自然环境背景噪声下基于低维深度特征的手机来源识别

苏兆品[1,2,3,4],吴张倩[2],岳 峰[2,4],武钦芳[2],张国富[1,2,3,4]

(1. 大数据知识工程教育部重点实验室(合肥工业大学),安徽合肥 230601; 2. 合肥工业大学计算机与信息学院,安徽合肥 230601;
3. 智能互联系统安徽省实验室,安徽合肥 230009; 4. 工业安全与应急技术安徽省重点实验室(合肥工业大学),安徽合肥 230601)

摘 要: 基于语音的手机来源识别是近年来多媒体取证领域中的一个研究热点,但已有研究大都局限于纯净语音或人工背景噪声语音. 本文以自然环境背景噪声下的手机语音为研究对象,提出一种基于低维深度特征的手机来源识别方法. 首先提取对数域的 Mel 滤波器组系数作为基本的声学特征,然后输入到时间卷积网络中进行训练,进一步提取能够表征语音设备的深度特征,并利用线性判别分析进行降维,去除高维深度特征中的冗余. 最后,将得到的低维深度特征输入到支持向量机中进行分类和识别. 在 47 种不同型号手机录制的 37600 条自然环境背景噪声语音样本库上的测试结果表明,本文所提方法在自然环境背景噪声下具有更优的识别性能,且对不同品牌、相同品牌不同型号、不同样本长度、不同数据集规模和不同采样率都具有很好的适应性.

关键词: 手机来源识别; 自然环境背景噪声; 低维深度特征; 时间卷积网络; 线性判别分析

中图分类号: TN912. 3 **文献标识码:** A **文章编号:** 0372-2112 (2021)04-0637-10

电子学报 URL: http://www. ejournal. org. cn **DOI:** 10. 12263/DZXB. 20200658

Source Cell-Phone Identification Under Background Noise Based on Low-Dimensional Deep Features

SU Zhao-pin[1,2,3,4], WU Zhang-qian[2], YUE Feng[2,4], WU Qin-fang[2], ZHANG Guo-fu[1,2,3,4]

(1. *Ministry of Education Key Laboratory of Knowledge Engineering with Big Data* (*Hefei University of Technology*), *Hefei*, *Anhui* 230601, *China*;
2. *School of Computer Science and Information Engineering*, *Hefei University of Technology*, *Hefei*, *Anhui* 230601, *China*;
3. *Intelligent Interconnected Systems Laboratory of Anhui Province* (*Hefei University of Technology*), *Hefei*, *Anhui* 230009, *China*;
4. *Anhui Provincial Key Laboratory of Industry Safety and Emergency Technology* (*Hefei University of Technology*), *Hefei*, *Anhui* 230601, *China*)

Abstract: Identifying cell-phones using recorded speech has become a hot topic in the field of multimedia forensics in recent years. However, most of the existing studies focus on the clean speech or the speech with unnaturally artificial noise. In this paper, the speech with background noise is taken into account and a source cell-phone identification method is presented on the basis of the low-dimensional deep features. First, the logarithmic Mel-filter bank coefficients are extracted as the main acoustic features and input to the temporal convolutional network for training and further extracting the deep features of speech devices. Then, the linear discriminant analysis is used to reduce the size of the high-dimensional deep features and remove the redundancy. Finally, the low-dimensional deep features are used as input to the support vector machine classifier. The experimental results on 47 models of mobile phones and 37600 speech samples with background noise show that the proposed method has better recognition performance and better adaptability to different brands, different models of the same brand, different sampling lengths, different sizes of the dataset, and different sampling rates.

Key words: source cell-phone identification; background noise; low-dimensional deep features; temporal convolutional network; linear discriminant analysis

收稿日期:2020-07-08;修回日期:2020-12-02;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61573125);教育部人文社会科学研究青年基金(No. 19YJC870021, No. 18YJC870025);安徽省重点研究与开发计划(No. 202004d07020011);中央高校基本科研业务费专项资金(No. PA2020GDKC0015, No. PA2019GDQT0008, No. PA2019GDPK0072)

1 引言

在众多的法律纠纷处理过程中,尤其是在民事诉讼过程中,手机语音有时起着关键性的作用,但是手机语音是否可以作为有效证据被法庭采纳的一个先决条件是需要确定其真伪,而对手机语音的来源设备识别是语音证据鉴真的一个根本前提,已成为近年来多媒体取证领域中的一个研究热点[1].

主流的手机来源识别研究大都从整体语音信号上提取设备特征. Zou 等[2]利用高斯混合模型和通用背景模型设计一种基于 Mel 频率倒谱系数(Mel Frequency Cepstral Coefficient,MFCC)和功率归一化倒谱系数的识别方法. Luo 等[3]利用不同生产商在音频采集管道上不尽相同带来的音频取证上的微小差异,提出一种新的带能量描述符(Band Energy Descriptor,BED)特征,并使用支持向量机(Support Vector Machine,SVM)进行手机设备识别. Qin 等[4]提出了一种基于常数 Q 变换域(Constant Q Transform,CQT)的语音特征,并使用卷积神经网络(Convolutional Neural Networks,CNN)进行训练. Jiang 和 Leung[5]以 MFCC 和线性频率倒谱系数作为特征向量,并基于加权多数表决的加权 SVM 进行分类. Li 等[6]基于深度自动编码网络学习深度表示特征来刻画每个手机在录音中留下的内在痕迹,并使用谱聚类将同一手机获得的录音合并到一个单一的簇中来进行分类. Verma 等[7]提取录音的低频和高频区域中存在的设备特定信息作为被动签名来识别录音的源手机. 上述方法能够充分挖掘语音信号中的设备关键信息,取得了较好的识别效果,但需要处理整个音频信号,计算开销很大.

为了排除语音部分的干扰,充分挖掘手机设备的本质特征,语音信号的非语音部分越来越受到关注. Qi 等[8]从背景噪声中提取语音特征,并对比分析了在不同深度学习分类器下的分类性能. Jin 等[9]从手机语音的自噪声中提取频谱形状特征和频谱分布特征. 裴安山等[10,11]将本底噪声作为手机的指纹用于识别,并通过使用自适应端点检测算法得到语音的静音段,然后将静音段中对数域的 Mel 滤波器组系数(Logarithmic Mel-Filter Bank Coefficients,Fbank)降维后作为分类特征. Baldini 等[12,13]在不同频率下用非语音声音刺激内置麦克风,利用手机内置麦克风的固有物理特性构建 CNN 对手机进行识别. 上述研究可以有效降低计算开销,但忽视了语音部分的特征,可能会丧失一些设备特定信息,从而影响识别效果.

总的说来,基于语音信号的手机来源识别其本质是挖掘语音信号中含有的设备元器件(主要是麦克风)自身的噪声(主要是高斯噪声)特征来进行识别,而这种特征存在于整个频率范围. 虽然 Qin 等[4]在 CQT 域提取特征,关注了语音信号在不同频率的分辨率在不同手机设备上的区别,但由于在执行 CQT 变化时部分频谱泄露,需要基于实验室环境中的纯净语音或人工背景噪声语音,且使用大量的语音样本进行训练才能取得较好的效果. 在音频取证中,充当证据的手机语音信号通常产生于人们交流和交易协商的生活和工作环境,而不是理想而又安静的实验室环境,也很少是纯粹的静音片段,而是包含了各种自然环境背景噪声. 特别的,自然环境背景噪声复杂多变,在不同天气、地点,背景噪声信号对手机语音信号产生的影响也不同. 有时候,强自然环境背景噪声甚至可能会完全掩盖设备本身的噪声. 在对这些手机语音进行特征提取时,如何降低自然环境背景噪声对手机设备本身噪声的干扰是一个难点问题.

基于上述背景,本文针对司法领域对手机语音证据的鉴真需求,以自然环境背景噪声下的手机语音为研究对象,首先提取手机语音中更能全面反映语音信息的 Fbank 特征,并输入到时间卷积网络(Temporal Convolutional Network,TCN)[14]中进行训练,更大程度的在整个频谱范围内挖掘能够表征语音设备的深度特征,然后利用线性判别分析(Linear Discriminant Analysis,LDA)[15]进行降维,去除高维深度(High-Dimensional Deep,HDD)特征中的冗余,并结合 SVM 提出一种基于低维深度(Low-Dimensional Deep,LDD)特征的手机来源识别方法,最后基于创建的自然环境噪声手机语音库对所提方法进行测试和验证.

2 基于 LDD 特征的手机来源识别框架

基于 LDD 特征的手机来源识别框架如图 1 所示. 针对手机语音训练集,首先提取语音信号中的 Fbank 特征,以保留相邻特征的相关性和更高的维度. 然后将 Fbank 特征作为 TCN 的输入,让 TCN 自主训练去提取语音信号中的设备噪声特征. 紧接着,将 TCN 中最后一个卷积层输出的 HDD 特征利用 LDA 进行降维,去除 HDD 特征中的冗余,从而得到语音信号中设备噪声的 LDD 特征. 最后,利用 LIBSVM 工具包[16]对提取的 LDD 特征进行训练建立手机设备多分类模型. 对于手机语音的测试集,也需要基于相同的步骤得到测试集的 LDD 特征,然后与 SVM 输出的多分类模型进行模型匹配即可完成手机来源识别.

为了更加清晰的说明图 1 所示的基于 LDD 特征的手机来源识别框架,在下面的小节中,本文将详细介绍识别框架中的一些关键步骤.

2.1 Fbank 特征提取

为了更好的保留语音中的设备特征信息,本文利用 LibROSA 工具包[17]提取 Fbank 特征作为基本语音特征,其提取流程简要描述如下.

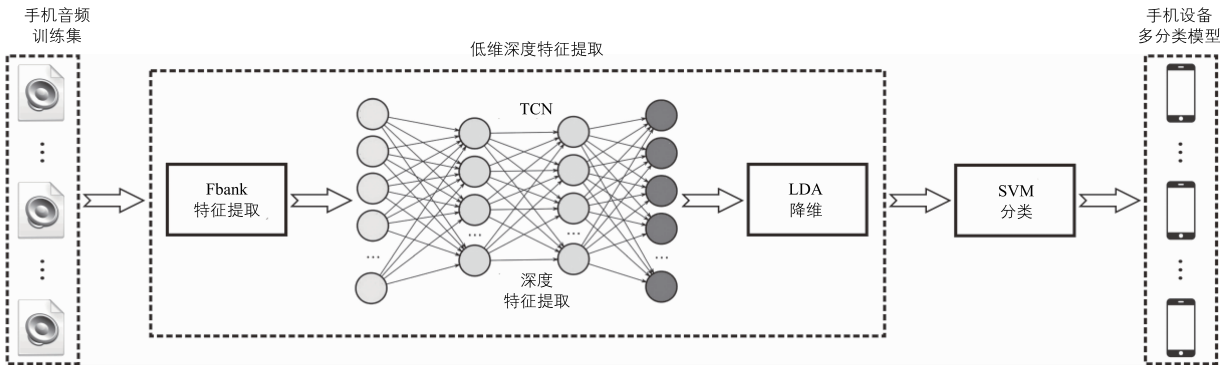


图1　基于LDD特征的手机来源识别框架

(1)分帧:语音信号具有短时平稳性,将语音信号分成短时帧,可以有效保证每个帧内的信号是相对稳定的,从而有利于后续的语音信号处理. 此外,为了避免相邻两帧的变化过大,需要在两相邻帧之间设置一段重叠区域. 在本文方法中,采用 LibROSA 工具包的默认推荐设置,即帧长 $N=2048$ 个采样点,帧移为 512 个采样点.

(2)加窗:为了在一定程度上消除分帧后出现的帧与帧之间的不连续性,减小频谱泄露,使用汉宁窗对每帧进行加窗处理:

$$s_i'(n)=\frac{1}{2}\left[1-\cos\left(\frac{2\pi n}{N-1}\right)\right]s_i(n) \tag{1}$$

其中,$0\leqslant n\leqslant N-1$;$s_i(n)$个表示第 i 帧的第 n 个采样点加窗前的采样值;$s_i'(n)$为加窗后的采样值.

(3)快速傅里叶变换:对加窗后的每帧信号,逐帧进行快速傅里叶变换(Fast Fourier Transform,FFT),然后根据输出的频点矩阵计算能量谱图:

$$P=|FFT(s_i')|^2 \tag{2}$$

(4)Mel 滤波器:为了将能量谱转换为更接近人耳机理的 Mel 频率,构造 Mel 滤波器组并与能量谱进行点积运算得到 Mel 频谱图. 在本文方法中,采用 LibROSA 工具包的默认设置,即 Mel 滤波器数设为 128.

(5)取对数:为了模拟人耳的对数式特性,对 Mel 频谱图取对数即可得到 Fbank 特征(dB):

$$\boldsymbol{X}'=10\lg[\max(amin,S)] \tag{3}$$

$$\boldsymbol{X}'-=10\lg[\max(amin,ref)] \tag{4}$$

$$\boldsymbol{X}=\max[\boldsymbol{X}',\max(\boldsymbol{X}')-80] \tag{5}$$

其中,S 为 Mel 频谱图,$amin=1e-10$ 是 $abs(S)$ 和 $ref=1$ 的最小阈值,$\boldsymbol{X}\in\mathbf{R}^{128\times M}$为一个实数矩阵,$M$ 为前面对语音信号分帧后的总帧数,每一帧都有 128 维特征.

2.2　基于 TCN 的深度特征提取

本文将提取的 Fbank 特征 $\boldsymbol{X}$ 作为 TCN 的输入,利用 TCN 学习设备噪声深度特征. 本文所设计的 TCN 网络结构如图 2 所示,在整体网络中还多次利用了加速神经网络训练的 BatchNorm 算法[18],以提高收敛速度和稳定性.

(1)首先,Fbank 特征输入后先经过一维卷积过滤:

$$\boldsymbol{Y}_0=\sigma_1(W_0*\boldsymbol{X}) \tag{6}$$

其中,W_0 是该层网络需要学习的参数;σ_1 为非线性激活函数 Tanh;$\boldsymbol{Y}_0$ 为是第一层网络的输出.

(2)进入残差模块进一步学习. 残差网络模型 ResNet[14]可以有效解决随着神经网络层数变多拟合效果反而变差的问题. 如图 2 所示,在 TCN 架构中引入残差单元 Res_unit. 每一个 Res_unit 中的卷积核个数是 128,且全部采用扩张卷积,其中最关键的参数 dilation rate(即 d)在连续残差单元间以 2 的指数形式增加,$d=2^n$,$n\in\{0,1,2,3,4\}$,这使得能够在不显著增加参数数量的情况下,可在很大程度上增加感受野. 每个 Res_unit 的输出通过添加到下一个 Res_unit 的输入进行合并. 令$\boldsymbol{Y}_j$ 代表第 j 层残差单元的输出,$j\in\{1,2,3,4,5\}$,则有:

$$\boldsymbol{Y}_j=\boldsymbol{Y}_{j-1}+F(W_j,\boldsymbol{Y}_{j-1}) \tag{7}$$

其中,W_j 是第 j 层 Res_unit 需要学习的参数;F 是在 Res_unit 中经历的非线性变换,如图 2 所示,将输入信号进

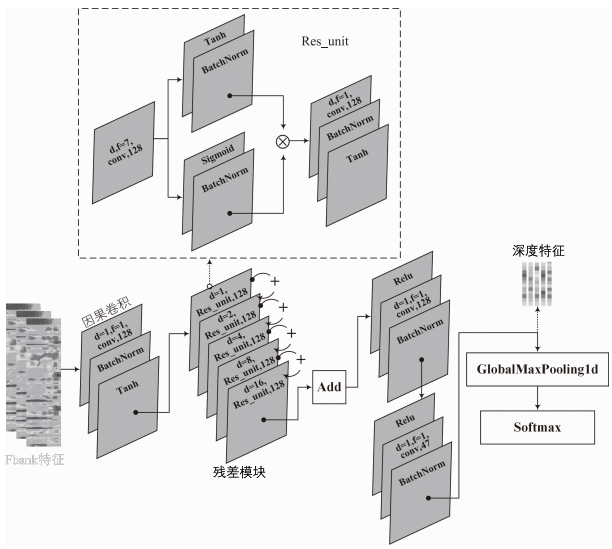


图2　TCN网络结构

行卷积之后分别利用不同的激活函数 Sigmoid 和 Tanh 进行线性变换,并将结果相乘,其中的 Sigmoid 与 Tanh 相乘,相当于给每一维特征加权,提高学习到的特征性能和模型泛化能力. 相乘后再经过一维卷积和 Tanh 激活函数:

$$F(W_j, \boldsymbol{Y}_{j-1}) = \sigma_1\{W_{j_2} * [\sigma_1(W_{j_1} * \boldsymbol{Y}_{j-1})\sigma_2(W_{j_1} * \boldsymbol{Y}_{j-1})]\} \tag{8}$$

其中,σ_2 是 Sigmoid 非线性激活函数,W_{j_1}和 W_{j_2}分别代表在第 j 层 Res_unit 中第一层 conv 和第二层 conv 的参数,满足 $W_j = W_{j_1} + W_{j_2}$. 在经过 5 个 Res_unit 的学习后,累加(Add)不同输出,再利用 σ_3 激活函数 Relu 进行非线性变换可得$\boldsymbol{Y}_5$:

$$\boldsymbol{Y}_5 = \sigma_3\left[\boldsymbol{Y}_0 + \sum_{j=1}^{5} F(W_j, \boldsymbol{Y}_{j-1})\right] \tag{9}$$

可见,TCN 中的所有 Res_unit 均累加上$\boldsymbol{Y}_0$. 利用 TCN 网络学习不同语音信号中有区别的语音特征,整个模型的表示能力在很大程度上要取决于通过 W_0 中的滤波器产生的$\boldsymbol{Y}_0$.

(3)在残差模块之后又添加两层卷积层:

$$\boldsymbol{Y}_6 = \sigma_3(W_6 * \boldsymbol{Y}_5) \tag{10}$$

$$\boldsymbol{Y}_7 = W_7 * \boldsymbol{Y}_6 \tag{11}$$

其中,两个卷积层的卷积核个数分别是 128 和 47.

(4)最后,应用全局平均池化,将数据由三维降到二维,减少训练参数的同时提高模型的泛化能力,并附加一个神经元数量等于类数量的 softmax 层:

$$\boldsymbol{Y}_8 = \text{GlobalMaxPooling}(\boldsymbol{Y}_7) \tag{12}$$

$$\hat{\boldsymbol{Y}} = \text{softmax}(\boldsymbol{Y}_8) \tag{13}$$

其中,$\hat{\boldsymbol{Y}} \in [0,1]$是 TCN 的预测结果.

需要特别指出的是,对于常规的 TCN 应用,最后一个隐含层的输出$\boldsymbol{Y}_7$ 是一个 HDD 特征,而全局平均池化层和 softmax 层是对这个 HDD 特征进行数据的强制降维以实现分类和识别. 但当数据集规模较小、且训练不够充分时,这种强制性的降维必然会导致 TCN 的分类准确率大幅下降. 因此,本文保留了完整的 TCN 训练结构,但并没有利用全局平均池化层和 softmax 层的设置,而是针对最后一个卷积层输出的 HDD 特征$\boldsymbol{Y}_7$,利用有效的降维算法来实现 LDD 特征的提取.

2.3 基于 LDA 的 LDD 特征提取

正如前述,TCN 提取的深度特征$\boldsymbol{Y}_7 \in R^{128\times 47}$维数大、相关性强,数据在每个特征维度的分布稀疏,若直接将$\boldsymbol{Y}_7$送入 SVM 训练,SVM 将耗费大量的机器内存和运算时间,而且特征的强相关性也会导致 SVM 训练效果有限. 因此,本文基于 LDA[15] 对 HDD 特征进行优化,进一步提取音频信号中的 LDD 特征,如图 3 所示,具体过程描述如下.

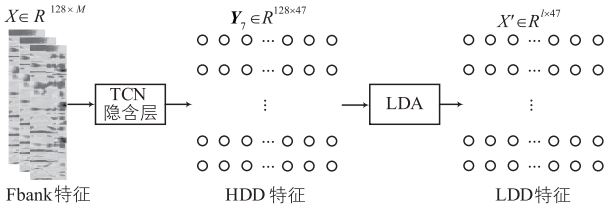


图3 LDD特征的提取

(1)首先,将 HDD 特征$\boldsymbol{Y}_7$ 重塑成一维,得到数据集 $X = \{\boldsymbol{x}_k\}, k \in \{1,\cdots,K\}$ 表示训练样本的类别,对应某一类别的手机,K 为总类别数. $\boldsymbol{x}_k$ 表示第 k 类样本的集合,其中的每个样本为一个 6016 维的特征向量.

(2)计算类内散度矩阵:

$$\boldsymbol{Q}_1 = \sum_{k=1}^{K}\sum_{x\in\boldsymbol{x}_k}(x - \bar{\boldsymbol{x}}_k)(x - \bar{\boldsymbol{x}}_k)^{\mathrm{T}} \tag{14}$$

其中,$\bar{x}_k$ 表示第 k 类样本均值.

(3)计算类间散度矩阵:

$$\boldsymbol{Q}_2 = \sum_{k=1}^{K} m_k(\bar{x}_k - \bar{\boldsymbol{x}})(\bar{x}_k - \bar{\boldsymbol{x}})^{\mathrm{T}} \tag{15}$$

其中,m_k 为第 k 类样本的样本数,$\bar{\boldsymbol{x}}$ 为所有样本均值.

(4)计算矩阵 $\boldsymbol{Q}_1^{-1}\boldsymbol{Q}_2$,并对其进行奇异值分解,得到奇异值 λ_k 及其对应的特征向量$\boldsymbol{w}_k, k \in \{1,\cdots,K-1\}$.

(5)对 λ_k 进行倒序排序,取前 l 个 λ_k 对应的特征向量组成投影矩阵 $\boldsymbol{W}$.

(6)计算样本集中每类样本$\boldsymbol{x}_k$ 在新的低维空间的投影:

$$\boldsymbol{z}_k = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_k \tag{16}$$

得到降维后的样本集 $X' = \{\boldsymbol{z}_k\}, k \in \{1,\cdots,K\}$. $\boldsymbol{z}_k$ 表示第 k 类样本的集合,其中的每个样本为一个 l 维特征向量. 本文将优化后的 LDD 特征 $X' \in R^{l\times 47}$输入到 SVM 中进行训练,建立手机设备多分类模型.

3 实验结果与分析

3.1 手机语音库的建立

为了搜集尽可能多的自然环境背景噪声下的手机语音,我们搭建了一个音频信号网络搜集平台,任意用户可通过该平台上传 5 ~ 15min 的 MP3 语音,而且用户性别、年龄、所处环境和语音内容均不受限制. 所处环境包括室内、操场、地铁站、马路边等,语音内容包括日常对话、电影对话、无线电广播等. 我们将搜集到的语音经过处理、筛选,构建了一个具有自然环境背景噪声的手机语音数据库,如表 1 所示,共包含 10 个品牌、47 种型号手机设备信息,每种型号手机对应不同的 ID.

手机设备信息(主要是麦克风)是固有特征,存在于语音信号的整个过程. 也就是说,无论是有声段还是

静音段都含有固有设备特征信息. 然而,本文主要面向自然环境背景噪声下的手机来源识别,所构建的手机语音数据库包含各种实际背景噪声,不存在理想环境下的绝对静音帧. 因此,本文将表 1 中每种型号手机的语音信号平均分割成 1s,每个 ID 手机共收集 800 条语音片段,其中 600 条用于训练,其余 200 条用于测试. 语音数据库一共包含 37600 条语音样本,其中训练库有 28200 条语音,测试库包括 9400 条语音. 本文的所有实验均是基于以上语音数据库进行测试和分析.

表 1　每种手机的品牌型号及其对应的 ID

ID	品牌与型号	ID	品牌与型号
1	Hongmi_2a	25	Oppo_r9plus
2	Hongmi_5plus	26	Oppo_a57
3	Honor_10	27	Oppo_findx
4	Honor_7x	28	Oppo_r7s
5	Honor_8lite	29	Samsung_galaxy-s4
6	Honor_9	30	Vivo_y35a
7	Honor_v9	31	Vivo_v3maxa
8	Honor_v10	32	Vivo_x20
9	HuaweiBND_al10	33	Vivo_x6
10	Huawei_aletl00	34	Vivo_x6s
11	Huawei_p9	35	Vivo_x9sl
12	Huawei_mate9	36	Xiaomi_4c
13	Huawei_nova3i	37	Xiaomi_5
14	Huawei_p8	38	Xiaomi_5sp
15	Huawei_nce-al00	39	Xiaomi_6
16	IPhone6_sp	40	Xiaomi_8
17	IPhone_7	41	Xiaomi_8lite
18	IPhone_8	42	Xiaomi_mix3
19	IPhone_6plus	43	Xiaomi_max2
20	IPhone_x	44	Xiaomi_max2s
21	Nubia_ui-v5	45	Xiaomi_note1
22	Oppo_a79	46	Xiaomi_note3
23	Oppo_r11	47	One_plusone
24	Oppo_r9	–	–

3.2　参数设置与评价指标

对于 TCN,我们选择交叉熵损失函数 categorical_crossentropy 和最常用的优化器 Adam,通常其学习率固定为 0.01. 特别的,我们测试了关键参数训练周期对 TCN 性能的影响,如图 4 所示,在训练周期大于 20 时,train 和 valid 的损失和精度基本趋于稳定. 因此,为了保证充分的学习,本文最终将训练周期设定为 30. TCN 的其他参数已在图 2 的网络结构中给出.

对于 LDA,作为一种有监督的降维方法,最多可以降到类别数 $K-1$ 的维数. 因此,LDD 的特征维数设置为 $l=46$. 对于 SVM,本文选择径向基函数作为分类器的核函数,将惩罚系数和径向基函数参数均设为 1.0[16].

为了充分评估所提方法的性能,本文引入如下四种在机器学习中常用性能指标[19]:准确率(Accuracy)、查准率(Precision)、查全率(Recall)和 F1 分数(F1-score). Accuracy 是使用的最普遍的,也是最直观的性能指标,表示预测正确的样本占所有样本的比例,表示了一个分类器的区分能力;Precision 是指在预测结果为正例的样本里,真实情况也为正例所占的比率;Recall 是指在真实情况为正例的所有样本中,预测结果也为正例的样本所占的比率;F1-score 是 Precision 和 Recall 的一个加权平均,兼顾了分类模型的查准率和查全率.

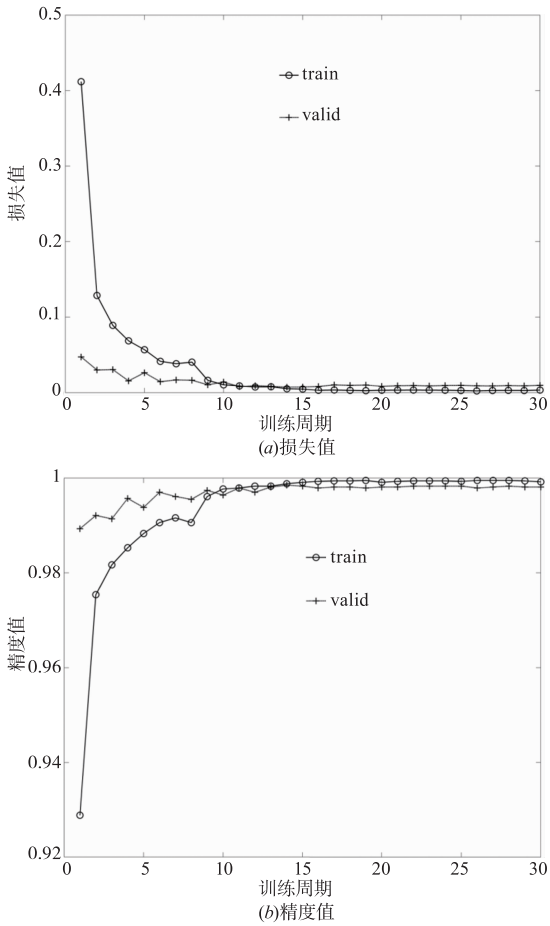
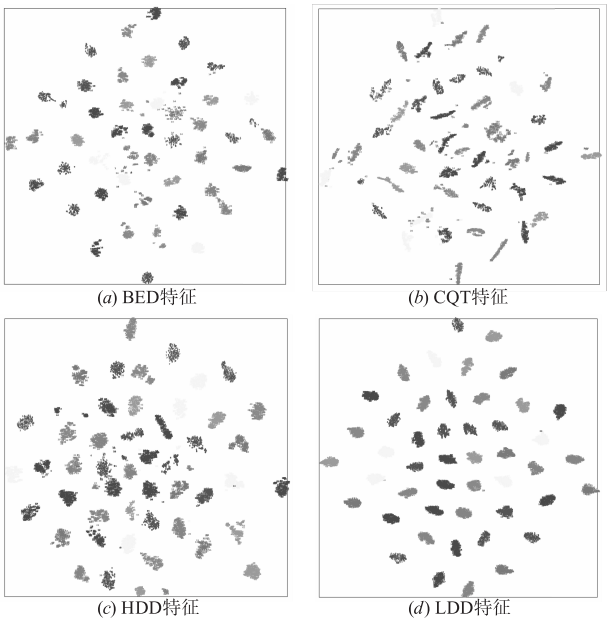


图4　训练周期对TCN 损失值和精度值的影响

3.3　不同语音特征的对比

在第一个实验中,为了验证本文所提的 LDD 特征的有效性,将其与已有的 BED 特征[3]和 CQT 特征[4]进行对比分析.

图 5 分别给出了 BED 特征、CQT 特征、TCN 提取出的 HDD 特征和基于 LDA 提取的 LDD 特征的 t-SNE(t-Stochastic Neighbor Embedding)[20]可视化结果. t-SNE 方法能够同时保持原有数据的全体与局部结构的特性,可以全面的反应不同特征的分类能力. 可以看出,BED 和 CQT 特征的分类效果已经较好,且二者的可分性不相上下,大多数设备可形成明显可分离的簇. 这是因为,BED 特征关注语音信号的傅里叶变换后的能量值差异,可以很好捕捉到不同品牌手机设备之间的细微



(*a*) BED特征 (*b*) CQT特征

(*c*) HDD特征 (*d*) LDD特征

图5 不同特征的t-SNE可视化结果

差异,而 CQT 关注的是中、低频频带的特征,与固定时频分辨率的短时傅里叶变换相比,具有更高的低频频率分辨率和高频时间分辨率. 但是,虽然不同品牌之间手机设备的差异较明显,但同一品牌不同型号手机设备之间的相似度较高,仍然有一些极其相似的手机设备无法分离出来,簇与簇之间非常接近. HDD 特征是基于 Fbank 的深度特征,其可分性与 CQT 和 BED 特征已经很接近,但有些簇仍然存在着关联,而 LDD 特征的可分性要显著优于其他三种特征,簇与簇之间的区别非常的明显. 这是因为,Fbank 特征具有很大的相关性,充分保留了语音信号中的有效信息,经过 TCN 暴力提取的 HDD 由于充分挖掘了 Fbank 保留的有效特征,可分性已经显著提升,再经过 LDA 去除冗余,让 LDD 特征的可分性更好.

3.4 不同降维方法的对比

为了验证 LDA 提取语音 LDD 特征的有效性,第二个实验将对比分析不同的降维方法. 在深度学习中,除了 LDA,常用的降维方法还有主成分分析(Principal Component Analysis,PCA)、独立成分分析(Independent Component Analysis,ICA)、因子分析(Factor Analysis,FC)和局部线性嵌入(Locally Linear Embedding,LLE)等[21],不同的降维方法适用不同的应用场景.

表 2 给出了不同降维方法与 SVM 结合下的平均识别准确率. 可以看出,LDA 方法获得了最好的识别准确率. 这是因为,SVM 对数据的冗余比较敏感,当训练数据越是线性可分的时候,SVM 的分类效果越明显,相对于其他的降维方法,LDA 可以充分利用先验知识,计算速度快,特别是当数据满足高斯分布的时候效果非常显著,而语音信号中含有的设备噪声主要就是高斯噪声.

表 2 不同降维方法的平均识别准确率

降维方法	LDA	PCA	ICA	FC	LLE
Accuracy(%)	**99.96**	97.2	95.54	98.68	74.05

3.5 不同识别方法的对比

为了进一步验证本文所提方法的有效性,本节实验将基于 LDD 特征的手机来源识别方法(后称 LDD + SVM)与 HDD + TCN(即直接使用 TCN 进行分类)、BED + SVM[3] 和 CQT + CNN[4] 进行对比实验分析.

表 3 不同识别方法的平均识别准确率

识别方法	BED + SVM	CQT + CNN	HDD + TCN	LDD + SVM
Accuracy(%)	97.34	97.36	96.87	**99.96**

表 3 给出了不同手机来源识别方法的平均识别准确率. 可以看出,在自然环境背景噪声手机语音库上,BED + SVM 和 CQT + CNN 的识别性能相当,均略好于 HDD + TCN,而本文的 LDD + SVM 方法识别准确率最高,离完全识别只差了 0.04%. 上述实验结果表明,HDD + TCN 方法单纯依靠 TCN 暴力提取 HDD,不能去除音频信号中的冗余,其识别性能远不如 LDD + SVM 方法,说明了通过 LDA 提取 LDD 的必要性.

对于同一品牌不同型号的手机,往往采用相同的语音采集芯片,更加难以区分. 表 4 给出了 BED + SVM、CQT + CNN 和 LDD + SVM 三种方法在相同品牌不同型号上 Precision、Recall 和 F1-score 的结果. 可以看出,在每个品牌的不同型号(对应不同 ID)上,LDD + SVM 的查准率和 F1 分数均要好于 BED + SVM 和 CQT + CNN. 此外,除了在 ID36 和 ID38 上,LDD + SVM 的查全率要稍微低于 BED + SVM 和 CQT + CNN 一点点外,在每个品牌的不同型号上,LDD + SVM 的查全率均显著优于 BED + SVM 和 CQT + CNN. 而且,可以很清楚的看到,LDD + SVM 几乎在每个型号上的每个指标都达到了 100%. 上述结果表明,LDD + SVM 对相同品牌不同型号手机的适应性要明显优于 BED + SVM 和 CQT + CNN.

表 4 三种识别方法在相同品牌不同型号上各指标的结果

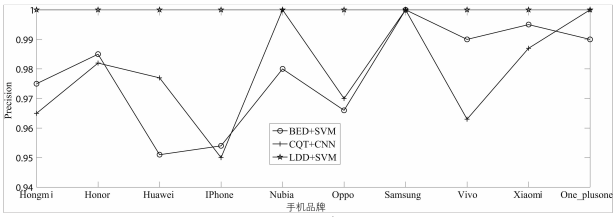
品牌	ID	Precision			Recall			F1-score		
		BED + SVM	CQT + CNN	LDD + SVM	BED + SVM	CQT + CNN	LDD + SVM	BED + SVM	CQT + CNN	LDD + SVM
Hongmi	1	1	1	1	1	1	1	1	1	1
	2	0.95	0.93	1	0.93	0.92	1	0.94	0.92	1

续表

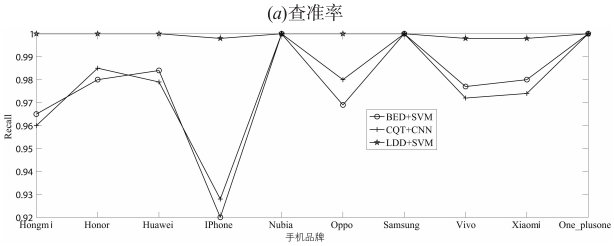
品牌	ID	Precision			Recall			F1-score		
		BED + SVM	CQT + CNN	LDD + SVM	BED + SVM	CQT + CNN	LDD + SVM	BED + SVM	CQT + CNN	LDD + SVM
Honor	3	0.96	0.97	1	0.94	0.96	1	0.95	0.97	1
	4	1	0.99	1	1	0.99	1	1	0.99	1
	5	0.99	0.98	1	0.96	0.97	1	0.98	0.97	1
	6	1	0.99	1	0.98	0.99	1	0.99	0.99	1
	7	1	1	1	1	1	1	1	1	1
	8	0.96	0.96	1	1	1	1	0.98	0.98	1
Huawei	9	0.97	0.93	1	1	0.99	1	0.98	0.96	1
	10	0.99	1	1	0.95	1	1	0.97	1	1
	11	0.99	0.95	1	0.94	0.91	1	0.97	0.93	1
	12	0.99	1	1	1	1	1	0.99	1	1
	13	0.98	0.98	1	1	0.98	1	0.99	0.98	1
	14	1	1	1	1	1	1	1	1	1
	15	0.74	0.98	1	1	0.97	1	0.85	0.98	1
IPhone	16	0.89	0.85	1	1	0.98	1	0.94	0.91	1
	17	1	0.98	1	0.66	0.77	0.99	0.79	0.86	1
	18	0.92	0.93	1	0.96	0.89	1	0.94	0.91	1
	19	0.96	0.99	1	0.98	1	1	0.97	1	1
	20	1	1	1	1	1	1	1	1	1
Nubia	21	0.98	1	1	1	1	1	0.99	1	1
Oppo	22	0.99	1	1	0.94	1	1	0.97	1	1
	23	1	1	1	0.95	0.98	1	0.98	0.99	1
	24	0.97	0.99	1	1	1	1	0.99	1	1
	25	0.99	0.99	1	0.94	0.97	1	0.96	0.98	1
	26	1	1	1	1	0.98	1	1	0.99	1
	27	0.92	0.97	1	0.98	0.99	1	0.95	0.98	1
	28	0.89	0.84	1	0.97	0.94	1	0.93	0.89	1
Samsung	29	1	1	1	1	1	1	1	1	1
Vivo	30	1	0.94	1	1	1	1	1	0.97	1
	31	0.97	0.97	1	0.91	0.96	1	0.94	0.97	1
	32	1	0.96	1	0.97	0.91	1	0.99	0.93	1
	33	0.98	0.92	1	0.99	0.98	1	0.98	0.95	1
	34	1	1	1	0.99	0.99	0.99	0.99	0.99	1
	35	0.99	0.99	1	1	0.99	1	0.99	0.99	1
Xiaomi	36	1	0.94	1	1	1	0.99	1	0.97	1
	37	0.99	1	1	0.98	1	1	0.98	1	1
	38	1	0.99	1	1	1	0.99	1	0.99	1
	39	1	0.98	1	0.99	0.99	1	1	0.99	1
	40	0.98	0.99	1	0.92	0.83	1	0.95	0.9	1
	41	1	1	1	1	0.98	1	1	0.99	1
	42	0.99	0.99	1	0.98	0.96	1	0.99	0.98	1
	43	0.99	0.97	1	0.98	0.96	1	0.99	0.97	1
	44	1	1	1	1	1	1	1	1	1
	45	1	1	1	0.94	1	1	0.97	1	1
	46	0.99	1	1	0.99	0.99	1	0.99	1	1
One_plusone	47	0.99	1	1	1	1	1	0.99	1	1
平均值		0.9766	0.9753	1	0.9728	0.9728	0.9991	0.9734	0.9738	1

此外,图 6 给出了 BED + SVM、CQT + CNN 和 LDD + SVM 三种方法在不同品牌上 Precision、Recall 和 F1-score 三个指标的结果. 从图中可以看出,在 7 个不同的手机品牌上,LDD + SVM 的查准率、查全率和 F1 分数均要明显好于 BED + SVM 和 CQT + CNN. 其中,仅在 Nubia、Samsung 和 One_plusone 三种手机上与 BED + SVM 和 CQT + CNN 旗鼓相当,这是因为数据库中这三种手机仅有一个型号. 上述结果验证了 LDD +

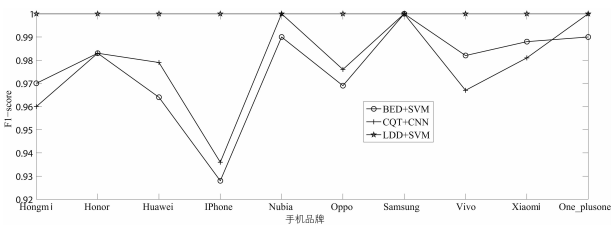
SVM 对不同手机品牌的适应性要优于 BED + SVM 和 CQT + CNN.



(*a*)查准率



(*b*)查全率



(*c*)F1分数

图6 三种识别方法在不同品牌上各指标的结果

3.6 样本规模的影响对比

基于机器学习的手机语音来源识别方法的性能往往会受到数据集规模的影响. 数据集规模越大,模型训练就越充分,识别的效果就会越好. 反之,数据集规模越小,模型训练越不充分,要想达到好的识别效果,对识别方法的性能要求就越高. 为了衡量不同识别方法在不同数据集规模上的适应性,图 7 给出了 BED + SVM、CQT + CNN 和 LDD + SVM 三种方法对每个手机 ID 上不同数据集规模时的平均识别准确率. 可以看出,随着每个手机 ID 上的数据集规模的减小,BED + SVM 和 CQT + CNN 的识别准确率均呈下降趋势,而本文 LDD + SVM 的识别准确率并没有发生明显变化,始终接近 100%,甚至在每个手机 ID 只有 200 条语音时,BED + SVM 和 CQT + CNN 的准确率均不到 92%,而 LDD + SVM 的准确率非常接近 100%. 上述实验结果表明,本文的 LDD + SVM 方法即使在小规模数据集下仍具有较好的识别性能,对不同规模数据集的适应性要更强.

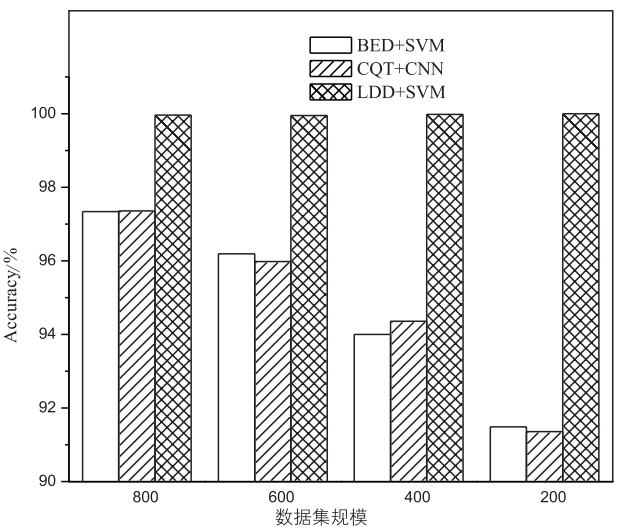


图7 不同识别方法在不同数据集规模上的平均识别准确率

3.7 样本长度的影响对比

手机设备信息是固有特征,存在于语音信号的整个过程. 音频越短,检测效率越高,可以更快的给出识别结果,但面临的挑战也越大. 尤其在司法取证领域,往往存在很多短音频(时长小于 3s),例如手机即时通讯工具中的语音聊天记录. 因此,为了衡量不同识别方法对不同样本长度的适应性,图 8 给出了 BED + SVM、CQT + CNN 和 LDD + SVM 三种方法对每个手机 ID 上不同样本长度时的平均识别准确率. 可以看出,样本长度的确对识别性能仅有很细微的影响. 这是因为,无论音频片段长还是短,都包含固有设备特征信息. 因此,如果测试音频的时长大于 1s,则完全可以对测试音频进行切分,只取前 1s 进行测试,即可判定其手机来源,从而大大降低分析的数据量.

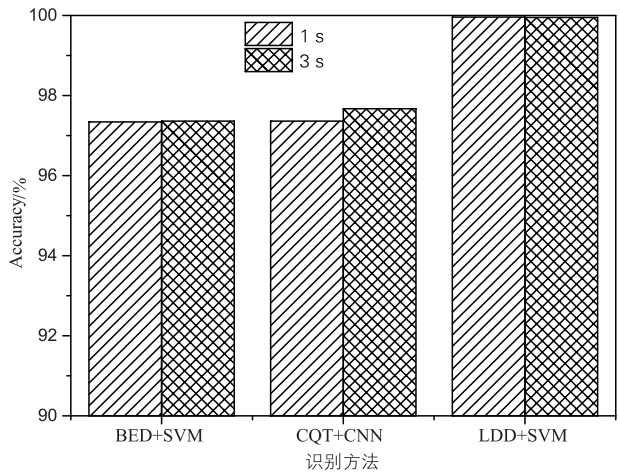


图8 不同识别方法在不同样本长度上的平均识别准确率

3.8 采样率的影响对比

本节实验将衡量不同采样率下本文 LDD + SVM 方法的有效性. 图 9 给出了 LDD + SVM 方法分别在采样率 8kHz、11.025kHz 和 22.05kHz 下的识别性能. 可以看出,虽然在 22.05kHz 下的整体识别性能略好一些,但与在 8kHz 和 11.025kHz 上的各个指标值差距非常小,均不低于 99.7%. 上述实验结果表明,LDD + SVM 方法对采样率并不太敏感. 这是因为 LDD + SVM 是基于 Fbank 特征(没有进行去相关和压缩处理),从各个频段

中深度挖掘最能表征设备噪声的信息.

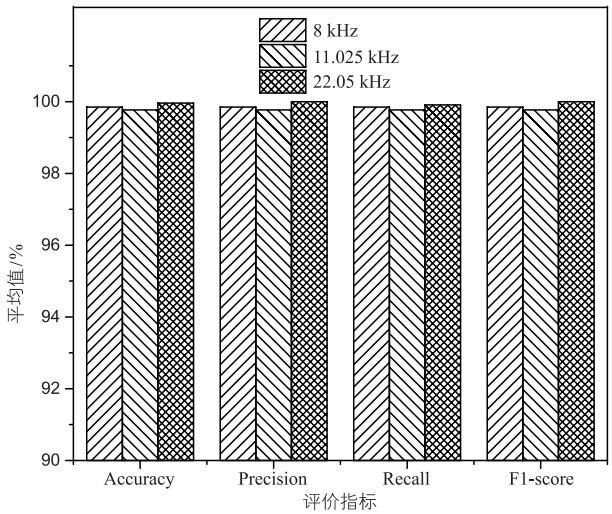


图9　LDD+SVM方法在不同采样率上的识别性能

4　结束语

基于手机语音的来源设备识别是多媒体取证领域中的一个热点问题,本文针对司法领域对自然环境背景噪声下的手机语音证据的鉴真需求,首先提取手机语音中的 Fbank 特征以保留完整的设备噪声信息,并输入到 TCN 中进行训练,进一步提取能够表征语音设备的深度特征,然后利用 LDA 进行降维,去除高维深度特征中的冗余,并结合 SVM 提出了一种基于低维深度特征的手机来源识别方法 LDD + SVM. 通过在 47 种型号手机设备录制的 37600 条自然环境噪声语音样本库上的实验表明,本文提出的 LDD + SVM 方法在准确率、查准率、查全率和 F1 分数四个主流指标上的整体表现要明显优于已有识别方法,不仅出错率降低,而且可以很好的适应不同品牌、相同品牌不同型号、不同样本长度、不同数据集规模和不同采样率,为自然环境背景噪声下的手机来源识别提供了一个有益的尝试. 但由于实验条件的限制,本文收录的手机型号覆盖范围还不够广泛,在未来仍需进一步扩充语音库,而且还需要尝试其它的手机语音格式.

参考文献

[1] 贺前华,王志锋,RUDNICKY A I,等. 基于改进 PNCC 特征和两步区分性训练的录音设备识别方法[J]. 电子学报,2014,42(1):191 – 198.
HE Qian-hua, WANG Zhi-feng, RUDNICKY A I, et al. A recording device identification algorithm based on improved PNCC feature and two-step discriminative training[J]. Acta Electronica Sinica, 2014, 42(1): 191 – 198. (in Chinese)

[2] ZOU L, YANG J, HUANG T. Automatic cell phone recognition from speech recordings[A]. Proceedings of the 5th IEEE China Summit and International Conference on Signal and Information Processing[C]. Xi'an, China: IEEE, 2014. 621 – 625.

[3] LUO D, KORUS P, HUANG J. Band energy difference for source attribution in audio forensics[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(9): 2179 – 2189.

[4] QIN T, WANG R, YAN D, et al. Source cell-phone identification in the presence of additive noise from CQT domain [J]. Information, 2018, 9(8): Article No. 205.

[5] JIANG Y, LEUNG F H F. Mobile phone identification from speech recordings using weighted support vector machine[A]. Proceedings of the 42nd Annual Conference of the IEEE Industrial Electronics Society[C]. Florence, Italy: IEEE, 2016. 963 – 968.

[6] LI Y, ZHANG X, LI X, et al. Mobile phone clustering from speech recordings using deep representation and spectral clustering[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(4): 965 – 977.

[7] VERMA V, KHATURIA P, KHANNA N. Cell-phone identification from recompressed audio recordings[A]. Proceedings of the 24th National Conference on Communications[C]. Hyderabad, India: IEEE, 2018. 1 – 6.

[8] QI S, HUANG Z, LI Y, et al. Audio recording device identification based on deep learning[A]. Proceedings of the IEEE International Conference on Signal and Image Processing[C]. Beijing, China: IEEE, 2016. 426 – 431.

[9] JIN C, WANG R, YAN D, et al. Source cell-phone identification using spectral features of device self-noise[A]. Proceedings of the 15th International Workshop on Digital Watermarking[C]. Beijing, China: Springer, 2016. 29 – 45.

[10] 裴安山,王让定,严迪群. 基于设备本底噪声频谱特征的手机来源识别[J]. 电信科学,2017,33(1):85 – 94.
PEI An-shan, WANG Rang-ding, YAN Di-qun. Cell-phone origin identification based on spectral features of device self-noise[J]. Telecommunications Science, 2017, 33(1): 85 – 94. (in Chinese)

[11] 裴安山,王让定,严迪群. 基于语音静音段特征的手机来源识别方法[J]. 电信科学,2017,33(7):103 – 111.
PEI An-shan, WANG Rang-ding, YAN Di-qun. Source cell-phone identification from recorded speech using non-speech segments[J]. Telecommunications Science, 2017, 33(7): 103 – 111. (in Chinese)

[12] BALDINI G, AMERINI I, GENTILE C. Microphone identification using convolutional neural networks[J]. IEEE Sensors Letters, 2019, 3(7): Article No. 6001504.

[13] BALDINI G, AMERINI I. Smartphones identification through the built-in microphones with convolutional neu-

ral network[J]. IEEE Access,2019,7:158685 – 158696.
[14] BAI S,KOLTER J Z,KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[OL]. https://arxiv. org/abs/1803. 01271,2018-04-19.
[15] ABBASIAN H,NASERSHARIF B,AKBARI A,et al. Optimized linear discriminant analysis for extracting robust speech features[A]. Proceedings of the 3rd International Symposium on Communications,Control and Signal Processing[C]. St Julians,Malta:IEEE,2008. 819 – 824.
[16] CHANG C C,LIN C J. LIBSVM:A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(3):Article No. 27.
[17] MCFEE B,RAFFEL C,LIANG D,et al. librosa:audio and music signal analysis in Python[A]. Proceedings of the 14th Python in Science Conference[C]. Austin,Texas,USA:SciPy Organizers,2015. 18 – 25.
[18] IOFFE S,SZEGEDY C. Batch normalization:accelerating deep network training by reducing internal covariate shift[A]. Proceedings of the 32nd International Conference on Machine Learning[C]. Lille,France:JMLR. org,2015. 448 – 456.
[19] GRANDINI M,BAGLI E,VISANI G. Metrics for multiclass classification:an overview[OL]. https://arxiv. org/abs/2008. 05756,2020-08-13.
[20] VAN DER MAATEN L,HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research,2008,9:2579 – 2605.
[21] KASUN L L C,YANG Y,HUANG G,et al. Dimension reduction with extreme learning machine[J]. IEEE Transactions on Image Processing,2016,25(8):3906 – 3918.

作者简介

苏兆品 女,1983 年 8 月出生,山东菏泽人. 副教授,硕士生导师,CCF 会员. 2004 年和 2008 年在合肥工业大学分别获得学士和博士学位. 主要从事音频信息隐藏、深度学习和进化计算等方面的研究工作.
E-mail:szp@ hfut. edu. cn

吴张倩 女,1995 年 1 月出生,安徽宿州人. 硕士研究生. 2017 年在安徽中医药大学获得学士学位. 从事音频取证和深度学习方面的研究.
E-mail:2465231972@ qq. com

岳 峰 男,1981 年 2 月出生,安徽合肥人. 副研究员,硕士生导师. 2004 年、2009 年和 2015 年在合肥工业大学分别获得学士、硕士和博士学位. 主要从事软件工程、音频信息隐藏和进化计算等方面的研究工作.
E-mail:yuefeng@ hfut. edu. cn

武钦芳 女,1996 年 8 月出生,安徽亳州人. 硕士研究生. 2018 年在安徽中医药大学获得学士学位. 从事音频取证和软件工程方面的研究.
E-mail:1841807170 @ qq. com

张国富(通信作者) 男,1979 年 3 月出生,安徽合肥人. 教授,硕士生导师,CCF、CAA 会员. 2002 年和 2008 年在合肥工业大学分别获得学士和博士学位. 2011 年至 2013 年在合肥工业大学信息与通信工程博士后流动站从事博士后研究,2015 年至 2016 年为英国伯明翰大学计算机学院访问学者,现为工业安全与应急技术安徽省重点实验室副主任,主要从事基于搜索的软件工程、音频安全和进化计算等方面的研究工作.
E-mail:zgf@ hfut. edu. cn