基于改进 PNCC 特征和两步区分性训练的录音设备识别方法

贺前华[1],王志锋[1,2],Alexander I Rudnicky[2],朱铮宇[1],李新超[1]
(1.华南理工大学电子与信息学院,广东广州 510640;2.卡内基梅隆大学计算机学院,美国匹兹堡 15213)

摘 要: 录音设备来源识别是通过分析已获取的数字语音信号从而确定其录制设备的一种技术,属于数字音频盲取证.本文提出了一种基于改进 PNCC 特征和两步区分性训练的录音设备识别方法,由于音频中的静音包含了完整的设备信息,且不受说话人和文本等因素的影响,因此从静音段提取改进的 PNCC 特征,利用了 PNCC 的长时帧分析去除背景噪声对设备信息的影响.在模型方面,以 GMM-UBM 为基准模型,并通过两步区分性训练调整集内设备模型和通用背景模型,提升模型区分能力.该方法对于 30 种设备闭集识别的平均正确识别率为 90.23%;对于 15 个集内和 15 个集外设备的测试,等错误率为 15.17%,集内平均正确识别率为 96.65%,验证了本文算法的有效性.

关键词: 数字音频取证; 录音设备识别; GMM-UBM; 区分性训练; PNCC
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2014)01-0191-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.01.031

A Recording Device Identification Algorithm Based on Improved PNCC Feature and Two-Step Discriminative Training

HE Qian-hua[1], WANG Zhi-feng[1,2], Alexander I Rudnicky[2], ZHU Zheng-yu[1], LI Xin-chao[1]
(1. *School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong* 510640, *China*;
2. *School of Computer Science, Carnegie Mellon University, Pittsburgh* 15213, *USA*)

Abstract: Recording device identification is a kind of blind digital audio forensic technique, which extracts digital evidence of device mechanism involved in the generation of the speech recording by analyzing the acoustic signal. This paper proposes a recording device identification algorithm which is based on improved PNCC feature and two-step discriminative training. Due to the fact that silence periods contain the device information and is not affected by speaker and texture factors, this paper extracts improved PNCC from silence periods, which uses long term analysis to remove the effect of background noise. GMM-UBM is set as the baseline system, which is improved by two steps discriminative training. The experimental result indicates that the average accuracy of recording device identification on 30 devices is 90.23%; for 15 inset and 15 outset devices testing, the EER is 15.17% and ACC is 96.65%, which proves the effectiveness of the proposed algorithm.

Key words: digital audio forensics; recording device identification; GMM-UBM; discriminative training; PNCC

1 引言

随着篡改和伪造数字多媒体的广泛传播,引起了诸多的安全问题和社会危害,为此数字多媒体取证技术得到广泛的关注和应用[1,2].设备来源识别技术是数字多媒体取证技术中非常重要的组成部分,是对多媒体的来源、真实性和可靠性进行验证的一门技术.现有的设备来源识别技术主要集中在数字图像的设备识别领域,例如相机[3]、打印机和扫描仪[4]等的来源识别,而录音设备的识别研究却非常少.录音设备识别的目标是直接由已获取的语音信号找到其中隐含的录音设备信息,从而确定其录音设备,属于音频盲取证技术.大多数情况下盲录音设备取证更符合实际情况,因此具有更好的实用价值.录音设备识别有多方面的应用:在司法取证方面,可验证数字音频证据来源的可靠性和真实性;在知识产权保护方面,可协助打击盗版;也可验证数字音频的真实性,如音频中包含多个录音设备信息或者所包含录音设备信息出现不连续性,说明这段音频是经过人为编辑的;针对说话人和语音识别中的信道不匹配,可用来检测训练和识别时的设备信道信息,建立不同录音设备之间的映射函数提高识别率.录音设备识别包含几个不同层面的问题:录音设备类型的识别、录音设备品牌的识

收稿日期:2012-11-15;修回日期:2013-02-23;责任编辑:孙瑶
基金项目:国家自然科学基金(No.60972132,No.61101160);广东省自然科学基金(No.9351064101000003,No.10451064101004651)

别、录音设备型号以及录音设备个体的识别,其中录音设备类型如手机、录音笔、PDA、mp3 和 mp4 等,录音设备个体指某一台特定录音设备.

录音设备识别研究在国内外还处于起步阶段,相关报道很少.C Kraetzer 等[5]对四种麦克风的录音进行了闭集分类识别,其特征为短时特征和 Mel 倒谱系数构成的组合特征,并用朴素贝叶斯分类器在语音帧层面上进行封闭集分类识别,四种麦克风的平均正确识别率为 60% ~ 75%.该方法直接用语音信号提取设备特征,易受说话人、文本、说话人情感等因素的影响,所提特征反映的是混合音的特征,而不仅仅是录音设备特征.因此,为避免其他因素的干扰,本文在静音的基础上提取改进 PNCC 特征作为设备特征,并利用两步区分性训练获得具有区分性的设备模型.

2 录音设备识别模型

由于不同录音设备中传感器和信号采集电路的差异,产生不同的设备噪声,这些设备噪声混合在语音信号中形成"机器指纹"[6],提取了这种"机器指纹"就可以进行录音设备识别,而这些"机器指纹"可以采用现代的统计学方法和模式识别技术进行提取和识别.

2.1 基于两步判决的录音设备识别模型

首先,假定有 N 个录音设备在系统中进行了注册,对于一段来自特定录音设备 D 的语音样本 S,第一步则要判决该样本和集内哪一个录音设备模型 $\Lambda_n(1\leqslant n\leqslant N)$最接近,记 Λ_{out}为集外反设备模型.记 $X=\{x_1,x_2,\cdots,x_T\}$为语音样本 S 的特征.那么第一步判决模型为:

$$\Lambda_* = \underset{1\leqslant n\leqslant N}{\operatorname{argmax}}\, p(X/\Lambda_n) \tag{1}$$

其中 Λ_*为集内和观测样本 X 最接近的模型.

第二步则要判决观测样本 X 是否真正来自集内设备模型 Λ_*,即判断该样本是否来自集外,可将其归结为一个基本的假设检验问题:

H_0:观测样本 X 来自于集内录音设备模型 Λ_*;

H_1:观测样本 X 来自于集外录音设备模型 Λ_{out}.

以上假设检验可通过一个似然度检测等价给出:

$$\frac{p(X/\Lambda_*)}{p(X/\Lambda_{\text{out}})}\begin{cases}\geqslant\theta, \text{接受 } H_0\\ <\theta, \text{接受 } H_1\end{cases} \tag{2}$$

其中 θ 为预设判决阈值,$p(X/\Lambda_*)$表示 X 来自集内设备模型 Λ_*的概率,$p(X/\Lambda_{\text{out}})$表示 X 来自集外录音设备模型 Λ_{out}的概率.

2.2 基于 GMM-UBM 的录音设备基准模型

本文采用 GMM-UBM[7]建立录音设备的基准模型,集外模型 Λ_{out}则采用通用背景模型 UBM 建模.本文设备通用背景模型 DEV-UBM(DEV 代表设备)建立方法如图 1,分别用两类常用麦克风(驻极式和电容式)数据训练子设备模型合并成一个 DEV-UBM,使得两类数据训练出来的子模型保持平衡不偏向于某一类设备,同时也可以降低计算复杂度减少训练时间.

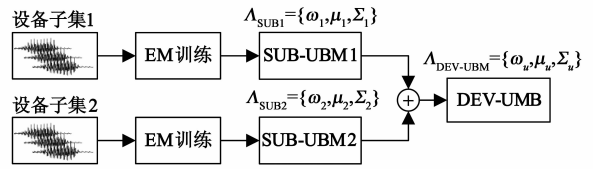


图1 设备通用背景模型的训练

在已获得的 DEV-UBM 上采用 MAP[8]算法进行自适应获得特定录音设备的基准模型 DEV – GMM.

2.3 两步区分性训练模型

实际情况下从录音样本中获取的静音量有限,难以获得准确描述设备空间分布的模型,使得 DEV-GMM 跟 DEV-UBM 非常靠近,各个设备模型之间的区分性很小,在短训练和测试样本(2 ~ 20s)情况下的识别率不高.因此本文将 GMM-UBM 作为设备基准模型,在其基础上根据 2.1 节两步判决模型,采用两步区分性训练使得集内各设备模型之间的分类误差最小,集内所有设备模型和通用背景模型的确认误差最小.

2.3.1 第一步区分性训练:最小分类误差训练[9],调整集内设备模型

此步训练的目的是调整集内的设备模型,使得各集内设备模型之间的分类误差最小,训练时只采用集内设备数据.首先,定义错误分类测度函数

$$d_n(X/\Lambda) = -G(X/\Lambda_n) + \log\Big[\frac{1}{N-1}\sum_{j\neq n}\exp\{\eta\cdot G(X/\Lambda_j)\}\Big]^{\frac{1}{\eta}},\quad 1\leqslant j\leqslant N \tag{3}$$

其中,Λ_n 是集内第 n 个录音设备的模型,$G(X/\Lambda)$为观测样本在 GMM 模型上的对数似然度得分,

$$G(X/\Lambda)=\frac{1}{T}\sum_{t=1}^{T}\log\sum_{i=1}^{M}\omega_i p_i(x_t) \tag{4}$$

$G(X/\Lambda_n)$代表正确分类的对数似然度,$\log\Big[\frac{1}{N-1}\sum_{j\neq n}\exp\{\eta\cdot L(X/\Lambda_j)\}\Big]^{\frac{1}{\eta}}$代表其它可能错判的对数似然度,当 $\eta\to\infty$,式(4)可简化为:

$$d_n(X/\Lambda) = -G(X/\Lambda_n) + \max_{j\neq n} G(X/\Lambda_j),\quad 1\leqslant j\leqslant N \tag{5}$$

为了方便求解,通过 S 型函数将 $d_n(X/\Lambda)$进行归一化平滑转换,获得损失函数,

$$l_n(X;\Lambda)=\frac{1}{1+\mathrm{e}^{-\alpha d_n(X/\Lambda)}} \tag{6}$$

其中 $\alpha(\alpha>0)$是 S 型函数的斜率系数,当 $d_n(X/\Lambda)$为一个非常大的负值,损失函数 $l_n(X;\Lambda_n)$趋近于 0,表明

分类正确.另一方面,当 $d_n(X/\Lambda)$为一个正值,损失函数介于 0 和 1 之间,则表示产生错误分类的似然度,即惩罚因子.全局的损失函数可以定义为

$$L(X;\Lambda)=\sum_{n=1}^{N} l_n(X;\Lambda)\delta_n(X) \tag{7}$$

当观测数据 X 由模型 Λ_n 产生时,$\delta_n(X)=1$,否则 $\delta_n(X)=0$.通过 GDP 方法[10],可将上述优化问题转换为利用损失函数的梯度对设备模型进行自适应更新,即

$$\Lambda^{k+1}=\Lambda^k-\lambda\nabla L(X;\Lambda)\mid_{\Lambda=\Lambda^k} \tag{8}$$

其中,Λ 模型初始值为基准设备模型,λ 为学习系数,$\nabla L(X;\Lambda)$为损失函数的梯度,k 为迭代次数.

2.3.2 第二步区分性训练:最小确认误差训练,调整通用背景模型

第二步区分性训练的目的是调整 DEV-UBM,使得所有集内设备模型和 DEV-UBM 的确认误差最小,训练时采用集内和集外设备数据(X_{in}为集内数据,X_{out}为集外数据).根据式(2),确认录音设备是否为集外设备时会产生两类错误(错误接受 FA 和错误拒绝 FR),定义错误确认测度函数:

$$d_0(X/\Lambda)=-G(X/\Lambda_n)+G(X/\Lambda_{\text{UBM}})-\theta \tag{9}$$

$$d_1(X/\Lambda)=G(X/\Lambda_n)-G(X/\Lambda_{\text{UBM}})+\theta=-d_0(X/\Lambda) \tag{10}$$

与 MCE 训练类似,通过 S 型函数将错误确认函数 $d_k(X/\Lambda)(k=0,1)$进行归一化平滑转换获得损失函数:

$$l_0(X;\Lambda_{\text{UBM}})=\frac{1}{1+e^{-\beta d_0(X/\Lambda)}} \tag{11}$$

$$l_1(X;\Lambda_{\text{UBM}})=\frac{1}{1+e^{\beta d_0(X/\Lambda)}} \tag{12}$$

全局损失函数可以定义为

$$L(X;\Lambda)=\sum_{X\in X_{\text{in}}} l_0(X;\Lambda)+\sum_{X\in X_{\text{out}}} l_1(X;\Lambda) \tag{13}$$

同样采用 GDP 方法[10],用损失函数的梯度对 DEV-UBM 进行自适应,同时获得最优判决阈值

$$\Lambda_{\text{DEV-UBM}}^{k+1}=\Lambda_{\text{DEV-UBM}}^{k}-\gamma\nabla L(X;\Lambda_{\text{DEV-UBM}})\mid_{\Lambda_{\text{DEV-UBM}}=\Lambda_{\text{DEV-UBM}}^{k}} \tag{14}$$

$$\theta_n^k=\theta_n^{k+1}-\rho\nabla L(X;\Lambda)\mid_{\theta=\theta_n^k} \tag{15}$$

2.4 集内和集外录音设备的识别与判决

第一步判决:采用对数似然度来计算输入语音数据的得分,输入的未知设备样本在每一个设备模型上都会获得一个得分,找出得分最高的模型

$$\Lambda_*=\arg\max_{1\leqslant n\leqslant N}\frac{1}{T}\sum_{t=1}^{T}\log p(x_t/\Lambda_n) \tag{16}$$

第二步判决:对式(2)进行等价转换,获得集内、集外判决公式为:

$$\Omega(X)=\frac{1}{T}\Big(\sum_{t=1}^{T}\log p(x_t/\Lambda_*)-\sum_{t=1}^{T}\log p(x_t/\Lambda_{\text{DEV-UBM}})\Big)\begin{cases}\geqslant\theta,\Lambda_*\\ <\theta,\Lambda_{\text{out}}\end{cases} \tag{17}$$

3 录音设备识别算法

语音中包含丰富的信息,例如说话人信息、文本信息、情感信息等,设备信息很容易被其它信息所掩盖,要单独从语音中提取出设备信道信息是非常困难的[11].本文利用静音段进行设备识别,因为静音包含了完整设备信息,且不受说话人、文本等信息的影响.并在静音的基础上通过 PNCC 特征[12]提取设备特征.

3.1 静音检测

如图 2 所示,首先采用双门限法[13]进行静音检测.为了保证有足够长度的静音数据进行训练和识别(实际的语音库中静音的长短差别很大),将相邻的不足 5s 的静音数据拼接起来构成时长不少于 5s 的静音样本.

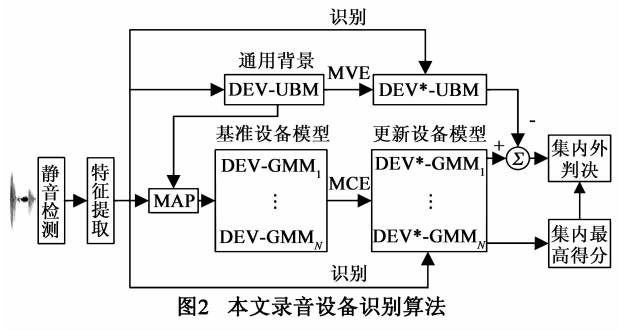


图2 本文录音设备识别算法

3.2 原始 PNCC 特征提取过程

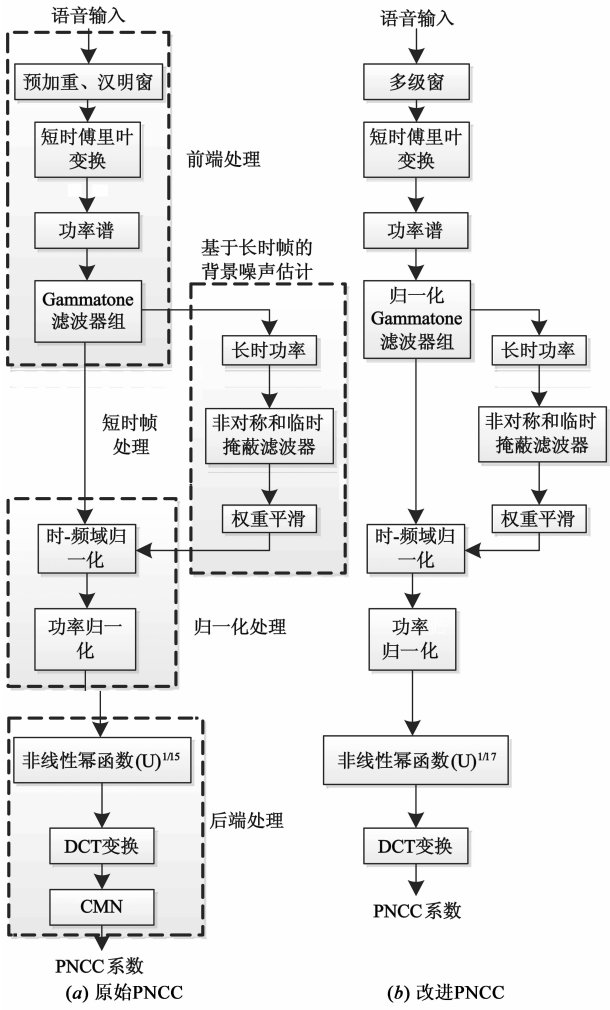
如图 3(a),PNCC 包含前端处理、基于长时帧的背景噪声估计、归一化处理、后端处理四个步骤.

3.3 改进的 PNCC 特征

虽然 PNCC 利用长时帧分析有效地去除了背景噪声的影响,但其中的一些处理方法同时去除设备和信道噪声.因此本文对 PNCC 进行了四个方面的改进,在去除背景噪声的同时保留设备信息.

3.3.1 预处理时不采用预加重

预加重是在语音信号数字化后进行的,处于数字语音信号处理的最前端.预加重实际上是一个高通滤波器,它的目的是抑制语音的低频部分,提升高频部分,使得语音信号的频谱变得平坦,以便于频谱分析和声道参数估计,对于语音识别、说话人识别等比较有效.录音设备特征提取则主要关注设备信道信息,作者前期工作显示录音设备信道信息相对于其它语音部分变化缓慢,大部分处于信号的低频部分[14],而预加重会抑制低频部分的信道信息,因此本文改进的 PNCC 没有采用预加重操作.



(*a*) 原始PNCC (*b*) 改进PNCC

图3 原始PNCC和改进PNCC特征

3.3.2 采用多级窗代替汉明窗

对语音信号进行短时分析时,需对语音信号进行加窗操作.通常采用单个窗函数,而单个窗会引入不平滑和频谱的失真,使得通过窗函数后的信号频谱估计存在较大的方差[15],而多级窗则可以有效地保留大部分频谱信息

$$S(f)=\sum_{j=1}^{K}\varphi(j)\left|\sum_{t=1}^{T}w_j(t)x(t)\mathrm{e}^{-\mathrm{i}2\pi tf/T}\right|^2 \quad (18)$$

其中 $w_j(t)$为多级窗函数,$\varphi(j)$为子窗权重系数,本文采用正弦多级窗函数

$$w_j(t)=\sqrt{2/(T+1)}\sin\frac{T\pi\mathrm{j}}{T+1} \quad (19)$$

$$\varphi(j)=\frac{\cos(2\pi\mathrm{j}M/2T)+1}{\sum_{j=1}^{K}[\cos(2\pi\mathrm{j}M/2T)+1]} \quad (20)$$

由图 4 可知,多级窗处理后的频谱比汉明窗更平滑,频谱估计的方差更小,信号中设备信息的频谱损失也会减小,有利于设备特征的提取.

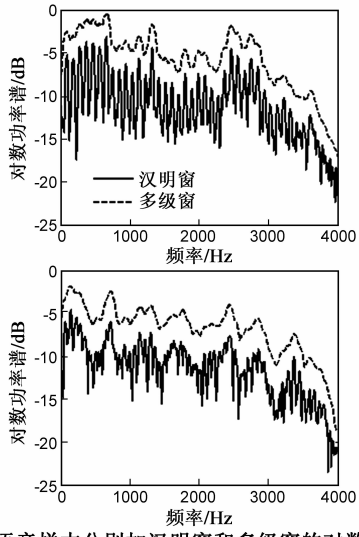
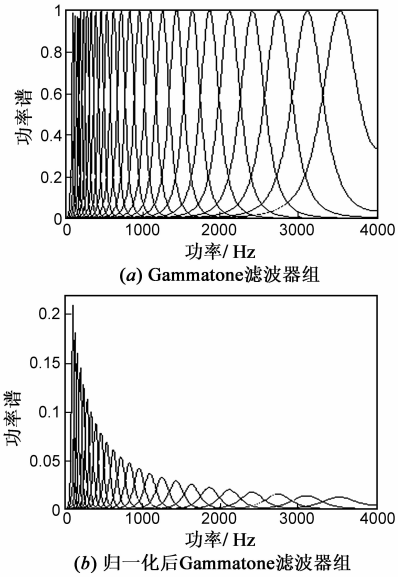


图4 对两个语音样本分别加汉明窗和多级窗的对数功率谱(K=8)

3.3.3 归一化 Gammatone 滤波器组

Gammatone 滤波器组非常接近人耳的听觉特性,对于加性背景噪声以及高斯白噪声有一定的抑制作用[12].但是原始的 PNCC 没有对 Gammatone 滤波器组进行归一化处理,使得高频部分的权重比较高,使得低频部分权重较低,而设备信息主要处于信号的低频部分[14].本文对滤波器组每个通道进行归一化,提升了低频部分的权重.本文归一化后 Gammatone 滤波器组如图 5 所示.



(*a*) Gammatone滤波器组

(*b*) 归一化后Gammatone滤波器组

图5 原始Gammatone滤波器组和归一化后滤波器组

3.3.4 不采用倒谱均值归一化 CMN

CMN 算法在每一个短时帧倒谱域上减去信道均值响应,用来消除信道卷积噪声在倒谱域造成的影响.提取设备特征的目标是要保留信号中设备信息,而原始

PNCC 中 CMN 操作会去除部分信道信息,对于最后录音设备信息的提取有较大影响,所以本文没有采用 CMN 操作.

3.4 模型训练以及参数设定

本文改进的 PNCC 是在 Kim 的 PNCC 工具箱[12]以及 Slaney 的听觉工具箱[16]上实现的,多级窗是在 Kinnunen 的工具箱[15]上实现的,这三个工具箱都是基于 matlab 平台.另外,DEV-UBM 采用 HTK 工具箱[17]进行训练,然后在 matlab 上实现 MAP 算法、MCE 以及 MVE 训练,从而获得特定设备的模型.

采用 NIST 2006 数据库的 ICSI 子库[18]训练 DEV-UBM,包含有 8 种不同的麦克风信道和 61 个说话人(28 男 33 女),共有 2223 段语音样本,每个样本时常约为 5min.包含的 8 种麦克风型号如表 1.

表 1　NIST 2006 ICSI 子库的 8 种麦克风

麦克风	型号	麦克风类别
M1	AT3035 (Audio Technica Studio Mic)	电容式
M2	MX418S (Shure Gooseneck Mic)	驻极式
M3	Crown PZM Soundgrabber II	电容式
M4	AT Pro45 (Audio Technica Hanging Mic)	电容式
M5	Jabra Cellphone Earwrap Mic	驻极式
M6	Motorola Cellphone Earbud	驻极式
M7	Olympus Pearlcorder	电容式
M8	Radio Shack Computer Desktop Mic	驻极式

将 NIST 2006 的数据根据麦克风类型分成两组:电容式(M1、M3、M4 和 M7)和驻极式(M2、M5、M6 和 M8),将这两组语音样本分别训练 64 个高斯的子 DEV-UBM,然后合并成 128 个高斯的 DEV-UBM.

本文 PNCC 幂函数指数取 1/17;Gammatone 滤波器组通道数取 24;MAP 中自适应系数取 16;MCE 区分性训练 α 设为 0.48,学习系数 λ 取 0.55;MVE 区分性训练 β 取 0.52,学习系数 γ 和 ρ 取 0.65.

对于闭集设备识别用平均设备正确识别率(ACC)做指标.对于集内和集外设备识别至少要两个指标表示:用错误接受率(FAR)和错误拒绝率(FRR)来描述系统所出现的错误,当 FAR = FRR 时,用等错误率(EER)做识别指标;对于被判为集内的样本(只考虑集内被判为集内的情况,因为集外被判为集内的情况在等错误率中有所体现)也会出现被判决到集内错误设备的情况,此时用集内正确识别率 ACC 来表示.

4 实验与结果

为了评估本文算法的有效性,选取了 NIST2008 SRE Supplemental 子库[18]、TIMIT[19]、ISLE[20]、863 汉语普通话连续语音库[21]、CADCC[21]五个数据库进行实验.这五个数据库在录制时采用了不同的录音设备,将各数据库中每种录音设备作为一个设备进行实验,这五个数据库中所包含的录音设备信息以及从中提取的静音样本数如表 2.

表 2　选取数据库所包含的录音设备信息以及选取的实验样本

数据库	设备(麦克风 + 采集设备)	数据编号	静音样本数
863 语音库	Sennheiser(+ DELL 声卡)	D-863	823
CADCC	CR722(+ 创新声卡)	D-CADCC	1220
TIMIT	Pressure-sensitive	D-TIMIT	744
ISLE	Knowles VR3565	D- ISLE	767
NIST 2008 SRE Supplemental 子库 (USB 声卡)	Shure MX185 Lavalier	D-NIST01	562
	Etymotic Link-It	D-NIST02	623
	Shure MX418S Podium	D-NIST03	579
	Crown PZM-6D	D-NIST04	593
	AT3035	D-NIST05	540
	Audio Technica Pro45	D-NIST06	694
	Panasonic Camcorder	D-NIST07	635
	R0DE NT6	D-NIST08	549
	AcoustiMagic	D-NIST09	573
	Lightspeed XLC20	D-NIST10	695

在采用上述五个数据库的同时,作者也建立了一个多录音设备数据库(MRDSD, Multi-Recording Device Speech Database),该数据库的部分数据已由中文语言资源联盟 CLDC 发布[22].有 40 人参与了数据库的录制(19 女,21 男),录制时采用了多个录音设备的组合,各个设备的组合以及从中提取的静音样本数如表 3.

表 3　MRDSD 所采用的录音设备以及选取的实验样本

数据采集设备	采用麦克风	数据编号	静音样本数
Creative 声卡	Shure565SD	D-CRE01	1316
	SOMIC	D-CRE02	1302
	SONY	D-CRE03	1331
	SAMSUNG	D-CRE04	1342
	SennheiserPX90	D-CRE05	1335
Realtek 声卡	Sennheiser PX90	D-REA01	1315
	SOMIC	D-REA02	1327
	SAMSUNG	D-REA03	1315
三星录音笔	Sennheiser PX90	D-SAM01	1334
	SOMIC	D-SAM02	1307
	SAMSUNG	D-SAM03	1329
索尼录音笔	Sennheiser PX90	D-SON01	1343
	SONY	D-SON02	1322
	SAMSUNG	D-SON03	1337
奥林巴斯录音笔	Sennheiser	D-OLY01	1313
	Shure	D-OLY02	1322

实验时,每类设备中随机抽取 4 个样本作为训练数据(训练的语音样本为 20s 左右),而其余的数据作为测试数据(测试集的每个样本为 5s 左右).本文设计了三组实验:

(1) 探索传感器和数据采集设备对设备识别的影响:选取同种数据采集设备不同种麦克风的数据集(D-CRE01 ~ D-CRE05),同种麦克风不同数据采集设备的数据集(D-CRE05、D-REA01、D-SAM01、D-SON01、D-OLY01)进行实验.

(2) 进行闭集设备识别实验,考察本文改进 PNCC 特征和两步区分性训练方法的有效性,并进行不同特征、不同模型、不同算法的比较.

(3) 进行开集识别实验,考察本文算法进行集内、集外设备识别的有效性:将 30 个录音设备分为三组进行测试:5 个集内和 25 个集外、15 个集内和 15 个集外、25 个集内和 5 个集外,随机进行分组.

实验 1 麦克风和数据采集设备对设备识别的影响实验

实验设计(1)选取的 9 种设备数据的识别结果如表 4.

表 4 选取 9 种录音设备闭集识别实验(%)

模型	D-CRE01	D-CRE02	D-CRE03	D-CRE04	D-CRE05	D-REA01	D-SAM01	D-SON01	D-OLY01
D-CRE01	**91.46**	1.57	2.57	1.83	3.48	0	0	0	0
D-CRE02	2.45	**91.73**	1.25	2.37	0	0	0.03	0.05	0
D-CRE03	1.76	1.98	**92.45**	1.25	3.18	0	0	0	0
D-CRE04	2.37	1.83	1.90	**91.52**	0	0	0	0	0
D-CRE05	1.96	2.89	1.83	3.03	**93.34**	0	1.26	0	2.81
D-REA01	0	0	0	0	0	**94.35**	2.32	2.96	0
D-SAM01	0	0	0	0	0	1.73	**94.39**	0	2.45
D-SON01	0	0	0	0	0	2.65	1.47	**94.63**	0
D-OLY01	0	0	0	0	0	1.37	0.53	2.36	**94.74**

上述录音设备识别矩阵的对角线表示每类设备的正确识别率,其他为错误识别结果,对 9 种录音设备的平均正确识别率为 93.18%,说明本算法对录音设备识别是有效的.其中 D-CRE01 ~ D-CRE05 是同一声卡采集的,D-CRE05、D-REA01、D-SAM01、D-SON01、D-OLY01 为同种麦克风采集的,说明麦克风和数据采集设备都能够提供区分设备的信息.另外,表 4 中的阴影部分是错误率产生的最主要地方,上半部分刚好是同种采集设备的识别结果,下半部分刚好是同种麦克风的识别结果,说明它们类内的错误率大于它们两类之间的错误率.另一方面,相同麦克风不同采集设备的平均识别率要高于相同采集设备不同麦克风识别的平均识别率,说明数据采集设备能够提供更具区分性的设备信息.

实验 2 选取 30 个设备的闭集识别实验,进行不同特征、不同模型、不同算法的比较.

(1)不同特征实验:30 个设备的数据进行闭集设备识别,首先考察不同特征的识别结果,分别对原始 PNCC、改进 PNCC、MFCC 等进行比较(都是基于两步区分性训练模型).

如表 5,本文改进 PNCC 的平均正确识别率比 MFCC 提高了 8.72%,比原始 PNCC 提高了 6.81%,是各特征中识别率最高的,说明了本文改进 PNCC 特征对设备识别的有效性.同时,本文对 PNCC 进行的四个方面的改进使得识别率相比于原始 PNCC 都有所提升,说明了本文改进步骤的有效性.在这四个改进步骤中,CMN 和不采用预加重对识别率的提升最为明显,说明了 CMN 和预加重对设备信息的影响也最为明显.

表 5 不同特征的平均正确识别率 ACC(%)

不同特征	30 个设备平均正确识别率(ACC%)
本文改进 PNCC	90.23
原始 PNCC	83.42
原始 PNCC + 不采用预加重	86.63
原始 PNCC + 多级窗处理	84.15
原始 PNCC + 滤波器组归一化	83.74
原始 PNCC + 不采用 CMN	87.35
MFCC	81.51

(2)不同模型实验:接下来的实验采用本文改进的 PNCC 特征,但是采用不同设备模型进行 30 个设备的闭集识别,其结果如表 6.

表 6 不同录音设备识别方法的比较(%)

不同设备模型	30 个设备平均正确识别率(ACC%)
区分性训练模型	90.23
GMM-UBM 设备模型	87.52
文献[5]中算法	60.47

本文的方法使得平均正确识别率提高了 30% 左右,文献[5]是将包含说话人语音的样本直接提取录音设备特征,这样提取的特征会受到说话人信息、文本信息等因素的影响.另一方面,在录音设备特别多的情况下,文献[5]中所用朴素贝叶斯分类器难以建立具有区分性的分类器.而本文的方法可以借助 GMM-UBM 提供足够多的高斯数来建立多个设备的模型,并通过两步区分性训练获得具有区分性的录音设备模型.两步区分性训练模型和 GMM-UBM 相比,有 2.71% 的提升,说明本文的区分性训练方法有效.

实验 3　集内和集外的录音设备识别实验

针对实验设计(3)部分的分组分别进行实验,获得的实验结果如表 7.

表 7　不同录音设备识别方法的比较(%)

数据	EER(%)/ACC(%)	
	GMM-UBM 模型	两步区分性训练模型
5 内/25 外	14.27/97.57	12.25/98.73
15 内/15 外	17.38/95.14	15.17/96.65
25 内/5 外	18.95/91.86	16.26/93.32

如表 7,两步区分性训练对于集内和集外设备识别的等错误率低于 17%,说明本文算法对集内和集外设备具有较好区分能力.而三组分组实验中两步区分性训练模型都要优于 GMM-UBM 模型,说明本文两步区分性训练使得设备模型之间的区分性变大,分类误差变小,因此集内的平均正确识别率都要比 GMM-UBM 高.另一方面,采用最小确认误差训练,集内模型和通用背景模型之间的距离变大,从而使得等错误率比 GMM-UBM 小.因此,对于集内集外录音设备的识别,本文的两步区分性训练模型有更好的识别效果.

5　结论

本文提出了一种基于改进 PNCC 特征和两步区分性训练的录音设备识别方法,由于音频中的静音包含了完整的设备信息,且不受说话人和文本等因素的影响,因此从静音段提取改进的 PNCC 特征,利用了 PNCC 的长时帧分析去除背景噪声对设备信息的影响.在模型方面,以 GMM-UBM 为基准模型,并通过两步区分性训练调整集内设备模型和通用背景模型,提升模型区分能力.第一步采用最小分类误差训练,调整集内设备模型,使得集内各模型之间的分类误差最小;第二步采用最小确认误差训练,调整通用背景模型,使得集内各模型和通用背景模型的确认误差最小.实验表明,对于 30 种设备闭集识别的平均正确识别率为 90.23%,与文献[5]相比,性能提升了 30%;对于 15 个集内和 15 个集外设备的测试,等错误率为 15.17%,集内正确识别率为 96.65%,验证了本文算法的有效性和可靠性.

对于录音设备辨识,本文的工作做了初步研究,只获得一些初步的结果,对于算法的鲁棒性,还需要扩充更多的录音设备数据做进一步的研究工作.另外,直接从包含语音的音频中提取设备信息有着更好的现实意义,这将是今后的一个研究方向.

参考文献

[1] 钟巍,孔祥维,尤新刚,等.基于态函数的离散分数余弦倒谱变换在取证话音信息隐藏中的应用[J].电子学报,2012,40(3):595-599.
Zong Wei, Kong Xiang wei, You Xingang, et al. Forensic speech information hiding using fractional cosine cepstrum transform[J]. Acta Electronica Sinica, 2012, 40(3): 595-599. (in Chinese)

[2] 周琳娜,王东明,郭云彪,等.基于数字图像边缘特性的形态学滤波取证技术[J].电子学报,2008,36(6):1047-1051.
Zhou Lin-na, Wang Dong-ming, Guo Yun-biao, et al. Exposing digital forgeries by detecting image blurred mathematical morphology edge[J]. Acta Electronica Sinica, 2008, 36(6): 1047-1051. (in Chinese)

[3] Kang Xiangui, Li Yinxiang, Qu Zhenhua, et al. Enhancing source camera identification performance with a camera reference phase sensor pattern noise[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 393-402.

[4] Chiang Peiju, N Khanna, A Mikkilineni, et al. Printer and scanner forensics[J]. IEEE Signal Processing Magazine, 2009, 26(2): 72-83.

[5] C Kraetzer, A Oermann, J Dittmann, et al. Digital audio forensics: A first practical evaluation on microphone and environment classification[A]. Proceedings of IEEE MMSEC[C]. New York: IEEE, 2007. 63-74.

[6] Wang Zhifeng, Wei Gang, He Qianhua. Channel pattern noise based playback attack detection algorithm for speaker recognition[A]. Proceedings of ICMLC[C]. Guilin: IEEE, 2011. 1708-1713.

[7] D A Reynolds, T F Quatieri, R B Dunn. Speaker verificatioing adapted gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1-3): 19-41.

[8] 何磊,武健,方棣棠,等.最大后验估计和最近邻线性回归结合的说话人自适应方法[J].电子学报,2000,28(11):55-58.
He Lei, We Jian, Fang Di-tang, et al. A novel speaker adaptation method based on map and NNLR[J]. Acta Electronica Sinica, 2000, 28(11): 55-58. (in Chinese)

[9] B H Juang, S Katagiri. Discriminative learning for minimum error classification[J]. IEEE Transactions on Signal Processing,

1992,40(12):3043 – 3054.

[10] S Katagiri, C H Lee, B H Juang. New discriminative training algorithms based on the generalized probabilistic descent method[A]. Proceedings of NNSP[C]. Los Angles: IEEE, 1991.299 – 308.

[11] 陈雁翔,刘鸣.基于发音特征的音视频说话人识别鲁棒性的研究[J].电子学报,2010,38(12):2920 – 2924.
Chen Yan-xiang, Liu Ming. Research on robustness of audio-visual speaker recognition based on articulatory features[J]. Acta Electronica Sinica, 2010, 38(12): 2920 – 2924. (in Chinese)

[12] Kim Chanwoo, R M Stern. Power-normalized cepstral coefficients (PNCC) for robust speech recognition[A]. Proceedings of ICASSP[C]. Tokyo: IEEE, 2012.4101 – 4104.

[13] Li Qi, Zheng Jinsong, A Tsai, et al. Robust endpoint detection and energy normalization for real-time speech and speaker recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2002, 10(3): 146 – 157.

[14] 王志锋,贺前华,张雪源,等.基于信道模式噪声的录音回放攻击检测[J].华南理工大学学报(自然科学版),2011,39(10):7 – 12.
Wang Zhifeng, He Qianhua, Zhang Xue-yuan, et al. Playback attack detection based on channel pattern noise[J]. Journal of South China University of Technology (Natural Science Edition), 2011, (10): 7 – 12. (in Chinese)

[15] T Kinnunen, R Saeidi, F Sedlak, et al. Low-variance multitaper MFCC features: A case study in robust speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(7): 1990 – 2001.

[16] M Slaney. Auditory toolbox version 2[R]. Interval Research Corporation Technical Report, 1998.

[17] Du Zhihui, Li Xiangyu, Wu Ji. Accelerating the training of HTK on GPU with CUDA[A]. Proceedings of IPDPSW[C]. Beijing: IEEE, 2012. 1907 – 1914.

[18] NIST. Speaker Recognition Evaluation[OL]. http://www.itl.nist.gov/iad/mig/tests/sre/, 2009 – 09 – 03.

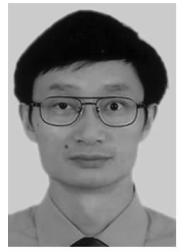
[19] LDC. TIMIT acoustic-phonetic continuous speech corpus [OL]. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp? catalogId = LDC93S1, 1993 – 09 – 20.

[20] E Atwell, P Howarth, C Souter. The ISLE corpus: Italian and German spoken learner's English[J]. ICAME Journal, 2003, 27(1): 5 – 18.

[21] Li Yanxiong, He Qianhua, Li Tao, et al. A detection method of lip-smack in spontaneous speech[A]. Proceedings of I-CALIP[C]. Nevada: IEEE, 2008.292 – 297.

[22] CLDC. Playback and Speaker Recognition Database[OL]. http://www.chineseldc.org/resource_info.php? rid = 154, 2012.

作者简介



贺前华 男.1965 年 2 月生,湖南邵东人.现任华南理工大学电子与信息学院教授、博士研究生导师、副院长,广东省"千百十"人才工程培养对象.研究领域主要有多媒体信息检索技术、数字音频侦测技术、信息安全身份认证技术、音视频双模态语音识别技术和嵌入式系统设计与应用等.

王志锋(通讯作者) 男.1985 年 2 月生,湖北武汉人.华南理工大学信号与信息处理专业博士,美国卡内基梅隆大学计算机学院联合培养博士.研究方向为语音信号处理、数字取证技术以及生物特征识别.已发表相关论文近 10 篇.
E-mail: eezfwang@gmail.com