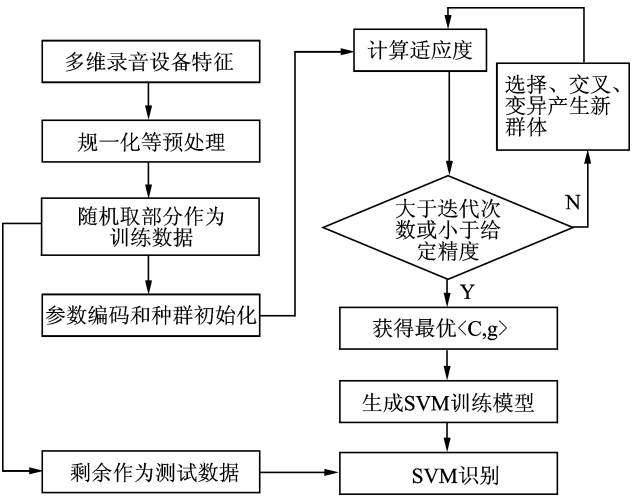

图 8 GA-SVM 识别模型

Fig. 8 GA-SVM recognition model

(2) GMM 模型

1995 年，Reynolds[70]将 GMM 引入与文本无关的说话人识别，GMM 基于 Bayes 理论，把音频训练和匹配等问题，转化成模型选择、参数训练、概率计算等子问题，取得了较好的性能，GMM 成为了说话人识别[71-72]的首选分类方法。

由于 GMM 模型可以对任意形式特征参量的统计分布进行描述，当混合度 M 足够大时，可以精确地逼近录音设备特征矢量的空间分布，但也存在着对训练数据能够覆盖整个录音设备特征集。与 HMM 相比，GMM 只有一个状态，不重点考虑特征的时序过程，只描述特征矢量的空间特性。从而减少了音频与时序的关系对录音设备识别的影响。多维空间矢量与高斯混合模型的概率输出是多个高斯概率密度函数的加权和，可以采用最大期望(Exception maximum，EM)算法来估计 GMM 参数。

(3) SRC

SVM 在小样本情况下具有较好的泛化能力，但录音设备识别需要较大 VC 维才能较好表征设备特征，从而削弱了 SVM 小样本优势，SRC 对大样本、小样本和高维数据都具有较好的识别性能[73-74]。Panagakis 等[50]将 SRC 引入录音设备识别中，取得了比 SVM 和 NN 都好的性能。语音信号具有较高的冗余度，SRC 通过假设语音信号为稀疏信号，可用过字典中的原子来进行线性描述，如图 9 所示。

(4) 基于 NN 的方法

NN 在录音设备识别中应用较晚，2014 年，Kotropoulos 等[75]将 RBF-NN 应用于手机识别中，通过 MOBIPHONE 数据库测试，识别率高达 97.6%，性能优于 SVM 和多层感知器，作为神经网络领域早期的代表技术之一，RBF 具有结构简单、学习收敛速度快、能够逼近任意非线性函数的优点，但 RBF 受限于基函数中心的选择，而且在小样本情况下性能不佳。

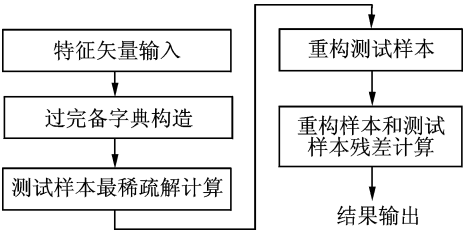

图 9 稀疏表示分类器工作流程

Fig. 9 Workflow of sparse representation classifier (SRC)

近年来，深度学习开始在语音信号处理领域广泛应用[76-79]。2012 年，Ossama 等[76]将卷积神经网络(Convolutional neural networks，CNN)引入语音识别的 NN-HMM 混合模型中，用频域中 CNN 准则来归一化声学特征方差，取得了较好的性能。Mitra 等[77]提出了一种时频域双层 CNN 用于语音识别，通过 Fisher 数据集测试，其参数远小于传统 CNN 性能优于且在抗噪声和背景干扰方面优于传统深度 NN。Hrúz 等[78]将 CNN 应用于说话人变化检测，通过 NIST 数据库测试，性能优于广义似然比距离。深度学习特别是 CNN 在语音

信号处理领域应用表现良好,但在录音设备识别中应用还未见报道。

5 录音设备数据库建设方面

录音设备数据库主要分两大类,固定电话音频数据库和手机音频数据库,这些数据库基本都是在TIMIT 数据库基础上采用音频设备进行翻录。从音频数量上看,HTIMIT、LLHDB 相对得到大部分学者认可,但由于手机、电话等电子产品更新换代较快,传统录音设备数据库面临淘汰危机,手机音频数据库不断需要更新。

(1) HTIMIT 数据库

1997 年,Reynolds[80]利用扬声器播放来自 TIMIT 数据库中的 384 个说话人(192 位男性和 192 位女性)音频,并分别通过 9 部固定电话和 1 个麦克风录制,语音信号采样率为 8 kHz,除 TIMIT 数据库语音外,每个手持设备(固定电话)来播放了 1 Hz 的扫描信号和高斯白噪声作为测试信号,从而形成了 HTIMIT 数据库,如图 10 所示。

固定电话面对的声音信号并非直接来自说话人现场声音,信号最大频率限制在 4 kHz 以下,由于扬声器、模数转换电路、电话局模拟器等影响,HTIMIT 数据库并不理想,但 HTIMIT 数据库包含了大量说话人的录音设备数据,是其他数据库所不能比拟的,后来的很多研究都是基于该数据库展开的[51]。

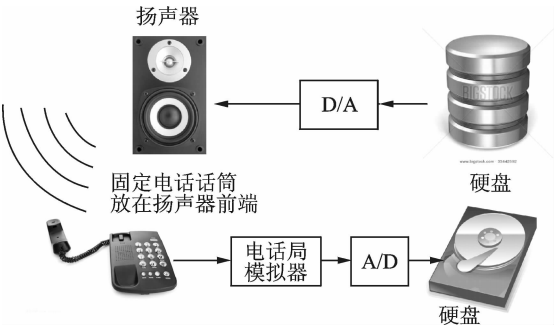

图 10 HTIMIT 数据库构建模型

Fig. 10 HTIMIT database building model

(2) LLHDB 数据库

Reynolds[80]利用同样的 9 部固定电话和 1 个高质量麦克风现场录制了 53 个说话人(24 位男性和 29 位女性)的话音,构建了 LLHDB 数据库(见图 11)。与 HTIIT 数据库相当,LLHDB 数据库采样率也为 8 kHz。

(3) Buchholz 数据库

2009 年,Buchholz 等[29]采用电动式、电容式和压电式 3 类共 7 个型号的麦克风针对 8 个语音源文件在 12 个不同类型的房间录制了 672 段音频文件,音频信号采样率为 44.1 kHz,量化比特为 16 bit,与 HTMIT 和 LLHDB 数据库相比,Buchholz 数据库的音频信号采样率大大提高了,使之蕴含了丰富的录音设备特征。

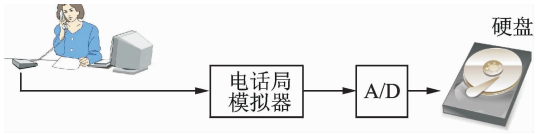

图 11 LLHDB 数据库构建模型

Fig. 11 LLHDB database building model

(4) LIVE RECORDS 数据库

2012 年,Hanilci 等[58]从 TIMIT 数据库中选取了 24 个说话人音频共 240 段,采用 14 部手机进行录制共产生 3 360 段音频,音频格式为自适应多速率(Adaptive multi-rate,AMR)格式,采样频率为 8 kHz,速率为 12.2 kbit/s,Zou 等[81]仿照 LIVE RECORDS 构建模式采用三星、索尼和 NOKIA 三个品牌 7 种型号手机,也建立了自己的手机音频数据库。

(5) Pandey & Aggarwal 音频数据库

2014 年,Aggarwal 等[7]、Pandey 等[30]采用三星、索尼和 NOKIA 等 5 个品牌 26 种型号手机进行现场录制音频,音频格式分 WAV 和 AMR 两种格式,其中 AMR 格式采用 FFMPEG 软件进行转换成 WAV 格式进行分析。

(6) MOBIPHONE 数据库

2014 年,Kotropoulos 等[75]建立了利用 7 个不同品牌 21 种型号手机建立了 MOBIPHONE 数据库,该数据库从 TIMIT 数据库中选取了 24 位男女(12 位男、12 位女),每个手机录制 10 段音频,音频信号采样频率为 16 kHz。

6 总结与展望

国内外对录音设备识别的研究还处于起步阶段,可以从以下几个方面开展深入研究:

(1)录音设备特征参数

从国内外现有的文献来看,由于版权保护等原因,虽然各设备厂商在录音设备设计方面,特别是在麦克风头、调理电路、压缩算法、语音增强算法等方面有诸多不同,但尚未有专门针对录音算法、音频电路等的特征参数出现,下一阶段的研究可以分为两大类:

一是研究分析现有音频类特征参数,从中提取录音设备个性特征。录音设备特征参数基本还是采用语音类特征参数,其中符合人耳听觉特性的类倒谱参数如 MFCC、PLP、RSFs 等参数具有较高的识别率,分析这些参数发现,这些类倒谱参数都具有将激励源与传输特性进行了线性分离,而录音设备特征蕴含在传输特性中,下一阶段的研究应该是如何实现录音设备特征与说话人特征、录音场合特征及噪声进行分离,基于噪声空间结构的方法提供了一个很好的思路,可以从信号源的角度进行处理解决。

二是从录音电路和算法特点寻找特征参数。除了在现有音频特征基础上进行特征分离外,针对录音电路特点,特别是对 50 Hz 以下和 15 kHz 以上频段的传输特性、信号中蕴含的电网谐波特性和针对压缩和语音增强算法对噪声处理的措施,从中研究分析提取录音设备个性特征。

(2)录音设备识别模型

在识别模型方面,GMM 在小样本性能较差,当特征呈非线性情况,分类能力大幅度下降;SVM 对大样本性能呈下降趋势,并且受惩罚因子、高斯宽度等参数影响较大。国内外研究文献显示,在录音设备识别中广泛采用的 SVM、GMM、SRC 等分类算法,都是模式识别中常用的分类方法,并没有有效针对录音设备识别情况进行专门设计,在特征降维、录音设备特征分离等方面有较大提升空间,而将深度学习理论应用于录音设备识别是一个值得探讨的话题。

(3)录音设备数据库建设

录音设备特征是通过音频信号进行呈现,单个或少数人的发音器官发出的声音和周围环境噪声可能并不能完全覆盖整个频段,不能完全携带录音设备特征,需要大量的音频信号来共同体现;由于电子产品更新换代太快,新的录音设备如手机不断推出,传统的录音设备数据库逐步退出历史舞台,录音设备数据库构建需要耗费大量的人力和时间,维护成本较高,而目前的录音设备数据库所覆盖的录音设备型号数量还是偏少;对各类人群如各年龄段、性别、方言等尚未形成全覆盖;目前的录音设备录制模式存在室内、室外、会议等多种模式,对音频数量和类别提出了更高的要求,因此,有必要建立综合各类(不同年龄男女、不同场合、不同方言、不同编码格式、不同手机品牌、不同手机型号)音频信号数据库,特别是同一型号不同手机在不同使用年限情况下的音频,对于音频设备识别尤其重要。

从音频取证角度研究录音设备并不局限于上述 3 个方面,还应该包括比如同一编码格式下不同录音设备差异性研究、同一设备不同编码格式下的特征变化研究、同一品牌不同型号的录音设备的不同特性区别、同一设备不同使用年限下的特征变化等等。这些实际上都依赖于录音设备数据库的建立健全。

参考文献:

[1] 工业和信息化部. 2017 年通信业统计公报[EB/OL]. (2018-02-02)[2018-07-31]. http://www.miit.gov.cn/n1146285/n1146352/n3054355/n3057511/n3057518/c6047251/content.html.

[2] 全国人民代表大会. 中华人民共和国民事诉讼法[EB/OL]. (2017-06-29)[2018-07-31]. http://www.npc.gov.cn/npc/ xin-

wen/2017-06/29/content_2024892.htm.
[3] Maher R C. Audio forensic examination[J]. IEEE Signal Processing Magazine, 2009, 26(2):84-94.
[4] 包永强, 梁瑞宇, 丛韫,等. 音频取证若干关键技术研究进展[J]. 数据采集与处理, 2016, 31(2):252-259.
Bao Yongqiang, Liang Ruiyu, Cong Yun, et al. Research progress on key technologies of audio forensics[J]. Journal of Data Acquisition and Processing, 2016, 31(2):252-259.
[5] Kraetzer C, Oermann A, Dittmann J, et al. Digital audio forensics: A first practical evaluation on microphone and environment classification[C]//Proc 9th Workshop on Multimedia and Security. Dallas, TX, USA:DBLP, 2007.
[6] Buchholz R, Kraetzer C, Dittmann J. Microphone classification using fourier coeffcients[C]//Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer,2010.
[7] Aggarwal R, Singh S, Roul A K, et al. Cellphone identification using noise estimates from recorded audio[C]//International Conference on Communications and Signal Processing. Guilin, Guangxi: IEEE, 2014:1218-1222.
[8] Rabiner L R, Sambur M R. An algorithm for determining the endpoints of isolated utterances[J]. Bell System Technical Journal, 1975, 54(2):297-315.
[9] Taboada J, Feijoo S, Balsa R, et al. Explicit estimation of speech boundaries[J]. Science, Measurement and Technology, IEEE Proceedings, 1994, 141(3):153-159.
[10] Marzinzik M, Kollmeier B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics[J]. IEEE Transactions on Speech & Audio Processing, 2002, 10(2):109-118.
[11] Yang S, Li Z G, Chen Y Q. A fractal based voice activity detector for internet telephone[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Vancouver, Canada: IEEE, 2003(1):I-808-I-811.
[12] Kobayashi Y, Niimi Y. Word boundary detection by pitch contours in an artificial language[C]//Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP. Denver, Colorado,USA: IEEE, 1980:900-903.
[13] Vahatalo A, Johansson I. Voice activity detection for GSM adaptive multi-rate codec[C]//Speech Coding Proceedings, 1999 IEEE Workshop on. Porvoo, Finland: IEEE, 1999:55-57.
[14] Lin C T, Lin J Y, Wu G D. A robust word boundary detection algorithm for variable noise-level environment in cars[J]. IEEE Transactions on Intelligent Transportation Systems, 2002, 3(1):89-101.
[15] Evangelopoulos G, Maragos P. Multiband modulation energy tracking for noisy speech detection[J]. IEEE Transactions on Audio Speech & Language Processing, 2006, 14(6):2024-2038.
[16] Haigh J A, Mason J S. Robust voice activity detection using cepstral features[C]//TENCON'93. Proceedings. Computer, Communication, Control and Power Engineering. 1993 IEEE Region 10 Conferenceon. Beijing, China: IEEE, 1993(3):321-324.
[17] 赵新燕,王炼红,彭林哲. 基于自适应倒谱距离的强噪声语音端点检测[J]. 计算机科学,2015,42(9):83-85,117.
Zhao Xinyan, Wang Lianhong, Peng Linzhe. Adaptive cepstral distance-based voice endpoint detection of strong noise[J]. Computer Science, 2015,42(9):83-85,117.
[18] Mcclellan S, Gibson J D. Variable-rate CELP based on subband flatness[J]. Speech & Audio Processing IEEE Transactions on, 1995, 5(2):120-130.
[19] Lorber J, Holt K S, Rendle-Short J, et al. Robust voice activity detection based on spectral entropy and two-stage mel-warped wiener filtering[C]//International Symposium on Intelligent Information Technology Application. Shanghai,China: IEEE, 2008:306-309.
[20] 李艳,成凌飞,张培玲. 一种基于改进谱熵的语音端点检测方法[J]. 计算机科学,2016,43(S2):233-236.
[21] Liu C H, Huang C C. Voice activity detector based on CAPDM architecture[J]. Electronics Letters, 2001, 37(1):68-69.
[22] Hussain A, Samad S A, Fah L B. Endpoint detection of speech signal using neural network[C]//TENCON 2000. Proceedings. Piscataway NJ, USA: IEEE, 2000(1):271-274.
[23] Ghisellicrippa T, Eljaroudi A. A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech[C]//International Conference on Acoustics, Speech, and Signal Processing. Tsukuba, Japan:IEEE, 2002:441-444.
[24] Couvreur L, Couvreur C. Wavelet-based non-parametric HMM's: Theory and applications[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway NJ, USA: IEEE, 2000: 604-607.

[25] Gazor S, Zhang W. A soft voice activity detector based on a Laplacian-Gaussian model[J]. Speech & Audio Processing IEEE Transactions on, 2003, 11(5):498-505.

[26] Selvakumari N A S, Radha V. A voice activity detector using SVM and Naïve Bayes classification algorithm[C]//International Conference on Signal Processing and Communication. Porto, Portugal: IEEE, 2018:1-6.

[27] 韩立华,王博,段淑凤. 语音端点检测技术研究进展[J]. 计算机应用研究, 2010, 27(4):1220-1226.
Han Lihua, Wang Bo, Duan Shufeng. Development of voice activity detection technology[J]. Application Research of Computers, 2010, 27(4): 1220-1226.

[28] Deller J R, Jr, Hansen J H L, Proakis J G. Discrete-timeprocessing of speech signals[M]. New York, NY, USA: Wiley-Interscience-IEEE, 2000.

[29] Buchholz R, Kraetzer C, Dittmann J. Microphone classification using Fourier coefficients[M]//Information Hiding. [S. l.]: Springer-Verlag, 2009:235-246.

[30] Pandey V, Verma V K, Khanna N. Cell-phone identification from audio recordings using PSD of speech-free regions[C]// Electrical, Electronics and Computer Science. Hangzhou, China: IEEE, 2014:1-6.

[31] Marple L. Exponential energy spectral density estimation[C]//Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP. Denver, Colorado, USA: IEEE, 1980:588-591.

[32] Hirsch H G. Improved speech recognition using high-pass filtering of subband envelopes[J]. Proc Eurospeech, 1991, 1(1): 413-416.

[33] Hirsch H G, Ehrlicher C. Noise estimation techniques for robust speech recognition[C]// International Conference on Acoustics, Speech, and Signal Processing. Orlando, Florida, SA: IEEE, 2002(1):153-156, 1995.

[34] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics[J]. IEEE Trans Speech & Audio Processing, 2001, 9(5):504-512.

[35] Freudenberger J, Stenzel S, Venditti B. A noise PSD and cross-PSD estimation for two-microphone speech enhancement systems[C]// Statistical Signal Processing, 2009. SSP'09. IEEE/SP, Workshop on. Piscataway NJ, USA: IEEE, 2009:709-712.

[36] Yousefian N, Loizou P C. A dual-microphone speech enhancement algorithm based on the coherence function[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 20(2):599.

[37] Doclo S, Jensen J. Maximum likelihood PSD estimation for speech enhancement in reverberation and noise[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(9):1595-1608.

[38] Taseska M, Habets E A P. Non-stationary noise PSD matrix estimation for multichannel blind speech extraction[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2017(99):1.

[39] Hendriks R C, Gerkmann T. Noise correlation matrix estimation for multi-microphone speech enhancement[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 20(1):223-233.

[40] Shin H S, Fingscheidt T, Kang H G. A priori SNR estimation using air- and bone-conduction microphones[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2015, 23(11):2015-2025.

[41] Cohen I, Berdugo B. Noise estimation by minima controlled recursive averaging for robust speech enhancement[J]. Signal Processing Letters IEEE, 2002, 9(1):12-15.

[42] Souden M, Chen J, Benesty J, et al. An integrated solution for online multichannel noise tracking and reduction[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 19(7):2159-2169.

[43] Schwartz B, Gannot S, Habets E A P. Two model-based em algorithms for blind source separation in noisy environment[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2017(99):1.

[44] Milner B, Darch J. Robust acoustic speech feature prediction from noisy mel-frequency cepstral coefficients[J]. IEEE Transactions on Audio Speech & Language Processing, 2010, 19(2):338-347.

[45] Al-Ali A K H, Dean D, Senadji B, et al. Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions[J]. IEEE Access, 2017, 5(99):15400-15413.

[46] Hermansky H. Perceptual linear predictive (PLP) analysis of speech[J]. Journal of the Acoustical Society of America, 1990, 87(4):1738-1752.

[47] Vachhani B B, Patil H A. Use of PLP cepstral features for phonetic segmentation[C]//International Conference on Asian

Language Processing. Urumqi, China: IEEE, 2013:143-146.

[48] Salehi H, Parsa V. Nonintrusive speech quality estimation based on perceptual linear prediction[C]//Electrical and Computer Engineering. Vancouver, Canada:IEEE, 2016:1-4.

[49] Daly I, Hajaiej Z, Gharsallah A. Speech analysis in search of speakers with MFCC, PLP, Jitter and Shimmer[C]//International Conference on Advanced Systems and Electric Technologies. Cebu, Philippines: IEEE, 2017:291-294.

[50] Panagakis Y, Kotropoulos C. Automatic telephone handset identification by sparse representation of random spectral features [C]// Proceedings of the on Multimedia and security. Coventry, UK:ACM, 2012:91-96.

[51] Garcia-Romero D, Espy-Wilson C Y. Automatic acquisition device identification from speech recordings.[C]// IEEE International Conference on Acoustics Speech and Signal Processing. [S. l.]: IEEE, 2010:1806-1809.

[52] Sanchis A, Juan A, Vidal E. A word-based Naïve Bayes classifier for confidence estimation in speech recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 20(2):565-574.

[53] Anzola J, Cuartas K A, Tarazona G M. Classification methodology of research topics based in decision trees: J48 and Randomtree[J]. International Journal of Applied Engineering Research, 2015, 10(8):19413-19424.

[54] Ananthakrishna T, Shama K, Niranjan U C. k-means nearest neighbor classifier for voice pathology[C]// India Conference, 2004. Proceedings of the IEEE Indicon. Kharagpur, India: IEEE, 2004:352-354.

[55] Jia M K K, Ambikairajah E, Epps J, et al. Speaker verification using sparse representation classification[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic: IEEE, 2011:4548-4551.

[56] Austin P C. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality[J]. Statistics in Medicine, 2010, 26(15):2937-2957.

[57] Matsui T, Tanabe K, Siniscalchi S M, et al. Penalized logistic regression with HMM log-likelihood regressors for speech recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2010, 18(6):1440-1454.

[58] Hanilci C, Ertas F, Ertas T, et al. Recognition of brand and models of cell-phones from recorded speech signals[J]. IEEE Transactions on Information Forensics & Security, 2012, 7(2):625-634.

[59] Kinnunen T, Rajan P. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013: 7229-7233.

[60] Kocal O H, Yuruklu E, Avcibas I. Chaotic-type features for speech steganalysis[J]. IEEE Transactions on Information Forensics & Security, 2008, 3(4):651-661.

[61] Campbell W M, Campbell J P, Gleason T P, et al. Speaker verification using support vector machines and high-level features [J]. IEEE Trans Audio Speech & Language Processing, 2007, 15(7):2085-2094.

[62] 任浩,叶亮,李月,等.基于多级 SVM 分类的语音情感识别算法[J].计算机应用研究,2017,34(6):1682-1684.
Ren Hao, Ye Liang, Li Yue, et al. Speech emotion recognition algorithm based on multi-layer SVM classification[J]. Application Research of Computers, 2017, 34(6): 1682-1684.

[63] 蒋海华,胡斌.基于 PCA 和 SVM 的普通话语音情感识别[J].计算机科学,2015,42(11):270-273.
Jiang Haihua, Hu Bin. Speech emotion recognition in mandarin based on PCA and SVM[J]. Computer Science, 2015, 42 (11): 270-273.

[64] 姜晓庆,夏克文,夏莘媛,等.采用半定规划多核 SVM 的语音情感识别[J].北京邮电大学学报,2015,38(S1):67-71.
Jiang Xiaoqing, Xia Kewen, Xia Xinyuan, et al. Speech emotion recognition using semi-definite programming multiple kernel SVM[J]. Journal of Beijing University of Posts and Telecommunication, 2015, 38(S1): 67-71.

[65] Adankon M M, Cheriet M. Learning semi-supervised SVM with genetic algorithm[C]//International Joint Conference on Neural Networks. Anchorage, Alaska: IEEE, 2007: 1825-1830.

[66] Yeh J P, Chiang C M. Reducing the solution of support vector machines using simulated annealing algorithm[C]//International Conference on Control, Artificial Intelligence, Robotics & Optimization. Prague, Czech Republic: IEEE Computer Society, 2017:105-108.

[67] Demidova L, Nikulchev E, Sokolova Y. Big data classification using the SVM classifiers with the modified particle swarm optimization and the SVM ensembles[J]. International Journal of Advanced Computer Science & Applications, 2016, 9(5): 294-312.

[68] Lei Z, Yang Y, Wu Z. Mixture of support vector machines for text-independent speaker recognition[C]//INTERSPEECH 2005—Eurospeech, European Conference on Speech Communication and Technology. Lisbon, Portugal: DBLP, 2005:2041-2044.

[69] Campbell W M. Generalized linear discriminant sequence kernels for speaker recognition[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Florida, USA: IEEE, 2002:I-161-I-164.

[70] Reynolds D A. Automatic speaker recognition using Gaussian mixture speaker models[J]. Speech Communication, 1995, 17(1/2):91-108.

[71] Singh M, Mishra J, Pati D. Replay attack: Its effect on GMM-UBM based text-independent speaker verification system [C]//IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering. Gorakhpur, India:IEEE, 2017:619-623.

[72] Maurya A, Kumar D, Agarwal R K. Speaker recognition for Hindi speech signal using MFCC-GMM approach[J]. Procedia Computer Science, 2018, 125:880-887.

[73] Abrol V, Sharma P, Sao A K. Greedy dictionary learning for kernel sparse representation based classifier [J]. Pattern Recognition Letters, 2016, 78(C):64-69.

[74] Liu Q. Kernel local sparse representation based classifier[J]. Neural Processing Letters, 2016, 43(1):85-95.

[75] Kotropoulos C, Samaras S. Mobile phone identification using recorded speech signals[C]//International Conference on Digital Signal Processing. Kuala Lumpur, Malaysia:IEEE, 2014:586-591.

[76] Ossama A H, Mohamed A R, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012:4277-4280.

[77] Mitra V, Franco H. Time-frequency convolutional networks for robust speech recognition[C]//Automatic Speech Recognition and Understanding. Okinawa, Japan: IEEE, 2016:317-323.

[78] Hrúz M, Zajíc Z. Convolutional neural network for speaker change detection in telephone speaker diarization system[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA: IEEE, 2017:4945-4949.

[79] Cakir E, Parascandolo G, Heittola T, et al. Convolutional recurrent neural networks for polyphonic sound event detection [J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2017, 25(6):1291-1303.

[80] Reynolds D A. HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Munich, Germany: IEEE Computer Society, 1997:1535.

[81] Zou L, Yang J, Huang T. Automatic cell phone recognition from speech recordings[C]//IEEE China Summit & International Conference on Signal and Information Processing. Xi'an, China: IEEE, 2014:621-625.

作者简介:

包永强(1973-),男,博士,教授,研究方向:语音信号处理、音频取证,E-mail:jybyq@163.com。

梁瑞宇(1978-),男,博士,副教授,研究方向:语音信号处理,E-mail:lly1711@163.com。

王青云(1972-),女,博士,教授,研究方向:语音信号处理。

(编辑:张 彤)