基于卷积神经网络的重录语音检测算法

赵雅珺,王 泳,张梦鸽
(广东技术师范大学,广东 广州 510665)

摘 要:使用重录语音冒充他人身份会为社会安全带来严重威胁。但是,目前对于重录语音检测的研究仍相对较少。已有的重录语音检测方法一般集中于传统的信号处理方法,其特征提取的算法较为复杂,具有较大的局限性。为此,提出一种基于卷积神经网络的重录语音检测算法。所提出的网络结构依据语音信号的时频特征进行特殊设计,与时频图的特征分布特点高度契合,能将训练参数分配到更合理的地方,从而能使用更有效的特征来训练更紧凑的参数,因而大大降低了模型过拟合风险。为了验证该算法的性能以及通用性,采用不同录制设备、录制环境及录制距离的重录语音对算法进行测试。实验结果表明,该算法对不同设备和场景下录制的语音均达到了 99.8% 以上的检测率。由于采用时长 0.2 秒极短语音段作为检测数据得到以上的准确率,说明算法在实际应用场景中具备广泛的适用性。

关键词:安全;重录语音检测;卷积神经网络;时频特征
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2020)02-0171-07
doi:10.3969/j.issn.1673-629X.2020.02.033

A Recaptured Speech Detection Algorithm Based on Convolutional Neural Network

ZHAO Ya-jun, WANG Yong, ZHANG Meng-ge
(Guangdong Polytechnic Normal University, Guangzhou 510665, China)

Abstract: The use of recaptured speech to impersonate other people's identity can pose a serious threat to social security. However, the research on the detection of recaptured speech is still insufficient. The existing efforts of recapture detection mainly focus on traditional signal processing methods of which the feature extraction algorithms are complicated and limited. Therefore, we propose a recaptured speech detection algorithm based on convolutional neural network. This network architecture is specifically designed based on spectrogram characteristics, which is highly consistent with the distribution of spectral features. It can allocate the training parameters to a more reasonable place, so that more effective features can be used to train more compact parameters, thus greatly reducing the risk of model over-fitting. In order to verify the performance and versatility of the proposed algorithm, it is tested in different scenes including different recording equipments, recording environments and recording distances. The experiment shows that the proposed algorithm can achieve accuracy rates higher than 99.8% in different recording scenes. As short speech segments of 0.2 second are used in the experiment to obtain the results above, it is indicated that it is widely applicable in practical applications.

Key words: security; recaptured speech detection; convolutional neural network; spectrogram

1 概 述

已有研究证明,语音转换(voice conversion, VC)、语音合成(speech synthesis, SS)及重录语音等欺骗性语音能有效地欺骗说话人识别(automatic speaker recognition, ASV)系统,从而冒充他人登入系统[1-5],对社会安全产生严重威胁。其中,VC 及 SS 需要目标说话人较多的语音信息及特征,再加上现有算法尚未完全成熟,实现成本及难度相对较高;而重录语音利用低廉的录音设备即可轻松获得,且重录语音基本包含目标人物语音的所有特征,因此,相对 VC 及 SS 更具威胁。为此,文中对重录语音的检测算法进行研究。

在已有的研究中,针对欺骗性语音安全性的研究主要集中在对 VC 及 SS 的检测算法上。Hanilci C 等提出了利用语音信号的线性预测残差提取相位特征进

收稿日期:2019-02-18 **修回日期**:2019-06-20 **网络出版时间**:2019-09-25
基金项目:国家自然科学基金(61672173);广东省普通高校特色创新项目(2015KTSCX083)
作者简介:赵雅珺(1992-),男,硕士研究生,研究方向为深度学习、智能信息处理、数字语音处理等;王 泳,博士,副教授,CCF 会员(10543M),通信作者,研究方向为信息取证及信息隐藏等。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190925.1521.038.html

行欺骗检测的方法[6]；Kamble M 等提出了基于能量分离算法的瞬时频率余弦系数，用于检测真假语音[7]；Muckenhirn H 等通过计算一阶和二阶频谱统计量并将它们提供给分类器来检测攻击[8]；Janicki A 等提出了利用线性预测（linear prediction）残差信号提取基于音频质量特征的算法[9]；Alam J 提出了一种基于无限脉冲响应常数 q 变换特征表示的欺骗检测算法[10]。此外还有运用卷积神经网络的检测算法[11-12]，以及运用高斯混合模型（GMM）、动态时间规整（DTW）模型、深度学习等其他方法的检测算法[13-20]。

然而，针对重录语音检测的报道相对较少。文献[21]提出了一种利用频域线性预测框架提取时间包络特征的方法，用于检测重播欺骗攻击。采用高斯混合模型（GMM）和卷积神经网络（CNN）两种建模方法，对真实和伪造数据的 GMM 进行训练，CNN 子系统用来区分真实和重放语音。其融合系统结果的误差率为 9.7%，有待提高。文献[22]则应用了线性预测（linear prediction）残差信号。该文指出线性预测残差信号是一种准周期脉冲序列，如果样本被改变，感知到的线性预测残差信号将是不同的，由此，将 RMFCC（residual mel frequency cepstral co-efficient）作为线性预测残差信号的代表特征，应用在重播攻击检测系统中。文献[23]利用分层散射分解系数和逆梅尔倒谱系数（IMFCC）分析频谱在低端和高端存在的差异，然后采用 2 级 GMM 后端来获得真实语音和重放语音之间的逻辑似然比，再使用 HTK[24] 和 VLfeat 工具包[25] 对 GMM 进行训练。文献[26]提出基于卷积神经网络检测重录语音的算法，该算法利用电网频率（ENF）及其谐波组成的组合作为 CNN 的输入，此算法要求录音设备必须插入电网，以从语音信号中提取 ENF；若录音设备自带电源，则无法使用此方法。在涉及安全问题时，更大的可能性是录音设备为自带电源的设备，因此该算法在实际应用中具有明显的局限性。

上述研究尚存在一些问题：传统算法提取特征过程比较复杂；算法均缺乏通用性，对训练和测试环境或设备不同时比较脆弱。这些工作具有启发性，但是也反映出此类语音取证面临的困境，包括在没有标准化的情况下如何设定取证场景、取证场景是否与真实世界相符、录音数量是否不足等问题。为此，文中提出了一种基于新的卷积神经网络且对不同场景鲁棒的重录语音检测算法。网络的数据输入形式采用语音信号经过分帧的时频图，网络结构层包括若干卷积层、池化层，在实验中先分别对不同录制语音设备、距离及环境等重录语音影响因子进行研究分析，然后提出最终的训练方法，并对所有的不同条件下的重录语音进行检测。实验结果表明，在不同的实验条件下，该算法均达到了较高的检测率，因此具有通用性。

2 数据预处理和网络结构

构建的卷积神经网络模型结构如图 1 所示。网络结构中每一层的参数情况如表 1 所示。

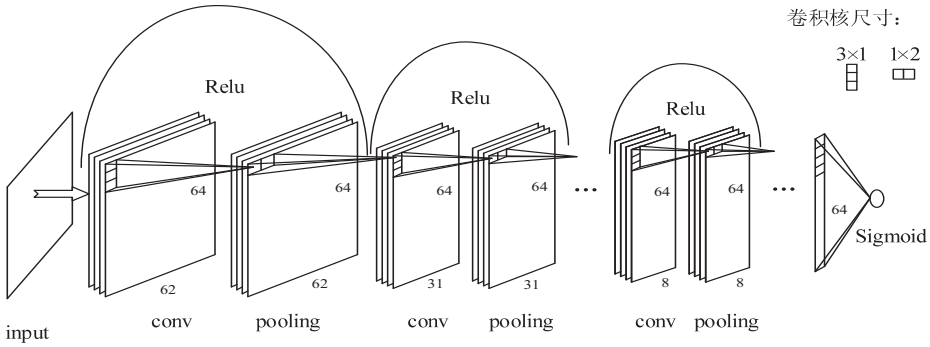


图 1 网络结构

表 1 网络结构的参数情况

结构层	输出尺寸	卷积核	参数量
Conv1	64×62	3×1,32	96
Pooling1	64×31	1×2	
Conv2	64×31	3×1,32	3 072
Pooling2	64×16	1×2	
Conv3	64×16	3×1,64	6 144
Pooling3	64×8	1×2	
Conv4	64×8	3×1,128	24 576
Pooling4	64×4	1×2	

续表 1

结构层	输出尺寸	卷积核	参数量
Conv5	64×4	3×1,128	49 152
Pooling5	64×2	1×2	
Conv6	64×2	3×1,128	49 152
Pooling6	64×1	1×2	
Conv7	64×1	3×1,128	49 152
Avg Pooling	64		
Full connection	64		64

该模型结构共有 7 层,每层包含一个卷积层与一个池化层,卷积层的输出通过 ReLU 函数进行激活,并在层与层之间加入残差连接[27],最后通过全局池化提取最终特征,并通过 sigmoid 预测检测结果。该结构最大的特点是采用在频率维度卷积及时间维度池化,具体设置为采用 3×1 卷积核和 1×2 池化。如此设置一方面最大化降低模型容量,极大减少过拟合的风险,降低模型对数据量的依赖性,另一方面,又与时频图的特征分布特点高度契合,将训练参数分配到更合理的地方,从而用更有效的特征来训练更紧凑的参数。

深度学习模型的性能对数据有极高的依赖性,以原始音频信号作为网络的输入数据,其特征分布过于稀疏,极大地提高了神经网络提取有效特征的难度。另一方面,重录设备会在原语音信号的频域上引入变化[21,23,26,28],此种变化可以作为区分重录语音及原始语音的重要依据。为此,文中的网络输入数据采用语音的时频图。时频图由短时傅里叶变换(short-time Fourier transform,STFT)生成,相对于直接输入语音数据,时频图对于重录设备引入的特征信息有相对密集的分布,更有利于神经网络特征提取,从而加快训练,提高精度。

语音重录包含三个过程:语音经过播放器播放,经过空气传播,再由录音设备录制。重录导致语音数据一定程度的失真,此失真包括幅度失真和时间轴上的线性伸缩,主要由播放时的 DA 变换与录制时的 AD 变换采用的设备、录制环境及录制距离等因素造成。幅度失真可以表示为能量变化和一个叠加噪声,线性伸缩的程度与使用的硬件如声卡性能及所采用的采样率有关。失真模型可表示为:

$$y(t) = \lambda x(\frac{t}{\alpha}) + \eta \tag{1}$$

其中,$y(t)$ 是重录语音;$x(t)$ 是原始语音;λ 是幅值变换因子;η 是叠加噪声。

对应的频域变化如式 2 所示。

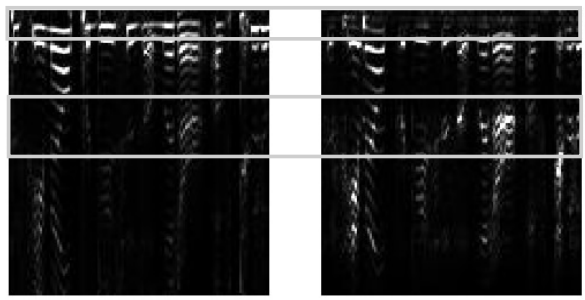
$$Y(j\omega) = \lambda \partial X(j\alpha\omega) + N(j\omega) \tag{2}$$

其中,∂ 是时间轴线性伸缩因子;$Y(j\omega)$、$X(j\omega)$、$N(j\omega)$ 分别为 $y(t)$、$x(t)$、η 的频域表示。

对于固定的录音设备,其特征是非常稳定的,即 λ、α 是常数,而叠加噪声与录制环境、录制距离及录制设备 AD 转换有关。

对于训练数据,该实验均为安静环境下录制,避免引入无关的环境噪声,因为环境噪声有很大的随机性,且深度学习作为数据驱动的技术,在数据中加入环境噪声会使模型在训练过程中将是否含有环境噪声作为检测的依据。这对于实验是非常不利的,模型检测的依据应该是与设备相关的、稳定存在的特征。在实验中叠加噪声 η 主要与不同的录制环境,以及不同的录制设备有关,因此对于特定录制设备、特定录制距离下 $H(j\omega)$ 的分布也是特定的。为了验证模型对含有环境噪声的录制语音检测的鲁棒性,文中也对含有环境噪声的录制语音进行了检测。

综上分析,对于文中采用的时频图,作为检测是否为重录语音的特征,其分布特点在相邻语音帧之间具有独立性并且在特定频段又具有一致性。



原始语音　　重录语音

图 2　时频图

如图 2 所示,左侧为原始语音时频图,右侧为一种场景下的重录语音时频图,方框内区域可以直观地看出,重录语音引入的变化在某些频段较为明显。频率分辨率的大小是影响特征提取的最关键因素,这是由短时傅里叶变换中窗函数长度决定的,窗长度越大,频率分辨率越高,特征表现越明显。因此,在传统信号处理方法检测重录语音特征时,为了提取充分的特征,往往需要很长的语音段,这极大地限制了其适用范围。文中采用 0.2 秒语音段作为实验数据,短时傅里叶变换采用 126 长度汉宁(Hanning)窗,步长为 50,时频图

的尺寸为(64×62)。模型适用于绝大多数应用场景,并且实验结果证明具有很好的效果。

文中在频率维度进行卷积,同时在时间维度进行池化。只在频率维度进行卷积(3×1),不考虑时间维度的相关性,能极大地减少卷积核参数量,使得模型有更强的抗过拟合能力,减少对数据量的过度依赖,同时在训练过程中由于卷积核的参数共享,时间维度具有同分布设备的特征信息重复训练卷积核参数,可以使训练更加充分。池化层采用时间维度的池化(1×2),频率维度不进行池化。池化能减少特征的维度,加快网络的计算,并且使网络结构对数据特征的伸缩、变形有更强的鲁棒性,但池化在减少数据维度的同时也会丢失很多的特征信息,对于时频图,特征分布不存在伸缩与变形,只在时间维度池化,既减少了特征维度,同时又不会导致频率维度特征的丢失,这在文中卷积神经网络训练过程中极为重要。通过多层卷积与池化计算,特征维度最终变为一维,长度与时频图频率相同。

3 实验结果

3.1 实验设置

实验采用 0.2 秒时长语音段,原始语音库由 300 个说话人共 100 分钟的语音,每人语音时长为 20 秒,均是经过裁剪处理,不包含明显的静音片段,抽样频率 16 kHz,量化精度 16 bits。已有的研究报道均不考虑训练样本和测试样本在不同场景下的录制,而这不符合实际场景。为此,该实验语音库用不同的录音设备及在不同的录音距离下重录,以测试算法的通用性。随机抽选 50 位发言人的语音作为测试数据,其余 250 人的语音用于训练,避免同一位发言者的录音出现在不同数据集,保证训练数据与测试数据的独立性。

具体录制过程如下:对于训练集,在安静环境下由不同距离和设备组合对原始语音库重录 4 次,由此获得 4 个重录语音库,它们分别包含 25 000 段语音。原始语音通过手提电脑联想 Y40-70AT-IFI 播放;重录设备是手提电脑戴尔(Inspiron)灵越 14(Ins14VD-258)和智能手机小米 2S。4 次录制的情况如表 2 中编号为 t 的数据。

对于测试数据,采用与训练集相同的录制设置。为了验证模型对具有环境随机噪声干扰的语音的鲁棒性,分别在室内安静环境与有一定随机噪声的室内环境下录制,测试集共包含 8 个语音库,每个语音库包含该库录制模式下共 25 000 条测试语音,如表 2 中编号为 s 的数据。

表 2 语音录制情况

录制次数	录制设备	录制距离/cm	录制环境
t_1	手提电脑	20	安静
t_2	智能手机	20	安静
t_3	智能手机	40	安静
t_4	智能手机	60	安静
s_1	手提电脑	20	安静
s_2	智能手机	20	安静
s_3	智能手机	40	安静
s_4	智能手机	60	安静
s_5	手提电脑	20	噪声
s_6	智能手机	20	噪声
s_7	智能手机	40	噪声
s_8	智能手机	60	噪声

数据输入网络之前需要进行预处理,过程如下:对每个语音段进行短时傅里叶变换,语音采样率为 16 kHz,量化精度 16 bit,采用 126 长度汉宁(Hanning)窗,步长为 50。全部数据在输入网络前要经过归一化处理,先计算整个训练集数据的均值 μ 与标准差 σ,然后对数据样本 x' 进行减均值,除以标准差来进行归一化,最后得到经过预处理的数据 x。

$$x = \frac{x' - \mu}{\sigma} \tag{3}$$

3.2 训练网络

文中网络误差函数为交叉熵损失函数,采用 Adam 优化算法进行训练,初始学习率设置为 0.001,并在训练过程中动态调整学习率,每训练 10 000 次将学习率减小一倍,每次训练批量大小为 32。为了在训练过程中监督训练效果,从训练数据中随机选取 2 000 条数据用于验证,通过对比训练数据损失函数与验证数据损失函数,为损失函数加入正则化项并设置正则化系数为 0.000 1 能有效防止过拟合。

表 3　超参数

(β_1、β_2 分别为 Adam 优化器参数)

学习率	β_1	β_2	正则化系数	最小批次	训练次数
10^{-3}	0.9	0.999	10^{-4}	32	50 000

表 3 列出了训练过程中的一些重要的超参数设置,在训练过程中不断监测训练损失与验证损失,并挑选训练损失小并且与验证损失较为接近时的模型作为测试模型。在该超参数设置下网络在训练过程中能够快速收敛,并且最终取模型得到相当高的精确度。

3.3　测试结果

重录语音检测涉及多个影响因子,包括录制设备、录制距离以及录制环境等。为了验证不同的影响因子对于网络的影响,分别对不同的录制语音进行实验,训练多个模型并分别对测试数据进行测试,以此来分析各因素对网络检测率的影响,并通过实验结果的分析,从而提出最终的训练模型。具体内容如下:分别以原始语音 t 为正样本与不同重录语音作为负样本的组合来训练网络,以 t_1 为负样本训练得到模型 M_1,以 t_2 为负样本训练得到模型 M_2,以 t_3 为负样本训练得到模型 M_3,以 t_4 为负样本训练得到模型 M_4,分别从 t_1、t_2、t_3、t_4 数据集中等比例采样组成训练集负样本训练得到模型 M_5 作为最终的模型。测试结果如下:

(1)验证录制设备对网络的影响。

分别以模型 M_1、M_2 对测试数据 s、s_1、s_2 进行测试,其中 s 为原始语音的测试数据。在录制环境与录制距离相同条件下研究录制设备对模型的影响,测试结果如表 4 所示。

表 4　不同录制设备的测试结果

模型	s	s_1	s_2
M_1	99.46	99.90	93.70
M_2	99.92	31.46	99.90

测试结果表明,在录制距离与录制环境条件相同时,重录设备对于模型的影响较大,相同设备的重录语音训练的模型对同设备下的重录语音有较高的检测率,而对其他设备的重录语音检测率不理想。如表 4 所示,使用 t_1 数据训练的模型对于 s_2 数据检测率为 93.7%,低于其相同录制条件下的测试数据,而对于 t_2 数据训练的模型对 s_1 数据的检测率低至 31.4%,甚至低于随机猜测。并且,不同设备对于模型的影响大小也不相同,由表 4 中可知采用电脑重录语音训练的模型比手机重录语音训练的模型有更好的鲁棒性。

(2)验证录制距离对网络的影响。

以模型 M_2、M_3、M_4 分别对测试数据 s、s_2、s_3 进行测试,在录制设备与录制环境相同条件下研究录制距离对模型的影响,测试结果如表 5 所示。

表 5　不同录制距离的测试结果

模型	s	s_2	s_3	s_4
M_2	99.92	99.90	30.12	2.74
M_3	99.90	99.87	99.96	96.44
M_4	99.86	76.64	98.08	99.96

测试结果表明,不同的录制距离对网络影响较大,相同录制距离的重录语音训练的模型对相同录制距离的重录语音检测率较高,均能达到 99.9% 以上,而录制距离不同的情况下检测率则较低,并且随着距离的差距增加检测率不断下降,20 cm 的重录语音训练的模型,对于 40 cm 的重录语音检测率仅为 30.12%,对于 60 cm 的重录语音检测几乎全部错误。对于 60 cm 的重录语音训练的模型也有相似的结果。由表 5 实验结果可知,40 cm 距离的重录语音对于 20 cm、60 cm 的重录语音有不错的检测率,原因是其特征与这两种录制距离的重录语音有更多的相似性,因此模型能够通过这些特征来进行判定,而随着距离差距的增大,特征变化更大,对应的模型只能识别该距离下的重录语音特征。

(3)验证录制环境对网络的影响。

分别以模型 M_1、M_2、M_3、M_4 测试其对应录制条件下安静与有噪声的重录语音,测试结果如表 6 所示。

测试结果表明,模型对于有少量随机噪声的重录语音检测率略低于安静环境下重录语音,但是影响有限。随机噪声会为重录语音引入新的特征,这对模型的检测会有一定的干扰,但同时,无论是否含有噪声,重录语音对设备和录制距离的特征是比较稳定的,并且占据极大的比例,这些特征是区分原始语音与重录语音更重要、更稳定的特征。由表 6 测试结果可知,文中提出的网络结构对于不同录制环境下的重录语音都有较好的检测率,高达 99.8% 以上,表明该网络结构对重录语音中的随机环境噪声有良好的鲁棒性。

表 6　不同录制环境的测试结果

模型	安静/%	有噪声/%
M_1	99.90	99.83
M_2	99.90	99.81
M_3	99.96	99.86
M_4	99.96	99.84

以上实验表明,录制设备、录制距离、录制环境等影响因素对于模型都有不同程度的影响,单一条件下的重录语音所训练的模型对不同条件下的重录语音鲁棒性较低。因此为了提高模型对不同录制设备、录制

距离以及录制环境下的重录语音的检测能力,需要对训练数据数据集进行合理的设置,训练集应更多地包含各种不同条件下的重录语音数据,这样网络结构才能学习更多不同录制设备、录制距离以及录制环境下的重录语音特征,从而提高模型的识别能力,提高对于各种场景下的重录语音的鲁棒性。

(4)多场景训练数据组合训练模型。

为了提高模型对不同场景下的重录语音的检测能力,采取多录音条件下的数据组成训练集,对模型进行训练,具体内容如下:分别从 t_1、t_2、t_3、t_4 数据库中等比例各随机抽取四分之一数据并与原始语音共同组成训练集,然后使用该训练集对模型进行训练得到模型 M_5,并分别对测试集数据进行测试。测试结果如表 7 所示。

表 7 三种录制条件综合下的测试结果

测试数据	实验结果/%
s	99.90
s_1	99.84
s_2	99.91
s_3	99.93
s_4	99.96
s_5	99.80
s_6	99.83
s_7	99.85
s_8	99.89

由测试结果可知,采用多条件重录语音所组成的训练集能极大提高模型的鲁棒性,对不同情况下的重录语音测试精确度都比较高,均能达到 99.8% 以上。采用多条件下的重录语音组成训练集,极大地丰富了训练数据的特征信息,通过充分训练,模型能够提取不同录制设备、录制距离等特征信息,同时也使原始语音与重录语音的特征更有分辨性。测试结果表明,在采用多条件下重录语音进行训练的模型对各种不同录制设备、录制距离以及录制环境下的重录语音都有良好的检测率,此条件下训练的模型具有更好的鲁棒性。

综上所述,不同的录制设备、录制距离以及录制环境都会对模型的检测造成不同程度的影响,采用单一条件下的重录语音训练的模型不具有通用性,泛化能力不足。因此文中网络结构采用不同录制条件下的重录语音的组合数据集进行训练,结果表明该网络结构有良好的鲁棒性,对于不同录制设备、录制距离以及录制环境下重录语音都具有极高的检测性能,表明提出的卷积神经网络能很好地解决重录语音攻击的检测问题,并且具有对较短语音段的检测能力。

4 结束语

之前对于重录语音的检测,更多集中于传统的信号处理方法,特征提取的算法有很大的局限性,算法复杂,同时为了提取充分的特征,对语音段的长度有较大的要求,这对算法的实用性是很大的限制。文中提出的卷积神经网络结构可以在极短语音段上提取充分的特征信息,依据语音信号的时频特征进行特殊设计,运用特殊的卷积核设置,与时频图的特征分布特点高度契合,并且模型参数量较少,大大降低了模型过拟合风险。同时对录制设备、录制距离以及录制环境等影响因子进行了实验研究,结果表明通过增加训练集数据的丰富性能极大地提高模型的鲁棒性,通过采用多场景下的重录语音混合数据进行训练,模型取得了最好的效果。为了验证该算法地性能以及通用性,网络分别对不同录制设备、不同录制距离及不同录制环境下的重录语音进行测试,其结果的精确度可达 99.8% 以上。实验结果表明,该网络能够有效地学习到标准信号处理无法解决的强大的特征表示,并能获得较高的识别精度;该卷积神经网络模型对于不同录制场景和设备的重录语音检测具有通用性。

参考文献:

[1] KAJAREKAR S S, BRATT H, SHRIBERG E, et al. A study of intentional voice modifications for evading automatic speaker recognition[C]//IEEE Odyssey - the speaker and language recognition workshop. San Juan: IEEE, 2006: 1-6.

[2] KÜNZEL H J, GONZALEZ-RODRIGUEZ J, ORTEGA-GARCÍA J. Effect of voice disguise on the performance of a forensic automatic speaker recognition system[C]//ODYSSEY04 - the speaker and language recognition workshop. [s. l.]: IEEE, 2004.

[3] ALEGRE F, JANICKI A, EVANS N. Re-assessing the threat of replay spoofing attacks against automatic speaker verification[C]//2014 international conference of the biometrics special interest group (BIOSIG). Darmstadt: IEEE, 2014: 1-6.

[4] SHCHEMELININ V, TOPCHINA M, SIMONCHIK K. Vulnerability of voice verification systems to spoofing attacks by TTS voices based on automatically labeled telephone speech [C]//International conference on speech and computer. Novi Sad, Serbia: Springer, 2014: 475-481.

[5] ANJUM Z K, SWAMY R K. Spoofing and countermeasures for speaker verification: a review[C]//2017 international conference on wireless communications, signal processing and networking (WiSPNET). [s. l.]: IEEE, 2017: 467-471.

[6] HANILCI C. Speaker verification anti-spoofing using linear prediction residual phase features[C]//2017 25th European signal processing conference (EUSIPCO). Piscataway, NJ: IEEE, 2017: 96-100.

[7] KAMBLE M R, PATIL H A. Novel energy separation based

instantaneous frequency features for spoof speech detection [C]//2017 25th European signal processing conference (EUSIPCO). Piscataway, NJ: IEEE, 2017: 106-110.

[8] MUCKENHIRN H, KORSHUNOV P, MAGIMAI-DOSS M, et al. Long-term spectral statistics for voice presentation attack detection [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2017, 25(11): 2098-2111.

[9] JANICKI A. Spoofing countermeasure based on analysis of linear prediction error [C]//16th annual conference of the international speech communication association. Red Hook, NY: Curran Associates Inc., 2016: 2077-2081.

[10] ALAM J, KENNY P. Spoofing detection employing infinite impulse response—constant Q transform-based feature representations [C]//2017 25th European signal processing conference (EUSIPCO). Piscataway, NJ: IEEE, 2017: 101-105.

[11] DINKEL H, QIAN Y, YU K. Small-footprint convolutional neural network for spoofing detection [C]//2017 international joint conference on neural networks (IJCNN). Anchorage, AK: IEEE, 2017: 3086-3091.

[12] LIANG H, LIN X, ZHANG Q, et al. Recognition of spoofed voice using convolutional neural networks [C]//2017 IEEE global conference on signal and information processing (GlobalSIP). [s. l.]: IEEE, 2017: 293-297.

[13] SAHIDULLAH M, THOMSEN D A L, HAUTAMAKI R G, et al. Robust voice liveness detection and speaker verification using throat microphones [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 44-56.

[14] GONCALVES A R, KORSHUNOV P, VIOLATO R P V, et al. On the generalization of fused systems in voice presentation attack detection [C]//16th international conference of the biometrics special interest group (BIOSIG). [s. l.]: IEEE, 2017.

[15] QIAN Y, CHEN N, DINKEL H, et al. Deep feature engineering for noise robust spoofing detection [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(10): 1942-1955.

[16] ARASHLOO S R, KITTLER J, CHRISTMAS W. An anomaly detection approach to face spoofing detection: a new formulation and evaluation protocol [J]. IEEE Access, 2017, 5: 13868-13882.

[17] LEE K, PARK C, KIM N, et al. Accelerating recurrent neural network language model based online speech recognition system [C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). [s. l.]: IEEE, 2018: 5904-5908.

[18] 欧国振,孙林慧,薛海双. 基于重组超矢量的 GMM-SVM 说话人辨认系统[J]. 计算机技术与发展,2017,27(7):51-56.

[19] 林舒都,邵　曦. 基于 i-vector 和深度学习的说话人识别[J]. 计算机技术与发展,2017,27(6):66-71.

[20] 李燕萍,陶定元,林　乐. 基于 DTW 模型补偿的伪装语音说话人识别研究[J]. 计算机技术与发展,2017,27(1):93-96.

[21] WICKRAMASINGHE B, IRTZA S, AMBIKAIRAJAH E, et al. Frequency domain linear prediction features for replay spoofing attack detection [C]//19th annual conference of the international speech communication association. Red Hook, NY: Curran Associates Inc., 2019: 661-665.

[22] SINGH M, MISHRA J, PATI D. Usefulness of linear prediction residual signal for development of replay attacks detection system [C]//2017 twenty-third national conference on communications (NCC). [s. l.]: IEEE, 2017: 1-4.

[23] SRISKANDARAJA K, SUTHOKUMAR G, SETHU V, et al. Investigating the use of scattering coefficients for replay attack detection [C]//2017 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC). [s. l.]: IEEE, 2017: 1195-1198.

[24] YOUNG S. The HTK hidden Markov model toolkit: design and philosophy [R]. Cambridge: Cambridge University, 1993.

[25] VEDALDI A, FULKERSON B. Vlfeat: an open and portable library of computer vision algorithms [C]//Proceedings of ACM multimedia. New York, NY: ACM, 2010.

[26] LIN X, LIU J, KANG X. Audio recapture detection with convolutional neural networks [J]. IEEE Transactions on Multimedia, 2016, 18(8): 1480-1487.

[27] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2016: 770-778.

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