引用格式:谢元坤,程皓楠,叶龙. 深度伪造音频检测综述[J]. 中国传媒大学学报(自然科学版),2024,31(03):26-33.
文章编号:1673-4793(2024)03-0026-08

深度伪造音频检测综述

谢元坤,程皓楠,叶龙*
(中国传媒大学,北京 100024)

摘要:随着生成式人工智能技术的快速普及和发展,社交媒体领域充斥着大量由语音合成、语音转换等技术生成的深度伪造音频。这些高自然度的深度伪造音频为真伪媒体内容分辨带来了巨大挑战。为了解决这一问题,国内外已经组织了多样化深度伪造音频检测挑战赛,以促进音频反欺骗领域的发展。区别于已有综述局限于音频真伪二分类,本文跨越传统二分类,对深度伪造音频检测领域的相关工作做出了全面的总结。即将深度伪造音频检测领域分为三个子领域:全局伪造音频检测、局部伪造音频定位、深度伪造音频溯源,分别对三个子领域现有的数据集领域问题、解决方法进行了梳理和总结。最后,提出了深度伪造音频检测领域可能面临的挑战,对下一阶段的研究进行展望,期望为未来研究人员提供可靠参考。

关键词:深度伪造音频检测;全局检测;局部定位;伪造溯源
中图分类号:TP393.2 **文献标识码**:A

Audio deepfake detection: a survey

XIE Yuankun, CHENG Haonan, YE Long*
(Communication University of China, Beijing 100024, China)

Abstract: With the rapid development of generative artificial intelligence technology, social media platforms have become inundated with a plethora of deepfake audio synthesized using techniques such as speech synthesis and voice conversion. These deepfake audios, capable of producing highly natural and realistic voices, pose significant threats. To address this issue, numerous deepfake audio detection challenges have been organized globally, aiming to foster the development of the audio anti-spoofing field. Distinguishing from existing surveys which limited to the binary classification of whole audio authenticity, this article transcends traditional binary classification and provides a comprehensive summary of audio deepfake detection. Specifically, this article divides the domain of audio deepfake detection into three sub-domains: global deepfake audio detection, local deepfake audio localization, and deepfake audio source tracing, systematically reviewing and summarizing existing datasets, domain issues, and solution approaches in each sub-domain. Finally, this paper outlines the potential challenges facing the field of deepfake audio detection and offers prospects for future research, aiming to provide reliable reference for future researchers.

Keywords: audio deepfake detection; global deepfake audio detection; local deepfake audio localization; deepfake audio source tracing

基金项目:国家自然科学基金(62201524);国家重点研发计划项目(2021YFF0900504)
作者简介(*为通讯作者):谢元坤(1994-),男,博士研究生,主要从事深度伪造音频检测研究。Email:xieyuankun@cuc.edu.cn;程皓楠(1994-),女,博士,副研究员,主要从事音频合成、音频处理研究。Email:haonancheng@cuc.deu.cn;叶龙(1983-),男,博士,教授,博导,主要从事智能音视频处理技术、虚拟现实技术研究。Email:yelong@cuc.edu.cn

1 引言

近年来，随着深度学习技术的快速发展，其在机器听觉领域的应用也日趋广泛。众多高效且精确的模型陆续超越了传统技术，极大提升了媒体内容的生产与创作效率。特别是在音频合成领域如语音合成（Text to Speech，TTS）和语音转换（Voice Conversion，VC），它们能够创造出欺骗人耳和人机交互设备的目标说话人语音。这些技术在智能设备的语音交互中有着重要应用，能够提供高度拟人化和流畅自然的交互功能，广泛用于开发朗读听书、无障碍信息播报等语音服务，极大地丰富了用户的交互体验。此外，这些音频深度伪造技术还能推动文娱产业的创新发展，在影视、游戏制作中虚拟角色声音模拟、个性化音频内容创作等领域被广泛应用。

然而，音频深度伪造技术的滥用潜在风险不容忽视。如图1所示，深度伪造音频所危害的对象一方面是人类听觉系统，一方面是机器听觉系统，这对全球民生、经济、政治与社会安全构成严重威胁。因此，目前亟需自动化且具有高准确性的深度伪造音频检测方法，以应对其产生的威胁。

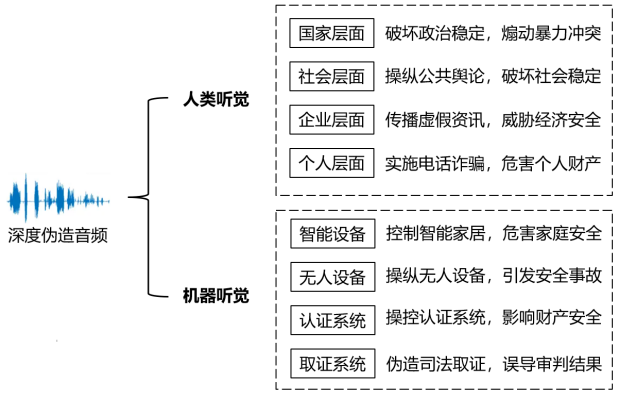

图1 深度伪造音频所产生的危害

为了有效防御恶意滥用的伪造语音对人类和机器造成的上述危害，近年来针对伪造语音的检测技术在同步发展，相继有学者对不同类型的伪造语音检测展开了多角度的研究。目前的深度伪造音频检测通常分为三项子任务，即全局伪造音频检测，局部伪造音频定位，深度伪造音频溯源。对深度伪造音频检测的早期研究基本围绕着全局伪造音频检测而展开。由英国爱丁堡大学等多个研究机构共同发起的ASVspoof系列挑战赛[1]自2015年起持续引领了伪造语音检测领域的发展。自ASVspoof 2019[2]举办以来出现了一系列先进的虚假音频解决方案。但研究者们很快发现这些解决方案通常泛化性不够，因此ASVspoof 2021专注于解决泛化性问题，旨在利用有限数据解决复杂场景的语音鉴伪问题。然而，ASVspoof系列比赛仅含有英文合成音频，并没有关于中文以及非二分类情形的讨论。最近，中科院自动所等多个研究机构举办了ADD挑战赛[3]，这是首个关于中文深度虚假音频检测的比赛。该比赛的三条赛道分别对应着深度伪造音频检测的三个子任务。

对深度伪造音频检测方法，常见的深度虚假音频检测的模型结构如图2所示，由前端特征提取和后端主干网络构成。对前端特征目前主要分为四类：手工特征，身份特征，原始输入，预训练特征。主干网络目前由卷积网络为代表的如轻量卷积网络LCNN[4]，以及时频域融合主干网络如AASIST[5]等所构成。通过使用Softmax或Sigmoid函数对主干网络输出的预测，最终完成音频真假二分类判断。

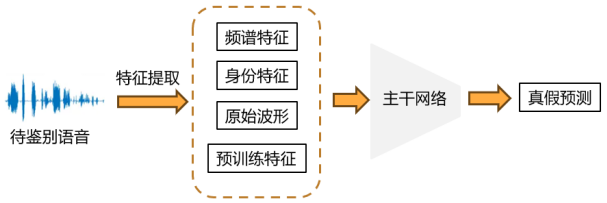

图2 深度伪造音频检测框架图

然而，在实际应用场景中，目前的主流二分类深度伪造音频检测模型在部分伪造和多分类溯源任务中表现并不理想。因此，目前深度伪造音频检测的研究主要聚焦于三个关键技术领域：全局伪造音频检测、局部伪造音频定位以及深度伪造音频溯源。

（1）全局深度伪造音频检测（二分类）：该技术涉及对整个音频信号进行综合性分析和判定，以确定音频中是否含有伪造成分。这需要综合考量音频的全局特征、时域和频域信息，并构建相应的模型以捕捉伪造音频的特征模式和异常行为。

（2）局部深度伪造音频定位（帧级别二分类）：除全局检测外，局部伪造音频定位技术尚处于初步研究阶段。该技术主要关注于定位音频信号中可能包含伪造成分的特定片段或局部区域。这些区域可能展示出操纵痕迹、接缝痕迹或其他形式的不一致性特征。

（3）深度伪造音频溯源（多分类）：深度伪造音频溯源旨在追踪伪造音频的起源或来源，从而提高鉴伪系统的可解释性。 该任务在全局深度伪造音频检测

之上提出了新的要求,旨在区分深度伪造的同时判断其采用的深度伪造算法。通过研究溯源技术,可以更深入地了解虚假音频的生成机制,并为检测和分析虚假音频提供更多的证据和信息支持。

本文从音频鉴伪的数据集出发,系统性地介绍当前深度伪造音频检测、伪造音频定位、伪造音频溯源三个子领域存在的关键性问题及目前的主流解决方法,并对整个深度伪造音频检测领域进行总结及展望。

2 深度伪造音频数据集

在数据集方面,本文对目前已存在的深度伪造音频数据集包括全局、局部、溯源三种任务进行梳理。其中,全局伪造音频检测数据集通常包括真实语音和通过TTS/VC算法生成的深度伪造语音,标签分为真假两类。局部伪造音频定位数据集通常包括真实音频和部分伪造的深度伪造音频,标签为帧级别真伪标签。深度伪造溯源数据集包括真实音频,和不同TTS/VC算法合成的音频,标签为伪造音频所使用的算法。现有数据集的详细信息如表1所示。

2.1 全局深度伪造音频数据集

ASVspoof2019LA(19LA)[2]:该数据集为深度伪造音频领域的基准数据集。该数据集真实源域来自于VCTK[6],伪造语音包含有19种TTS/VC算法。具体而言,其训练集和验证集包含有6种伪造方式(A01-A06),测试集包含全部不可见的13种伪造方式(A07-A19)。这种训练集和测试集的划分方式要求深度伪造音频检测模型具有一定的泛化性,能够对不可见的伪造方式进行有效鉴别。

ASVspoof2021LA(21LA)[7]:该数据集以19LA为基准所建立,其目标旨在检测真实环境下音频的真伪。该数据集使用19LA的测试集模拟了不同的真实通讯环境包括多样性的编解码方式、传输通道、比特率、采样率等场景。值得一提的是该数据集全部为测试集,在21LA挑战赛要求中,并未给定特定的训练集,仅让使用19LA的训练、验证协议进行模型训练。这要求了参赛者的鉴伪模型具有一定的鲁棒性,能够在不同的模拟环境中正确检测真实和伪造语音。

表1 深度伪造音频检测数据集总览

数据集	年份	开源	语言	类型	种类	条件	采样率	真实	伪造
全局伪造音频数据集									
ASVspoof 2019LA	2019	是	英语	语音	19	干净	16kHz	10256	90192
ASVspoof 2021LA	2021	是	英语	语音	19	编码	混合	14816	133360
ASVspoof 2021DF	2021	是	英语	语音	100+	编码	混合	14869	519059
WaveFake	2021	是	英/日	语音	7	干净	16kHz	0	117985
ITW	2022	是	英语	语音	未知	噪声	16kHz	19963	11816
TIMIT-TTS	2022	是	英语	语音	12	噪/码	16kHz	0	5160
ADD2022T1.2	2022	否	中文	语音	未知	噪声	16kHz	36953	123932
Latin-American	2022	是	西班牙	语音	6	干净	48kHz	22816	758000
CFAD	2023	是	中文	语音	12	噪/码	16kHz	38600	77200
ADD2023T1.2	2023	否	中文	语音	未知	噪/码	16kHz	172819	113042
FSD	2024	否	中文	歌声	5	噪声	48kHz	13157	23919
Singfake	2024	否	多语	歌声	未知	噪/码	混合	15488	11466
SceneFake	2024	是	英语	语音	10	噪声	16kHz	19838	64642
局部伪造音频数据集									
ASVspoof 2019PS	2022	是	英语	语音	9	干净	16kHz	12483	108978
HAD	2022	是	中文	语音	未知	干净	44.1kHz	53612	753612
Psynd	2022	否	英语	语音	1	编码	24kHz	30	2371
ADD2022T2	2022	否	中文	语音	未知	干净	16kHz	23897	127414
ADD2023T2	2023	否	中文	语音	未知	噪/码	16kHz	55468	65449
伪造音频溯源数据集									
VFD	2022	否	中文	语音	8	干净	混合	0	63200
LibriSeVoc	2023	是	英文	语音	6	干净	24kHz	13201	79206
ADD2023T3	2023	否	中文	语音	7	噪/码	混合	14907	95383

ASVspoof2021DF（21DF）[7]：该数据集为ASVspoof2021比赛新创建的赛道，该赛道旨在有效鉴别Deepfake语音。21DF数据集的真实源域来自三个不同域：VCTK，VCC2018[8]和VCC2020[9]，伪造语音由超过100种的不同TTS/VC算法进行合成而来。该数据集还包含不同的压缩方式，如mp3、m4a、ogg等。这要求鉴伪模型具有一定的泛化性，能够检测不同源域生成的语音。

WaveFake[10]：该数据集是双语数据集，真实语音包含英语源域为LJspeech、日语源域为JUST。该数据集使用6种最新不同的基于GAN的声码器进行合成，作者在这篇文章中还探索了使用单伪造语音进行训练，在多种伪造方式上的测试效果，并探索了鉴伪模型在跨语言的场景下的鉴伪性能。

In-the-Wild（ITW）[11]：ITW数据集是从不同社交媒体网站上下载的原始英语数据集。该数据集包含有公众人物及政治家在不同场景下的演讲，这涉及丰富的真实场景如不同的户外场景、不同的流媒体录制等。由于ITW的源域多样性（不同的录制设备、混响、噪声等）使得ITW在音频鉴伪领域通常为测试泛化性的数据集。

TIMIT-TTS[12]：TIMIT-TTS是一个合成语音数据集，包含12种最先进的TTS算法。所有选定的TTS算法都是频谱图生成器，这些算法主要保留了生成语音的差异，主要归因于声码器。为了降低音频质量并隐藏一些伪影，采用了各种后处理技术，包括添加高斯噪声、应用MP3编解码器和添加混响压缩等。

ADD2022T1.2[3]，**ADD2023T1.2**[13]：该数据集为ADD挑战赛中Track1深度伪造音频合成攻防的防御赛道。该数据集在生成数据中使用不同的真实场景中的噪声以及背景音乐以模拟真实鉴伪场景，该数据集仅提供给参赛方，目前尚未开源。

Chinese Fake Audio Detection（CFAD）[14]：CFAD是第一个开源的中文伪造音频数据集，真实源域数据为6种不同的常用中文数据集，合成数据包括12种TTS/VC的合成方法，并且使用不同的噪声和压缩方式对数据集进行了扩充，以模拟真实场景。

FSD[15]，**Singfake**[16]：不同于以上深度伪造语音的数据集，该两种数据集面向场景为检测深度伪造的歌声。其中FSD为第一个中文伪造歌声数据集，该数据集使用目前主流的4种歌声转换方法和1种歌声合成方法，制造了歌声伪造数据集。Singfake不同于FSD，其面向野外场景，其数据集来源为互联网中大量存在的真实歌声以及伪造歌声，这包含了复杂的真实源域以及不同的伪造类型。值得注意的是上述歌声鉴伪的最佳解决方案为对音源分离后的干声部分进行语音鉴伪。

SceneFake[17]：该数据集为语音伪造数据集，但不同于上述已有伪造数据集，该数据集的伪造类型为语音场景伪造，即篡改语音说话人所在的场景。该数据集使用干净环境下的19LA的数据集结合DCASE 2022挑战赛中的声学场景进行场景的伪造。

2.2 局部伪造音频定位数据集

ASVspoof2019PS（19PS）[18]：该数据集为第一个局部伪造音频定位数据集。其面向场景为局部深度伪造，即将深度伪造音频的片段插入到真实音频中以实现语句的语义篡改。该数据集使用19LA的数据集为基准，将不同的伪造语音片段插入真实语音中。该数据集文章对数据集进行20ms-640 ms总共5个不同分辨率的标注并提出了多分辨率的帧级别定位解决方案。

Half-Truth（HAD）[19]：HAD数据集包含部分伪造的语音，其中一些话语中的几个词语使用TTS生成技术进行了修改。该数据集旨在评估反欺骗方法并定位部分伪造的音频。被替换的关键词包括人物、地点、组织和时间等实体。

Partial Synthetic Detection（Psynd）[20]：该数据集中的数据样本是真实话语，其中注入了与目标发言者紧密相似的合成语音片段，这些片段是由多说话人TTS算法生成的。在训练、验证和初步测试数据中，每个话语都包含一个单独的伪造片段。数据集包含有特殊情况的测试集涉及完全伪造、完全真实和多个伪造片段。

ADD2022T2[3]，**ADD2023T2**[13]：该数据集为ADD挑战赛中Track2中的赛道数据集，其面向真实场景包含有不同噪声和压缩格式干扰。在ADD2022T2中仅要求对真实和部分伪造音频进行二分类，而在ADD2023T3中要求对部分伪造语音能够进行10ms级别的细粒度定位。

2.3 伪造音频溯源数据集

Vocoder Fingerprints Detection（VFD）[21]：该数据集是第一个探索使用不同声码器是否会留有指纹的工作，这种基于声码器的指纹为溯源的多分类模型提供检测基础。VFD基于AISHELL3使用8种伪造方法进行伪造，实验结果显示仅使用手工特征和Resnet的组合能够达到较高的溯源准确率，这表明不同声码器确实会留下较明显的指纹。

LibriSeVoc[22]:该数据集使用6种TTS/VC常用的声码器合成了总计146小时的数据集,其真实源域为TTS领域常用的LibriTTS[23]数据集,包含有多说话人。该文章认为不同的声码器具有不同位置的伪影,使用所提出的二分类和声码器分类器损失完成真假溯源的同时进行声码器的多分类任务。

ADD2023T3[13]:该数据集为ADD挑战赛的Track3所提供的数据集,目标任务为区分真伪的同时对伪造语音合成算法进行多分类溯源。更具有挑战性的是,在训练和验证集中只提供了6种伪造方式,而在测试集中有第7种不可见的伪造方式,这要求溯源模型在多分类的同时具有一定的分布外检测(Out-of-distribution,OOD)能力。

3 深度伪造音频检测方法

3.1 全局深度伪造音频检测

全局深度伪造音频检测主要考虑语音信号、声纹特征和频谱分布等生物信息的差异特征进行鉴别。由于人类听觉系统对相位相对不敏感,传统的声码器通常不会重建语音的相位信息,因此导致真实语音和合成语音之间相位谱的差异,可用于检测伪造语音。随着深度学习技术的发展,目前基于深度学习的伪造语音检测已成为主流方法。

深度伪造语音检测方法围绕特征和主干网络进行探究。特征提取方法包括频谱特征,身份特征,原始波形和预训练特征。频谱特征通常使用滤波器对语音信号提取功率谱[24]、幅度谱、相位[25]等信息表征,这些特征也称为手工特征。身份特征包括采用i-vector[26]、x-vector[27]等身份特征信息对真实与伪造语音提取相应的说话人身份信息。无论手工特征还是身份特征,对原始波形来讲必然要损失部分信息。Tak等[28]首先将说话人识别领域基于原始波形的主干网络Rawnet2[29]引入到语音鉴伪工作中,其后团队研究的主干网络AASIST[30]进一步用图注意力机制结合时频域信息,该网络也是目前语音鉴伪最先进的主干网络之一。随着目前预训练大模型的流行,最新的工作通常将大规模自监督预训练特征应用到语音鉴伪领域。由于大模型通常由数十万小时的语音数据训练而来,其隐变量往往能代表真实语音的信息,相比于只在某些小型训练集训出的手工特征,预训练特征拥有着良好的泛化性和鲁棒性。Wang等[31]首先探究了自监督特征如Wav2Vec2[32]、HuBERT[33]等大模型在语音鉴伪模型上的效果,实验证明预训练特征的泛化性要显著好于手工特征。Wang等[34]使用wav2vec2提取语义信息,使用预训练的HuBERT提取时长信息,并使用预训练的conformer提取发音信息,而后运用多头注意力机制融合三种信息进行真假判别,实验表明该方法仅使用单个公开数据集训练,在域外数据集得到了大幅性能提升。

诸如上述前端特征加后端主干网络构成的检测系统往往检测效果有限,缺少对未知伪造方式的泛化性,因此为了提高鉴伪模型的泛化性,在特征与主干网络不变的基础上,一些新的损失函数,新的训练策略孕育而生。Zhang等[35]认为不同类别的语音伪造方式间的分布并不相似,并随着时间推移伪造方法越来越多,为了提高对未知伪造方法检测率,设计了一种新的损失函数OC-Softmax。该方法学习了一个紧凑的真实语音决策边界以及宽松的虚假语音决策边界,仅关注真实语音以提高鉴伪模型泛化性。Ma等[36]使用连续学习方法训练伪造语音检测系统,该模型可帮助减少对过去知识的遗忘,同时通过在真实语音中增加额外的正样本对齐约束,保持真实语音特征表示分布的一致性。最近,数据增强的方法随着鉴伪数据集增加不断涌现,Cohen等[37]详细研究了数据增强在语音鉴伪领域的效果,包括不同压缩方案的数据增强,不同信道的数据增强对鉴伪模型产生的影响,以及介绍了频谱平均化的数据增强方案。Kawa等[38]提出了攻击不变数据集,该数据集融合了目前主流的三个语音鉴伪数据集,使用三种数据集联合训练并用目前先进的前后端进行了效果检验,证明数据集融合可提升域外泛化性。Xie等[39]在其基础上对三个域数据进行域不变表征学习,学习了一个真实域聚合,虚假域分散的域不变特征空间提升了鉴伪模型的泛化性。随后,Xie等[40]进一步结合离线的语音识别大规模自监督特征wav2vec2的最后一层以及改进的LCNN提升泛化性的同时加强了面对噪声数据的鲁棒性。

目前的整体深度伪造音频鉴伪技术主要使用合适的前端特征以及主干网络,关注在域内的测试集上表现。然而在真实场景下,语音鉴伪模型的性能会受到如噪声,录制设备,信道,编码方式等影响,这对当前的深度伪造音频检测模型提出了新的挑战。未来的伪造音频检测模型需要有强泛化性,强鲁棒性,以应对在各种场景下的检测。

3.2 局部伪造音频定位

不同于全局深度伪造音频检测,局部深度伪造音

频定位的研究目前处于起步阶段。在2021年，Yi等[19]建立了第一个局部深度伪造音频数据集，它将部分伪造的语音插入到整体语音中，并使用目前先进的主干网络进行检测，该数据集在不久后应用在中文深度伪造语音检测比赛ADD2022的部分伪造检测赛道。与此同时，Zhang等[18]以19LA数据为基础建立了第一个英文的部分伪造数据集19PS，以上两种部分伪造数据集的建立标志着部分深度伪造检测研究的开始。此后，Zhang等[41]使用SE-LCNN的网络提高了局部伪造音频的鉴别准确率。Lv等[42]使用wav2vec2作为前端，ECAPA-TDNN作为后端取得了ADD2022比赛的局部伪造音频检测的第一名。尽管上述研究对局部深度伪造音频检测的句级别真伪判断已经取得了不错的水平，但很少有研究能够既能判断语句的真假又能对局部伪造进行定位。Zhang等[18]在提出19PS数据集后对该数据集继续展开研究，该方法将该数据集进行了五种帧级别的标注，并使用wav2vec2作为前端，主干网络采用多层感知机模块堆叠，在句级别和帧级别的测试集检测均取得显著的检测效果。Cai等[43]对ADD2022数据集继续开展研究，该方法使用波形边界侦测方法侦测真假语音的边界，最终在ADD2022上取得了6.58%的句级别二分类最佳等错误率，并能实现帧级别的真假定位。Xie等[44]学习了一个域不变特征空间，使用余弦相似度将真实帧与虚假帧相分离并提出动态卷积方法进一步优化计算效率，该方法在19PS数据集当中取得了最低的错误率。

尽管目前对于局部深度伪造音频鉴伪技术的研究已经取得了一定的进展，但大量研究还停留在语句级别的真假判断，在帧级别语音的真假检测以及伪造区域的定位方面仍存在不足之处。由于局部伪造技术的复杂性和隐蔽性，准确地检测和定位帧级别的伪造仍然是一个具有挑战性的问题。对整体深度音频伪造存在的问题如泛化性，鲁棒性不佳等，在局部伪造检测上依然存在，并且由于虚假片段过短使得检测难度相比全局检测大幅提高。

3.3 深度伪造音频溯源

在深度伪造溯源独有的数据集出现前，已经有研究者们对19LA数据集进行溯源的研究。Nicolas等[45]通过研究手工声学特征将19LA的19种伪造方法进行多分类。该方法分别提取了语音的基频信息，失真信息，说话人性别信息，语音持续时长和响度，信号的幅度能量，噪声的幅度等声学信息对进行分类。结果显示使用jitter提取的失真信息分类效果最佳。Zhu等[46]同样对19LA伪造方法进行溯源，该方法首先将19LA的伪造方式分为三大类，包括基于波形生成器，说话人表征和转换类别的伪造方式。而后通过一个多任务的伪造方法检测系统进行溯源围绕19LA的溯源，结果显示其通过完成多任务分类的同时提升了真假二分类判断的准确率。Reddy等[47]提出了一种检测欺骗语音的伪影并识别相应语音生成算法的系统，该方法在19LA的子集上实现了99.58%的多分类准确率。从实验中，该方法发现基于语音源VS特征的系统更加关注音素的转换，而基于声道系统VTS特征的系统更加关注语音信号的静止段。该方法最终基于VS和VTS的系统上执行融合，以利用互补信息来提升溯源性能。Sun等[22]使用伪影研究伪造语音的溯源，该方法为二分类rawnet2模型引入了一个多任务学习框架，该模型与声码器溯源模块共享前端特征提取。该篇文章还提出了LibriSeVoc的新溯源数据集，该数据集使用DiffWave等基于扩散模型的新声码器进行伪造语音合成，还在ASV2019和WaveFake两个公开数据集上进行了测试，结果显示其提出的模型在三种数据集上的表现优于其基线模型。在ADD2023T3中，该比赛要求既能进行真假二分类也能进行伪造的多分类，并且在测试集中含有训练集和验证集均不可见的伪造方式。围绕ADD2023T3，Lu等[48]提出了最佳的解决方案，其使用基于KNN的OOD检测器，通过计算测试集样本与训练集KNN簇类的余弦相似度来判断测试集样本是否来自分布外，该方法通过融合5个基线系统后端的分数和基于KNN的分布外检测方法获得了89.63%的F_1分数，该分数为ADD2023T3的第一名。

4 总结与展望

本文首先将深度伪造音频检测领域细分为三个子领域即全局伪造音频检测、局部伪造音频定位、伪造音频溯源。围绕三个子领域，文章回顾了领域内的数据集和主流伪造音频检测方法。在本节中将总结了深度伪造领域已知的挑战，为深度伪造音频领域未来研究提供方向。

4.1 多样性的音频鉴伪数据集

目前，音频鉴伪研究主要基于ASVspoof系列展开，尽管后续出现了许多数据集工作，如WaveFake和ITW，但这些数据集无法填补目前鉴伪数据集与真实场景测试之间的性能差距。因此，未来的数据集需着

重模拟真实场景,包括丰富的说话人多样性、语种多样性、合成方法多样性、噪声传输方式、压缩格式等多样性,甚至包括音频类别的多样性。目前的伪造音频检测不应只局限于语音,基于音频大语言模型可完成如语音、歌声、音效、音乐等合成任务,建立相关鉴伪数据集是推进鉴伪研究的第一步。

4.2 具有泛化性的伪造音频检测模型

目前,许多音频鉴伪模型常常面临泛化性能差的问题,这通常与测试集与训练集之间存在的域偏移现象有关。如何利用有限的鉴伪数据集来建立具有高泛化性能的音频鉴伪模型是该领域急需解决的问题。这涉及到诸如数据增强、连续学习、域不变表征学习、分布外检测、域自适应等方法。另一方面,对音频伪造检测任务,虽然已有一些研究关注测试域外的泛化性能,但这些工作缺乏统一性,并未形成一致性的比较标准。因此,需要建立统一的跨数据集检测协议,以推动泛化性能的伪造音频检测研究。

4.3 具有可解释性的伪造音频检测模型

对伪造音频检测方法,通常会采用诸如语音识别预训练前端和主干网络进行分类判别。然而,虽然语音识别预训练前端在音频鉴伪领域具有有效性,但对音频真伪的具体解释通常无法由鉴伪网络给出。为了增强伪造音频检测模型的可解释性,可以采用注意力机制热力图和特征可视化等技术。另一方面,结合伪造音频溯源子领域,可以进一步挖掘伪造音频的本质和原因。

参考文献(References):

[1] Wu Z, Kinnunen T, Evans N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge[C]// Sixteenth Annual Conference of The International Speech Communication Association, 2015:2037-2041.

[2] Nautsch A, Wang X, Evans N, et al. ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(2): 252-265.

[3] Yi J, Fu R, Tao J, et al. Add 2022: the first audio deep synthesis detection challenge[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 9216-9220.

[4] Lavrentyeva G, Novoselov S, Tseren A, et al. STC antispoofing Systems for the ASVspoof2019 challenge[C]// Interspeech, 2019: 1033-1037.

[5] Jung J, Heo H S, Tak H, et al. Aasist: audio anti-spoofing using integrated spectro-temporal graph attention networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 6367-6371.

[6] Yamagishi J, Veaux C, MacDonald K, et al. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)[DB/OL]. Edinburgh Scotland: The Centre for Speech Technology Research (CSTR), University of Edinburgh, (2019-11-13)[2024-04-28]. https://datashare.ed.ac.uk/handle/10283/3443.

[7] Liu X, Wang X, Sahidullah M, et al. Asvspoof 2021: towards spoofed and deepfake speech detection in the wild[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2507–2522.

[8] Lorenzo-Trueba J, Yamagishi J, Toda T, et al. The voice conversion challenge 2018: promoting development of parallel and nonparallel methods[C]// The Speaker and Language Recognition Workshop, 2018: 195-202.

[9] Zhao Y, Huang W C, Tian X, et al. Voice Conversion Challenge 2020:-intra-lingual semi-parallel and cross-lingual voice conversion[DB/OL]. arXiv:2008.12527, 2020.

[10] Frank J, Schönherr L. WaveFake: a data set to facilitate audio deepfake detection[C]// NeurIPS Datasets and Benchmarks, 2021.

[11] Müller N, Czempin P, Diekmann F, et al. Does audio deepfake detection generalize?[DB/OL]. arXiv:2203.16263, 2022.

[12] Salvi D, Hosler B, Bestagini P, et al. TIMIT-TTS: a text-to-speech dataset for multimodal synthetic media detection[J]. IEEE Access, 2023, 11: 50851-50866.

[13] Yi J, Tao J, Fu R, et al. Add 2023: the second audio deepfake detection challenge[DB/OL]. arXiv:2305.13774, 2023.

[14] Ma H, Yi J, Wang C, et al. CFAD: a Chinese dataset for fake audio detection[DB/OL]. arXiv:2207.12308, 2022.

[15] Xie Y, Zhou J, Lu X, et al. FSD: an initial Chinese dataset for fake song detection[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024: 4605-4609.

[16] Zang Y, Zhang Y, Heydari M, et al. Singfake: singing voice deepfake detection[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024: 12156-12160.

[17] Yi J, Wang C, Tao J, et al. Scenefake: an initial dataset and benchmarks for scene fake audio detection[J]. Pattern Recognition, 2024, 152: 110468.

[18] Zhang L, Wang X, Cooper E, et al. The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 813-825.

[19] Yi J, Bai Y, Tao J, et al. Half-truth: a partially fake audio detection dataset[C]// Interspeech, 2021.

[20] Zhang B, Sim T. Localizing fake segments in speech[C]// 26th International Conference on Pattern Recognition (ICPR), 2022: 3224-3230.

[21] Yan X, Yi J, Tao J, et al. An initial investigation for detecting vocoder fingerprints of fake audio[C]// 1st International Workshop on Deepfake Detection for Audio Multimedia, 2022: 61-68.

[22] Sun C, Jia S, Hou S, et al. Ai-synthesized voice detection using neural vocoder artifacts[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 904-912.

[23] Zen H, Dang V, Clark R, et al. Libritts: a corpus derived from librispeech for text-to-speech [DB/OL]. arXiv: 1904.02882, 2019.

[24] Alzantot M, Wang Z, Srivastava M B. Deep residual neural networks for audio spoofing detection[C]// Interspeech, 2019.

[25] Sanchez J, Saratxaga I, Hernáez I, et al. Toward a universal synthetic speech spoofing detection using phase information[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(4): 810-820.

[26] Jung J, Shim H, Heo H S, et al. Replay attack detection with complementary high-resolution information using end-to-end dnn for the ASVspoof 2019 challenge[C]// Interspeech, 2019: 1083-1087.

[27] Chen T, Khoury E. Spoofprint: a new paradigm for spoofing attacks detection[C]// IEEE Spoken Language Technology Workshop (SLT), 2021: 538-543.

[28] Tak H, Patino J, Todisco M, et al. End-to-end anti-spoofing with RawNet2 [C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 6369-6373.

[29] Jung J, Kim S, Shim H, et al. Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms[DB/OL]. arXiv:2004.00526, 2020.

[30] Jung J, Heo H S, Tak H, et al. AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 6367-6371.

[31] Wang X, Yamagishi J. Investigating self-supervised front ends for speech spoofing countermeasures [DB/OL]. arXiv: 2111.07725, 2021.

[32] Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for selfsupervised learning of speech representations[C]// 34th Conference on Neural Information Processing Systems (NeurIPS), 2020, 33: 12449 - 12460.

[33] Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[34] Wang C, Yi J, Tao J, et al. Detection of cross-dataset fake audio based on prosodic and pronunciation features[C]// Interspeech, 2023.

[35] Zhang Y, Jiang F, Duan Z. One-class learning towards synthetic voice spoofing detection[J]. IEEE Signal Processing Letters, 2021, 28: 937-941.

[36] Ma H, Yi J, Tao J, et al. Continual learning for fake audio detection[C]// Interspeech, 2021: 886-890.

[37] Cohen A, Rimon I, Aflalo E, et al. A study on data augmentation in voice anti-spoofing[J]. Speech Communication, 2022, 141: 56-67

[38] Kawa P, Plata M, Syga P. Attack agnostic dataset: towards generalization and stabilization of audio deepfake detection[C]// Interspeech, 2022: 4023-4027.

[39] Xie Y, Cheng H, Wang Y, et al. Domain generalization via aggregation and separation for audio deepfake detection[J]. IEEE Transactions on Information Forensics and Security, 2023, 19: 344 - 358.

[40] Xie Y, Cheng H, Wang Y, et al. Learning a self-supervised domain-invariant feature representation for generalized audio deepfake detection[C]// Interspeech, 2023: 2808-2812.

[41] Zhang L, Wang X, Cooper E, et al. Multi-task learning in utterance-level and segmental-level spoof detection [C]// Automatic Speaker Verification and Spoofing Countermeasures Challenge(ASVSPOOF), 2021: 9-15.

[42] Lv Z, Zhang S, Tang K, et al. Fake audio detection based on unsupervised pretraining models [C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 9231-9235.

[43] Cai Z, Wang W, Li M. Waveform boundary detection for partially spoofed audio[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023: 1-5.

[44] Xie Y, Cheng H, Wang Y, et al. An efficient temporary deepfake location approach based embeddings for partially spoofed audio detection[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024: 966-970.

[45] Müller N M, Dieckmann F, Williams J. Attacker attribution of audio deepfakes[C]// Interspeech, 2022: 2788-2792.

[46] Zhu T, Wang X, Qin X, et al. Source tracing: detecting voice spoofing[C]// Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2022: 216-220.

[47] Reddy T U K, Varun S C, Sreekanth K P K S, et al. Evince the artifacts of spoof speech by blending vocal tract and voice source features[DB/OL]. arXiv:2212.02013, 2022.

[48] Lu J, Zhang Y, Li Z, et al. Detecting unknown speech spoofing algorithms with nearest neighbors[C]//Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis, 2023: 89-94.

编辑：王谦