基于信道信息的回放攻击检测研究

柯宏宇[1],高奕宁[1],郝雪营[1],黄　涛[1,2]
(1. 武汉邮电科学研究院,湖北 武汉 430074;
2. 武汉烽火众智数字技术有限责任公司,湖北 武汉 430074)

摘　要:生物识别具有广阔的研究前景,说话人识别作为生物识别的重要组成部分,涉及人们日常生活的许多方面。随着高保真录音及回放设备的普及,说话人识别系统的安全性面临回放攻击的严重挑战,由于回放攻击语音与真实语音具有相同的声纹,导致常规说话人识别很难有效鉴别声音的真实性,且生活中存在的噪声,会在一定程度上干扰系统的识别,这也对系统的鲁棒性提出了要求。因此,该文提出一种基于信道信息的录音回放攻击检测方法,提取 Legendre 系数及其统计特征为主要判别依据,同时使用语音基频特征与 MFCC 特征作为辅助特征,并使用一种基于支持向量机的决策融合算法进行判别,给予特征不同的权重。实验结果表明,多种特征相结合的方式,相较于现有其他方法,能在有效检测回放语音攻击的同时,提升系统的鲁棒性,在噪声环境下识别率平均提高了 1.5%。

关键词:语音识别;信号处理;信道攻击;机器学习;决策融合
中图分类号:TN391.42　　**文献标识码**:A　　**文章编号**:1673-629X(2021)06-0118-05
doi:10.3969/j.issn.1673-629X.2021.06.021

Research on Replay Attack Detection Based on Channel Information

KE Hong-yu[1],GAO Yi-ning[1],HAO Xue-ying[1],HUANG Tao[1,2]
(1. Wuhan Research Institute of Posts and Telecommunications,Wuhan 430074,China;
2. Wuhan Fiberhome Zhongzhi Digital Technology Co.,Ltd.,Wuhan 430074,China)

Abstract:Biometrics has broad research prospects. As an important component of biometrics,speaker recognition involves many aspects of people's daily lives. With the popularity of high-fidelity recording and playback equipment,the security of speaker recognition systems is facing serious challenges from playback attacks. As the playback attack voice has the same voiceprint as the real voice,it is difficult for conventional speaker recognition to effectively identify the authenticity of the voice. In addition,the noise in life will interfere with the recognition of the system to a certain extent,which also puts forward requirements for the robustness of the system. Therefore,we propose a detection method for recording and playback attacks based on channel information,extracting Legendre coefficients and their statistical features as the main criterion,using speech fundamental frequency features and MFCC features as auxiliary features,and using a support vector machine-based decision fusion algorithm judges by giving different weights to the features. Experiment shows that compared with other existing methods,the combination of multiple features can effectively detect playback speech attacks while improving the robustness of the system. The recognition rate in a noisy environment is increased by an average of 1.5%.

Key words:speech recognition;signal processing;channel attacks;machine learning;decision fusion

0　引　言

近年来,人工智能快速发展,促进了人机交互应用的加深。生物识别作为人机交互的重要一环,具有广阔的研究前景[1]。该技术利用人体与生俱来的较稳定特征进行身份验证,包括指纹、声纹、虹膜等,其中声纹识别具有非接触、高可靠、低成本等优势,成为了目前主流身份判定特征之一。然而,随着具备高保真录音功能电子设备的普及,清晰度较高的录音获取变得简单,这在一定程度上降低了不法分子偷录语音假冒认证的难度。如何在声纹识别任务中,有效区分输入语音是否为回放语音,对守护公民财产安全具有重大意义。目前,关于回放攻击检测的研究,大多与说话人识别联系在一起,缺乏对这一问题的单独探究。该文针对偷录语音与真实语音在信道中存在的信道噪声长时

收稿日期:2020-08-03　　**修回日期**:2020-12-03
基金项目:湖北省重大科技专项(2018AAA063)
作者简介:柯宏宇(1996-),男,硕士研究生,研究方向为语音信号处理、无人机通信;黄　涛,硕士,正高职高级工程师,研究方向为网络安全、人工智能。

统计特征差异,提出一种有效的检测手段,从模型鲁棒性、有效性两个方面对回放攻击展开研究。

1　研究背景

回放语音攻击可分为 4 类:录音重放、波形拼接、语音合成和语音模仿[2]。后三类攻击模式需对说话人声道模型建模,由于个体间的声道差异性较大,语音模仿的普适性较差,且合成拼接技术精度难以保证,因此实际案例应用较少。录音重放与真实语音具有相同的声纹信息与语音特征,因此最具威胁。尽管语音识别研究始于二十世纪五十年代,但是直到 1999 年才首次使用一男一女的语音样本评估录音重放攻击对系统的破坏性[3]。文献[4]使用远场偷录的语音进行录音回放攻击,实验结果表明,该录音回放检测系统在信噪比较低环境中的错误接受率(false acceptance rate,FAR)较高。为提高识别精度,文献[5]提出基于语谱图的检测算法,并在后续工作中引入了均值和方差参数进行相似度比对[6],有效降低等错误率(equal error rate,EER)。文献[7]在语谱图上引入中点相对位置这一概念,并着重研究麦克风采集距离对识别的影响,同时比对了不同信噪比下的检测结果。针对远场偷录所产生的低频无关因素,文献[8]提出了一种基于光谱比率(spectral ratio,SR)、低频比率(low frequency ratio,LFR)和调制系数构成特征集的语音检测算法,并使用支持向量机(support vector machine,SVM)进行分类,提高了不同场景下的识别正确率。除了采用语音特征参数对录音回放进行研究,有部分研究者从信道信息着手。文献[9]基于高通滤波器和统计帧,文献[10]采用经验模态分解滤波器,均实现了信道特征的提取,并在录音回放检测时获得了较好效果;文献[11]通过借鉴高斯混合模型和通用背景模型(Gaussian mixture model-universal background model,GMM-UBM)在说话人识别中的应用模式,成功提取了语音静音段特征,有效降低了 EER,但是实验规模较小,有待进一步扩充。除此之外,文献[12]采用了自适应子带谱熵法进行静音区提取,并改进了梅尔倒谱系数(Mel frequency cepstrum coefficient,MFCC)提取过程,包括在预处理时不进行预加重,加窗时使用多级窗代替单级窗,以及采用归一化 Mel 滤波器组进行特征提取等措施,实验结果表明,系统 EER 有效降低,但该研究假设环境安静无干扰,而这与实际使用存在差异。2018 年,文献[13]在总结现有对抗措施后,提出采用线性预测(linear prediction,LP)参数替代传统的光谱相关信息,实验结果证明,相较于已有参数,LP 参数具有更强的鲁棒性。但是选取单个参数作为性能指标进行训练时,所需训练数据量较大才可得到泛化性能较好的系统模型,且容易出现过拟合现象。

基于此,该文提出了一种基于决策融合的信道信息回放检测算法,提取 Legendre 系数及其统计特征,语音基频特征以及 MFCC 特征,并使用三个 SVM 进行决策,而后以一定权重融合以上三个参数进行总体决策,实现回放攻击检测。

2　相关工作

本节将针对文中所提问题,简要回顾语音信号的一般处理流程,包括语音信号的预处理与一些常用语音特征的提取方法。

2.1　预处理

语音信号包含人类发声器官本身以及采集设备带来的混叠,通常存在高次谐波失真、高频分量不足等缺陷。实际中,需要进行预处理以平滑信号,为后续处理提供良好基础。常用预处理手段包括:预加重、端点检测、分帧、加窗处理四部分。预加重能消除发声过程中声带和嘴唇对高频语音信号的抑制效应[14],从而使高频段信号的能量衰减得到补偿。具体的预加重公式如下:

$$H(z) = 1 - \alpha z^{-1}$$

其中,α 表示预加重系数,依据经验,文中设置 $\alpha = 0.98$。端点检测是指在输入信号中检测语音的起止位置,将语音的沉默片段去除[15]。端点检测可以在减少计算量的同时消除无关变量对系统识别的影响,常用检测指标包括信号能量和短时过零率等统计特性。分帧可以将长时、非稳态信号分成短时、近似平稳信号,进而可采用语音短时分析技术,通常采用的帧长为 10 ms ~ 30 ms,为保证信号过渡的连续性,帧移往往小于帧长,文中将帧长设置为 10 ms。加窗是指将语音帧与一个窗函数相乘,减小语音信号的截断效应,使语音帧两端平滑过渡到零。

2.2　特征提取

生理学研究表明,人的听觉系统是一个出色的说话人识别系统,对不同频率的声波有不同程度的灵敏度,其敏感程度可以由对数函数较好的表征。为了更好地拟合人耳听觉特性,通常采用倒谱系数刻画语音特征,倒谱系数由对语音信号的功率谱取对数得到,目前已广泛应用于语音识别领域。常用的语音倒谱系数特征包括线性预测倒谱系数(linear predictive cepstrum coefficient,LPCC)、梅尔倒谱系数(Mel frequency cepstrum coefficient,MFCC)、逆梅尔倒谱系数(inverted-mel frequency cepstrum coefficient,IMFCC)、耳蜗倒谱系数(cochlear frequency cepstrum coefficient,CFCC)等[16]。其中,MFCC 源于对人耳听觉特性的分析,计算较为方便,因而使用广泛。实际频率 f 与 Mel

频率间对应关系可由下式表示：

$$F_{\text{mel}} = 2\ 595\lg(1 + f/700)$$

其中，f 单位为赫兹，梅尔频率单位为 Mel。具体说来，在对 MFCC 特征进行提取时，可以依据两者间的对应关系，划分出三角滤波器组，即 Mel 滤波器组，该滤波器组在以赫兹为频率的轴上呈非等距分布，而在 Mel 频率轴上呈等间距分布。滤波器组一般由若干个三角滤波器排列构成，滤波器组带宽大致范围为 4 000 赫兹，包含人耳听觉敏感频率范围 3 000 赫兹至 4 000 赫兹。MFCC 滤波器组分布如图 1 所示。

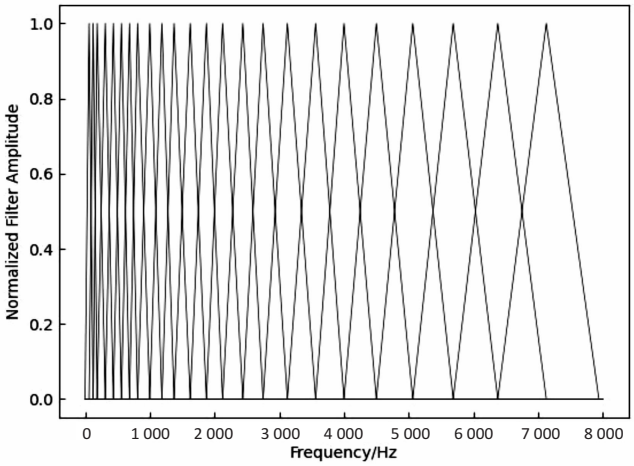

图 1 MFCC 滤波器组分布图

3 录音回放检测系统

该文提出一种基于信道信息的多参数回放攻击检测系统，系统整体框图如图 2 所示。对预处理后的语音信号提取 Legendre 多项式系数与其统计特征用以拟合信道模式噪声，同时提取基频特征与 MFCC 特征作为辅助特征，用于描述信道信息，在最后进行融合决策。

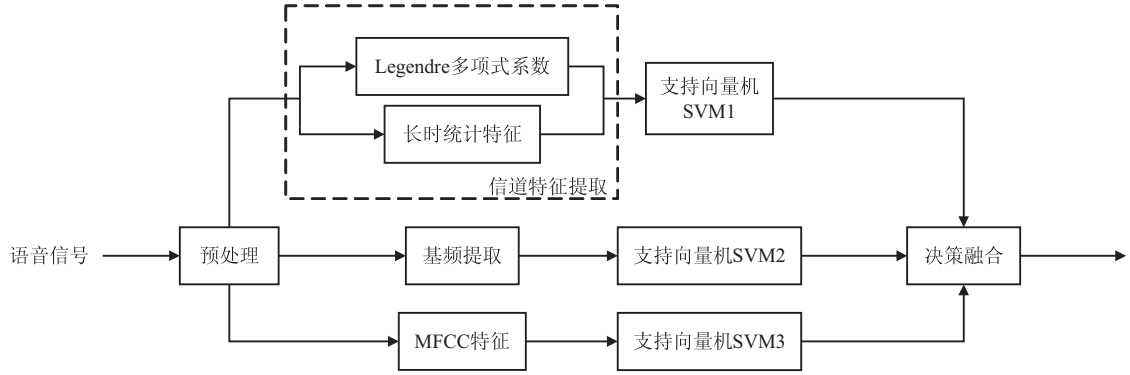

图 2 系统整体框图

3.1 噪声参数

该文采用 Legendre 多项式拟合信道模式噪声。Legendre 多项式是一种正交基底，较好地反映了帧间的关联，在作为录音回放检测指标时有较强的鲁棒性[17]。目前常采用六阶多项式系数对信道模式噪声进行模拟，其拟合表达式如下：

$$f(x) = \sum_{n=0}^{\infty} L_n P_n(x)$$

其中，L_n 表示多项式系数，n 表示阶数，$P_n(x)$ 则为 Legendre 多项式通项公式：

$$P_n(x) = \frac{1}{2^n \cdot n!} \cdot \frac{d^n}{dx^n}(x^2 - 1), x \in [-1, 1]$$

目前常采用六阶多项式 $(L_0, L_1, L_2, L_3, L_4, L_5)$ 系数对噪声进行模拟。零阶矢量表示信道模式噪声直流分量；一阶矢量表示信道噪声分布曲线斜率；二阶矢量表示信道噪声分布曲线曲率；高阶矢量则表示信道噪声分布曲线细节信息。考虑到信道短时特征随时间变化较为缓慢，该文采用 12 阶向量表征信道模式噪声特征，其中前六阶参数表征零阶到五阶 Legendre 多项式系数，后六阶参数加入 Legendre 多项式系数的长时统计特征，分别表征信道模式噪声的最大值，最小值，均值，中值，极差与标准差。

3.2 基频特征

基音是指话音中频率最低的分音，其频率被称为基频，可以用于反映说话人生物学特征，如年龄、性别等，是一种较为稳定的特征，目前常应用于刑侦破案中。常用的提取方法主要分为时域法、频域法以及统

计法[18]。时域法包含两类,分别为自相关算法以及平均幅度差算法。自相关算法通过自相关函数求取基频特征,自相关函数是用于计算语音信号序列的功率谱密度,可以反映语音信号在时间上的关联性,其公式表示如下:

$$R_n(k) = \sum_{m=0}^{N-k-1} S_n(m) \cdot S_n(m)$$

其中,$S_n(m)$ 为采样后的语音信号表达式,N 为窗长,k 为采样点数。由于相关函数在基音周期整数倍处取得极值,因此,通过计算相邻两个最大峰值间距,并将距离参数由时域变换到频域,即可得出基频值。同时噪声信号经自相关运算后主要集中于零点低频段,故该算法可以一定程度上区分噪声与输入语音[19];平均幅度差算法与自相关算法原理类似,不同之处在于自相关函数计算功率谱时为求乘积,算法时间复杂度往往较高,为了规避较大的运算量,可以采用平均幅度差计算方式求取基频。语音信号的短时平均幅度差函数公式表示如下:

$$F_n(k) = \sum_{m=0}^{N-k-1} |S_n(m+k) - S_n(m)|$$

其中,$S_n(m)$ 为某采样点的幅度,$S_n(m+k)$ 为相邻采样点的幅度,N 为窗长,k 为采样点数。该算法原理是周期信号中,相距为周期整数倍的采样点的幅值相等。除了计算方式的区别,平均幅度差算法所关注的性能指标是波谷而非自相关算法中的波峰。这是因为波谷相较于波峰更加陡峭,错判率更低,且采用中心削波后准确率更高[20]。

频域法以倒谱法为主,该方法利用语音信号倒谱特征提取基频,由于语音信号倒谱特征中含有声门激励周期,即基频信息,通过计算该周期即可得出基频[21]。在倒谱域中,由于激励信息与声道响应为加性关系,但由于所处频段不同,所以波形上分离度明显,计算基频精度较高,但是计算量过大,不适用于实时性要求较高的场合。

统计法是通过机器学习方法,提取时域特征或者频域特征后,分析自相关函数的周期性或者相邻采样点间幅度差,算出基频值后,得出基频值与输入语音时频域特征间的对应关系,生成训练模型,进而在新输入语音时可直接求出其基频值[22]。为对抗噪声带来的干扰,同时更好地确保说话对象的唯一性,该文融合基频特征作为一个辅助指标,减少语音回放信道攻击对检测系统的影响。

3.3　决策融合

一般的机器学习方法将训练重心放在单个性能指标上,忽略了其他可能优化性能指标的信息。而实际应用场景中测试集与训练集往往存在一定差异。因此测试时,训练模型如果仅采用单个指标进行决策,出现拟合失真的概率往往较高[23]。决策融合是一种通过共享多个性能指标的表征,同时使各指标之间相互影响的策略,具有较好的泛化性能。该文采用如下公式进行决策融合:

$$f(x) = \alpha x_1 + \beta x_2 + \gamma x_3$$

其中,α、β、γ 分别为各个决策的融合权重,x_1、x_2、x_3 分别为 Legendre 多项式决策结果,基频决策结果以及基于 MFCC 特征的决策结果。由于信道模式噪声特征在安静无噪声场景下已具有较好的录音回放检测表现,而该文在此基础上进一步考虑了多种信噪比条件下的录音回放检测,因此本实验中,$\alpha = 0.7$、$\beta = 0.2$、$\gamma = 0.1$,采用信道模式噪声作为主要判别依据,基频特征权重次之,最后是 MFCC 特征参数权重。经过调试,最终的接受阈值设置为 0.75。

4　实验测试

本节将对文中实验中涉及的数据集构造以及实验方法进行说明。实验计算机的 CPU 为 AMD Ryzen 7 3800X 8-Core,32G 内存,Windows 10 操作系统。实验平台为 MATLAB 2017b。

4.1　数据集

由于目前针对录音重放的开源数据集较少且不易直接获得,文中基于语音数据集 AISHELL-2019B-EVAL[24]对所需数据进行了制作,用以研究不同偷录设备翻录语音对检测的影响。制作时通过运行转录程序播放原数据集语音,同时采用监测麦克风进行收声,具体转录设备信息如表 1 所示。

表 1　基于 AISHELL 数据集语音样本制作详情

播放设备	偷录设备	样本数
SONY SRS-HG10	PHILIPS SHM1008	3 440

在信号处理中,信号功率与噪声功率的比值称为信噪比,其定义式如下:

$$\mathrm{SNR} = 10\lg(S/N)$$

其中,S 为信号功率,N 为噪声功率,SNR 单位为 dB。为确保系统性能的鲁棒性,将表 1 所获得数据按 0 dB、3 dB、5 dB、10 dB、20 dB 的信噪比与白噪声进行混合后,作为现有方法的对照组进行后续实验。

4.2　实验结果及分析

在对输入语音进行预加重、分帧、加窗等预处理流程后,计算信道模式噪声特征。同时对比文献[9,11-12]的方法,实验结果如表 2 所示。可以看到,噪声的引入对回放语音检测有一定的影响,随着信噪比的降低,识别精度总体呈下降趋势,其中,噪声对文献[11]的方法影响较为严重,原因之一在于低信噪比环境下

无法有效进行端点检测。文献[9]采用信道模式噪声统计特征作为判别依据,随着输入语音信噪比的增加,识别率稳定上升。但是由于决策指标单一,相比而言,文中提出的决策融合算法,能在有效对抗干扰的同时,提高模型在噪音环境中的表现。

表 2 不同信噪比下对比识别精度结果

方法	0	3	5	10	20
文献[9]	0.842 7	0.878 7	0.918 3	0.984 3	0.997 6
文献[11]	0.573 3	0.682 1	0.767 7	0.805 7	0.843 9
文献[12]	0.521 8	0.624 8	0.764 2	0.842 7	0.903 6
文中方法	0.877 4	0.892 3	0.929 5	0.987 4	0.993 6

实验结果表明,该文所提出的基于决策融合的信道信息检测方法简洁有效,系统的识别精度在不同信噪比环境下较为稳定,实现了攻击检测目标。

5 结束语

提出了一种回放攻击检测算法,并在噪声环境下研究了模型的鲁棒性,取得较为稳定的效果。除此之外,该模型是轻量级的,因此可以部署在移动端,具有一定实际应用价值。一部分研究认为,信道信息主要集中在高频部分,为了在高频上获得较高的分辨率,挖掘高频部分的有效信息,一些新的滤波器组或特征被设计并用于实践,该文也对部分特征进行了实验,识别效果有待进一步提升。如何提取更有效更稳定的特征,也是未来工作的一个方向。

参考文献:

[1] 叶 硕,褚 钰,王 祎,等. 语音识别中声学模型研究综述[J]. 计算机技术与发展,2020,30(3):181-186.

[2] 蒋 晔. 基于短语音和信道变化的说话人识别研究[D]. 南京:南京理工大学,2013.

[3] LINDBERG J,BLOMBER M. Vulnerability in speaker verification-a study of technical impostor techniques[C]//Sixth European conference on speech communication and technology (EUROSPEECH'99). Budapest,Hungary:[s. n.],1999.

[4] VILLALBA J,LLEIDA E. Speaker verification performance degradation against spoofing and tampering attacks[C]// FALA2010 conference. Vigo, Spain:[s. n.], 2010: 131-134.

[5] SHANG Wei, STEVENSON M. A playback attack detector for speaker verification systems[C]//2008 3rd international symposium on communications, control and signal processing. Saint Julian's,Malta:IEEE,2008:1144-1149.

[6] SHANG Wei,STEVENSON M. Score normalization in playback attack detection[C]//2010 IEEE international conference on acoustics,speech and signal processing. Dallas,TX: IEEE,2010:1678-1681.

[7] GALKA J,GRZYWACZ M,SAMBORSKI R. Playback attack detection for text-dependent speaker verification over telephone channels[J]. Speech Communication,2015,67: 143-153.

[8] VILLALBA J, LLEIDA E. Preventing replay attacks on speaker verification systems[C]//2011 Carnahan conference on security technology. Barcelona:IEEE,2011:1-8.

[9] 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.

[10] 王志锋. 基于信道信息的数字音频盲取证关键问题研究[D]. 广州:华南理工大学,2013.

[11] 张利鹏,曹 翠,徐明星,等. 防止假冒者闯入说话人识别系统[J]. 清华大学学报:自然科学版,2008(S1):699-703.

[12] 王茂蓉,周 萍,景新幸,等. 基于信道信息的录音假冒者检测系统研究[J]. 计算机仿真,2016,33(2):460-464.

[13] MISHRA J,SINGH M,PATI D. Exploring linear prediction residual signal for developing countermeasures to playback attacks[C]//2018 IEEE international students' conference on electrical,electronics and computer science (SCEECS). Bhopal:IEEE,2018:1-6.

[14] 张 震,王化清. 语音信号特征提取中 Mel 倒谱系 MFCC 的改进算法[J]. 计算机工程与应用,2008,44(22):54-55.

[15] 叶 硕,彭春堂,杜珍珍,等. 基于 DTW 的孤立词语音识别系统设计[J]. 长江大学学报:自科版,2018,15(17):33-37.

[16] 刘丽岩. 基于 MFCC 与 IMFCC 的说话人识别研究[D]. 哈尔滨:哈尔滨工程大学,2008.

[17] 代亚丽. 防录音回放攻击的说话人认证算法及系统设计[D]. 武汉:武汉理工大学,2014.

[18] 曹洪林,孔江平,王英利. 说话人基频与生理参数关系初探[J]. 清华大学学报:自然科学版,2013,53(6):848-851.

[19] 戚园园. 哼唱检索中的基音频率提取研究[D]. 北京:北京邮电大学,2013.

[20] 成新民,曾毓敏,赵 力. 一种改进的 AMDF 求取语音基音的方法[J]. 微电子学与计算机,2005,22(11):162-164.

[21] 陈敏敏,张云刚,王 智. 基频提取算法的研究与评价[J]. 微型电脑应用,2012,28(9):14-16.

[22] 陈 萧,徐 波. 改进的用于口语处理的基频提取算法[J]. 清华大学学报:自然科学版,2017,57(1):95-99.

[23] 许棣华,王志坚. 基于多任务学习的邮件过滤系统的研究[J]. 计算机技术与发展,2010,20(10):137-140.

[24] QIN X,BU H,LI M. HI-MIA:a far-field text-dependent speaker verification database and the baselines[C]//2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7609-7613.