

深度伪造音频生成与鉴伪技术综述

曾志平¹, 张旭龙¹, 瞿晓阳¹, 肖春光², 王健宗¹

1. 平安科技(深圳)有限公司, 广东 深圳 518063;
2. 深圳市宝安区教育事业发展中心, 广东 深圳 518101

摘要

随着深度学习技术的快速发展, 伪造音频生成技术, 尤其是深度伪造音频的生成, 已成为数字媒体安全领域的一大挑战。这些技术生成的音频与真实语音高度逼真, 难以分辨, 从而对社会安全和个人隐私保护提出了更高要求。鉴于此, 深度伪造音频鉴伪技术的研究显得尤为迫切。尽管该领域已取得一定进展, 但随着数据量、算力和新模型的不断涌现, 新的挑战也随之而来。首先, 概述了伪造音频生成技术从传统音频编辑到基于深度学习生成模型的演进历程。接着, 深入分析了基于声学特征的鉴伪方法和基于端到端模型的鉴伪策略, 详细讨论了深度声学特征鉴伪、预训练神经网络特征鉴伪, 端到端模型优化、泛化提升策略及实时检测的强化。最后, 展望了未来工作, 强调了需要重点关注的研究方向, 以推动深度伪造音频鉴伪技术的发展, 更有效地应对日益复杂的伪造音频威胁, 保障数字媒体的安全性与可靠性。

关键词

深度伪造音频; 伪造音频生成检测; 深度音频生成模型; 鉴伪技术

中图分类号: TP181

文献标志码: A

doi:10.11959/j.issn.2096-0271.2025064

Survey of deep fake audio generation and detection techniques

ZENG Zhiping¹, ZHANG Xulong¹, QU Xiaoyang¹, XIAO Chunguang², WANG Jianzong¹

1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China
2. Shenzhen Baoan District Education Information Center, Shenzhen 518101, China

Abstract

With the rapid development of deep learning technology, fake audio generation technology, especially deep fake audio generation, has become a major challenge in the field of digital media security. The audio generated by these technologies is highly realistic and difficult to distinguish from real speech, which puts forward higher requirements for social security and personal privacy protection. In view of this, the research of deep forgery audio authentication technology was particularly urgent, although some progress has been made in this field, but with the continuous emergence of data volume, computing power and new models, new challenges also follow. Firstly, the evolution of fake audio generation techniques was outlined, transitioning from traditional audio editing to deep learning-based generation models. Subsequently, an in-depth analysis was conducted on both acoustic feature-based and end-to-end model-based fake audio detection strategies, delving into details such as deep acoustic feature detection, pre-trained

neural network feature detection, end-to-end model optimization, generalization enhancement techniques, and the enhancement of real-time detection. Lastly, future work was envisioned, highlighting the key research directions that require attention, in order to advance the development of deep fake audio detection techniques, more effectively counter increasingly complex fake audio threats, and ensure the security and reliability of digital media.

Key words
deep fake audio, audio forgery detection, deep audio generative model, detection technique

0 引言

随着互联网的快速发展和普及，短视频在各个领域呈现爆炸性增长。便捷编辑工具的广泛普及赋予了个人创作个性化多媒体内容的能力，而人工智能生成内容（artificial intelligence generated content，AIGC）的兴起进一步加速了该趋势。在此背景下，深度伪造技术凭借其能够生成高度逼真的多媒体音频和视频内容的能力，引起了社会各界的广泛关注。深度伪造技术，特别是在音频和视频领域，利用深度学习和生成对抗网络（generative adversarial network，GAN）^[1]等先进的人工智能技术，能够合成出逼真、难以辨识真伪的多媒体内容。这些技术在娱乐、电影制作和营销等领域展现出巨大的应用潜力，但同时也引发了人们对其可能被滥用于非法目的的严重担忧。在多媒体音频领域，文本转语音（text-to-speech，TTS）技术和语音转换（voice conversion，VC）技术，如 ChatTTS^[2]和 Seed-TTS^[3]，能够生成与人类语音几乎无法分辨的高质量、自然音频。现有的 TTS 技术主要针对端到端系统进行优化，可实现自然而富有表现力的语音合成，在确保实时性的同时，还能预测和控制细粒度的韵律特征，包括停顿、插入词和笑声等^[2,4]。在语音合成的上下文学习、说话者特征相似度模拟以及语

音自然度等方面，Seed-TTS 展现出卓越的性能，并可通过自我蒸馏和强化学习，进一步提升其性能和可控性^[3]。随着大语言模型（large language model，LLM）的兴起，利用 LLM 的能力，语音合成系统在生成语音的质量和自然度方面取得了显著提升^[5-6]。此外，VC 技术也同步发展，实现了自然流畅^[7-8]、风格多样性^[9]和富有情感^[10]等技术特性，使转换后的语音越来越接近真实的人类语音。VC 技术能够生成更加真实、自然、难以区分于真实音频的语音。快速更新迭代的 TTS 技术和 VC 技术^[11-12]，虽然推动了语音合成转换领域的发展但也给分辨伪造音频和真实人类语音带来了前所未有的挑战。

深度生成音频技术在合法语音合成领域被广泛应用，其中包括 TTS 技术和 VC 技术，这些技术主要致力于改善语音交互的体验。然而，深度伪造音频生成技术则被用于恶意目的，旨在伪造或篡改音频内容，进行欺诈或误导行为。深度伪造音频生成技术的潜在滥用对个人隐私、社会信任和信息安全构成了严重威胁，操纵音频内容可能导致个人声誉受损、财产面临风险，甚至对政治稳定构成威胁。例如，通过伪造公众人物的语音进行虚假宣传或诈骗，可能会误导公众、引发恐慌或被用于政治操纵^[13]。此外，随着社交媒体和在线通信平台的兴起，错误信息和虚假新闻的传播速度加快，深度伪造音频生成技术可能会加剧该问题，使真实与虚构更加难以区分，

从而对社会秩序和公共安全造成破坏^[14]。

因此，开发有效的伪造音频鉴伪技术变得至关重要。这包括传统的声学特征分析^[13-14]、基于深度学习的端到端检测系统的构建^[15]等。这些技术的进步有助于提高伪造音频的检测准确性和效率，从而保护个人和社会免受其害。然而，这一领域的研究仍面临诸多挑战。例如，可用的伪造音频数据集相对有限，且标注工作烦琐，这给鉴伪模型的训练带来了挑战。此外，随着伪造技术的不断进步，检测模型需要不断更新以适应新的伪造手段^[16]。实时检测的需求也要求检测系统具有高效的计算能力，以适应实时音频流的鉴伪^[17-18]。未来的研究方向包括多模态融合鉴伪、提高鉴伪模型的可解释性以及增强模型在不同数据集和现实世界条件下的泛化能力。总之，深度伪造音频生成技术的发展对多媒体音频领域而言，既是机遇，又是挑战，人们在享受其带来的便利和创新应用的同时，也必须警惕其潜在的滥用风险，并采取相应的措施，以防范和减轻这些风险。

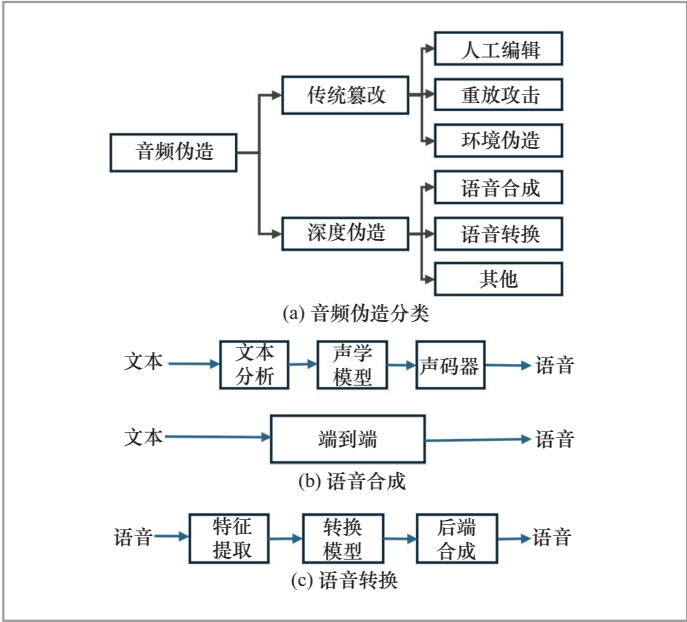


图1 伪造音频生成的方法

本文综述了深度伪造音频生成与鉴伪技术，探讨了这一领域的最新进展和未来研究方向。首先，介绍伪造音频生成技术的发展，包括传统的音频编辑技术、基于深度学习的音频生成和现有的伪造音频鉴伪数据集。接着，详细介绍各种伪造音频鉴伪技术，包括基于声学特征分析和基于端到端模型的鉴伪方法。最后，展望未来研究方向，包括多模态融合鉴伪、模型可解释性分析以及跨数据集泛化。本文旨在为相关研究人员和从业者提供该领域的全面视角，并促进未来技术的发展。

本文的主要贡献如下。

- 本文以一种比较新颖的分类方法详细地介绍了音频可能受到的欺诈伪造与相应的鉴伪措施，系统地介绍了伪造音频生成与鉴伪技术从早期的典型技术到最新的研究成果的发展历程。
- 本文以多样化的方式直观展示了伪造音频生成的可能性和手法，针对这些伪造方法，找出其相应的鉴伪方法，并对这些鉴伪方法进行了翔实介绍和分析。
- 根据深度伪造音频的特性与现状，本文从多角度对未来的发展趋势进行了展望，并针对一些关键问题和实现途径提出了指导性建议。

1 伪造音频生成技术

伪造音频生成技术的发展经历了从依赖手工操作的传统编辑方法，到现代深度伪造技术的兴起。伪造音频生成的方法如图1所示，包括人工编辑、重放攻击、环境伪造、语音合成、语音转换以及特殊场景下的环境伪造等。

传统的伪造音频生成方法主要依赖于音频编辑软件，通过手动剪辑、拼接和调

节音频片段来创建伪造音频。这种方法虽然可以实现基本的伪造音频生成，但在音质和自然度上存在明显缺陷，且需要大量的时间和较高的专业技能。除了基本的编辑技术，重放攻击也是一种传统手段，它涉及将已有的录音在不改变内容的情况下重新播放，以此来误导或欺骗听众。这种方法简单直接，但容易被识破，因为它缺乏音频内容的原创性。环境伪造是一种特殊的伪造音频生成技术，它通过模拟或改变音频中的背景环境声音来匹配或掩盖伪造的语音内容，从而增加伪造音频的真实感和说服力。

随着人工智能和深度学习技术的飞速发展，伪造音频生成技术进入了一个全新的阶段——深度伪造音频生成技术。该技术依赖于深度学习模型的强大能力，能够生成高质量、逼真的伪造音频。深度伪造音频生成技术除了包含语音合成和语音转换这两种核心应用，还包括情绪伪造、深度场景伪造以及部分伪造等其他技术。

本节综合探讨伪造音频生成技术的演变，首先回顾传统音频篡改技术，包括人工编辑、重放攻击和环境伪造等手段，分析它们在实施方式、应用场景及局限方面的具体特点；随后，深入分析深度学习技术的融入如何使伪造音频生成达到新的高度，特别是TTS技术和VC技术，这些技术能够生成与真实录音难以分辨的高质量伪造音频；另外介绍了目前用于评估和推动伪造音频鉴伪技术发展的深度伪造音频鉴伪数据集，讨论它们在该领域的重要作用以及各自的优势和局限性。通过这些数据集，研究者可以更全面地评估和提升伪造音频鉴伪技术的性能，以应对不断进步的伪造手段。

1.1 传统篡改音频技术

在深度学习兴起之前，伪造音频的生

成主要依赖于传统的篡改音频技术，使用各种音频编辑软件，通过手动操作实现对音频的篡改和伪造。传统篡改音频技术包括人工编辑、重放攻击和环境伪造。

人工编辑是最基本的音频篡改手段，涉及直接使用音频编辑软件对音频文件进行剪辑、拼接、删除或插入操作。编辑者可以通过精细的剪辑技术移除不需要的部分或插入额外的音频片段，以改变音频信息的上下文或内容。高级的编辑技术还包括对音频进行淡入淡出处理，以减少剪辑点的突兀感。变调和变速^[19]是通过调整音频的音高（pitch）和播放速度（tempo）来改变说话者的语调或语速。这种技术可以用来模拟不同的情感表达或掩盖说话者的身份。然而，不恰当的变调或变速可能会导致音频听起来机械或不自然。

重放攻击是一种特殊的篡改音频手段，攻击者通过录制目标音频的播放内容，并对其进行重新录制，从而生成一个新的音频文件。这种方式常被用于欺骗声纹识别系统，因为重新录制的音频保留了原始音频的主要声学特征，同时掩盖了直接录制的痕迹。该技术可以用于测试声纹识别系统的安全性，也可以用于误导监听者，使其认为音频是在特定环境下实时录制的。

环境伪造则是通过加入噪声和失真处理技术来添加背景噪声、混响效果等，以模拟不同的声学环境，如会议室、街道或家庭环境等，使伪造音频听起来更加真实。这种方法通过掩盖音频中的不自然过渡或不一致性，掩盖篡改痕迹，使伪造内容更加难以被识破，从而增加音频的可信度。这种方法进一步加大了音频分析的难度，尤其是在缺乏清晰参考的情况下。

尽管上述方法可以实现基本的音频篡改，但存在明显的缺陷。首先，传统篡改音频技术往往需要大量的时间和较高的专

业技能，难以实现大规模应用。其次，剪辑和拼接音频容易留下可检测的痕迹，如不自然的停顿和音质变化等。此外，传统篡改方法在音质和自然度上难以与现代深度伪造音频技术相比。

1.2 现代深度伪造音频技术

随着深度学习技术的兴起，伪造音频生成技术已经能够自动且高效地生成高质量、难以检测的伪造音频，这在很大程度上超越了传统篡改音频技术的能力和局限。现代深度伪造音频技术不仅能够模拟声音，还能够创造出全新的、逼真的音频内容，且这些内容听起来几乎无法与真实录音分辨。

TTS 技术旨在将文本转换为自然的语音^[20]。图 1 (b) 是目前主流的两种基于神经网络的语音合成方法。一种是标准的语音合成方法，需要先分析文本，提取规范化文本特征向量，输入声学模型后变成对应的声学特征向量，最终声码器将声学特征通过反变换得到相应的声音波形。这种传统的 TTS 技术使用拼接合成和参数合成方法，生成的语音通常不够自然且缺乏流畅性。另一种是端到端的方法，可以直接从文本生成高质量的音频。近年来，基于深度学习的 TTS 技术（如 Seed-TTS^[3]、Tacotron^[20]、WaveNet^[21]）等取得了显著进展，这些技术通过端到端的训练流程，可以直接从文本生成高质量的语音，解决了传统方法的许多不足。Seed-TTS 引入了一个非自回归（non-autoregressive, NAR）变体，命名为 Seed-TTSDiT，该变体采用基于完全扩散的架构，不依赖于预先估计的音素持续时间，而是通过端到端处理来实现语音生成。Tacotron 系列模型采用了序列到序列（sequence-to-

sequence）架构，通过注意力机制实现了文本到语音的直接映射。WaveNet 则是一种生成模型，通过逐点预测波形样本，能够生成高度自然的语音。结合上述两种语音合成方法，最新的 TTS 技术可以在不同语言、音色和语速条件下，生成逼真且自然的伪造语音。例如，Vocos^[4]作为一种新型的神经声码器，在提高音频合成速度和效率的同时，能够保持高质量的音频输出。Vocos 模型能够直接生成音频片段的傅里叶频谱系数，相较于传统的在时域中操作的方法，这种方法更加高效且与人类的听觉感知更加一致。此外，LLM 在多个自然语言处理（natural language processing, NLP）任务中表现出色。利用 LLM 的预训练能力，TTS 技术可以在大量的文本数据上进行预训练，然后在语音数据上进行微调。这种方法可以使 TTS 模型更好地捕捉语言模式，从而提高语音生成的质量和自然度^[22]。TTS 技术^[23-25]通过结合 LLM，与传统的自回归语言模型（如 VALL-E）相比，在处理难以合成的句子时，不仅表现出更高的鲁棒性，还能增强模型对复杂文本内容的处理能力，实现更加个性化和可控的语音合成^[5]。例如，Llama-VITS 模型通过结合 Llama2 生成的语义嵌入，优化了 VITS 模型在 TTS 合成中的表现，特别是在自然语音生成和情感语音表达上表现出色，显著增强了情感表达能力^[6]。

VC 技术旨在将一种语音转换为另一种语音，即在保持语音内容不变的情况下，改变说话者的音色和音调^[6]。图 1 (c) 描述了 VC 技术的基础框架。传统的语音转换方法依赖于语音信号处理技术，通过分析和合成语音特征实现转换，但效果往往不够理想。深度学习的引入彻底改变了 VC 技术的发展方向。基于自动编码器的模型

和GAN在语音转换领域展现了强大的能力。自动编码器通过编码器-解码器架构将语音特征压缩成低维表示,再从中重建出目标语音特征。GAN则通过生成器和判别器的对抗训练,使生成的语音更具真实性。例如,StarGAN-VC^[26]模型采用了多域语音转换的策略,通过一个单一模型实现多种说话者之间的语音转换。CycleGAN-VC^[17]则通过循环一致性损失,确保转换后的语音能够保持内容一致,同时具备目标音色。Cai等^[8]提出了一种基于可逆神经网络的并行语音转换框架,该框架不仅能够实现语音转换,还能追溯源说话人,提高了系统的安全性和可靠性。Shan等^[27]则提出了Phoneme Hallucinator模型,仅使用目标说话者短暂的语音片段,就能生成多样化且高保真的目标语音。此外,作为一种新颖的风格语音转换方法,PromptVC^[9]利用潜在扩散模型根据自然语言提示生成风格向量,从而实现更直观和灵活的语音转换控制。Choi等^[28]提出了DDDM-VC,这是一种解耦去噪扩散模型,用于经过验证的鲁棒语音转换,该模型通过解耦表示和先验混合实现了对每个属性的独立风格转换。Qi等^[10]探索了PAVITS,这是一种端到端的情感VC模型,它通过增强内容自然性和情感自然性来满足人类感知的需求。总的来说,上述这些方法通过提高语音转换的自然度、情感表达、风格多样性和安全性,使转换后的语音越来越接近真实的人类语音,VC技术能够生成更加真实、自然、难以区分于真实音频的语音。

在深度伪造音频技术的领域中,除了语音合成和语音转换这两种常见的应用,还涵盖了情绪伪造、场景伪造和部分伪造等其他类型的技术。其中,情绪伪造技术能够将说话者的情绪从一种状态转变为另

一种,如从高兴变为伤心,这是通过分析和修改音频信号中的情感特征来实现的。场景伪造技术则能够将音频中的背景声音从一个环境替换为另一个,如将安静的办公室背景音替换为机场的嘈杂声,这种技术可能被用于基于环境的地点欺骗。部分伪造技术则允许对现有音频进行局部篡改,如删除或替换一句话中的某些词,而不影响其他部分,实现对特定语句的精准篡改。

现代深度伪造音频技术在娱乐、教育和医疗领域均展现了其巨大的应用潜力,如用于配音、语音助手和虚拟角色的音频生成,以及帮助创建个性化的学习材料和康复训练等。然而,深度伪造音频技术的滥用也带来了严峻的安全和伦理问题,如虚假信息的传播和身份欺诈等。随着伪造音频生成技术从传统的编辑方法发展到依赖深度学习的高级伪造技术,伪造音频的质量和自然度得到了极大的提升。在此过程中,语音合成和语音转换作为深度伪造音频的两大核心技术,不仅在不断进步,也对伪造音频鉴伪技术提出了更高的要求。

深度伪造音频鉴伪数据集是评估和推进伪造音频鉴伪技术发展的关键资源。现有的伪造音频鉴伪代表性数据集见表1。其中,ASVspoof数据集^[29-31]专注于自动说话人验证(automatic speaker verification, ASV)系统的防伪攻击,提供了多种类型的语音伪造样本,包括语音转换和文本到语音,其多样性和覆盖范围广泛,为开发和评估防伪措施提供了重要的基础。然而,ASVspoof的主要限制在于其模拟场景和实验室环境,可能与实际应用中的复杂情况存在差距。In-The-Wild^[32]数据集收集自互联网的真实世界数据,包含大量自然场景下的伪造和真实语音,反映了真实世界中的多样性和复杂性,使鉴伪模型能够在更加接近现实的条件

表1 现有的伪造音频鉴伪代表性数据集

方法	年份	语言	类型	伪造方法数量	背景环境	真/假句子数量	特点
ASVspoof ^[29-31]	2019	英	TTS、VC	19	clean	10 000/90 000	说话人验证和伪造检测
	2021	英	TTS、VC	19	codec	14 000/133 000	说话人验证和伪造检测
	2021	英	TTS、VC	100+	codec	14 000/519 000	深度伪造检测
In-The-Wild ^[32]	2022	英	-	-	-	20 000/12 000	真实世界伪造
WaveFake ^[33]	2021	英、日	TTS	-	clean	0/118 000	伪造音频生成检测
FoR ^[34]	2019	英	TTS	7	clean	108 000/87 000	使用多种不同 TTS 算法
HAD ^[35]	2021	中	TTS	-	clean	53 000/753 000	中文部分伪造
FakeSound ^[36]	2024	中	TTS	-	-	0/3 000	精确时间戳关键段注释
ADD ^[37-39]	2022	中	TTS、VC	-	noisy	37 000/124 000	低质量嘈杂环境
	2023	中	TTS、VC	-	codec、noisy	55 000/65 000	低质量、部分伪造
MLAAD ^[40]	2024	多种语言	TTS	54	clean	0/76 000	多语言伪造音频生成检测

进行评估和训练。尽管如此，由于其数据来源的多样性，In-The-Wild 数据集在标注质量和一致性上可能存在一定的问题。WaveFake 数据集^[33]专注于由现代语音生成模型（如 WaveNet 和 WaveGAN）生成的伪造语音，提供了高质量的音频样本，适用于细粒度分析和生成模型对检测影响的研究。WaveFake 数据集的主要优势在于其高保真度和多样化的伪造样本，但由于其主要关注生成模型的特定类型，可能在覆盖面上有所局限。FoR 数据集^[34]是一个英语语音数据集，包含真实和伪造的语音，由多种 TTS 算法生成。该数据集的主要优势在于其多样化的预处理方法，但其局限性在于主要关注英语语音，跨语言适用性有限。HAD 数据集^[35]是一个高质量的半真音频检测数据集，包含各种语音生成和转换技术生成的伪造音频，支持多种伪造检测算法的开发和测试。HAD 数据集中的部分假音频只涉及更改话语中的几个单词，这些单词的音频是采用最新的先进语音合成技术生成的。FakeSound 数据集^[36]

是一个为深度伪造通用音频检测任务设计的自动化和多样化的音频数据集，它通过一个创新的自动化操纵流程生成，包含精确的注释，支持多任务学习和不同难度级别的测试集，旨在推动深度伪造音频鉴伪技术的研究和发展。ADD 数据集^[37-39]则关注于低质量伪造音频的鉴伪，这对于识别和防御在不理想录音条件下生成的深度伪造音频至关重要。ADD 数据集的局限性在于其伪造技术可能被快速迭代，需要定期更新，以保持其前沿性和实用性。MLAAD 数据集^[40]包含多种语言的伪造语音，支持跨语言的伪造检测，其设计强调跨语言检测模型的性能评估，为国际化研究提供了宝贵资源。然而，多语言数据集在标注和一致性处理上可能面临更大的挑战。

每个数据集都有其独特的优点，这为深度伪造音频鉴伪技术提供了宝贵的资源，同时也存在一些局限性，需要通过持续的研究和数据集的更新来克服。这些数据集可以在不同的场景和条件下全面评估伪造音频鉴伪技术的性能。这种综合评估方法

不仅有助于提升单一检测模型的准确性和鲁棒性，还能为开发更先进的多模态检测系统提供支持。通过持续的研究和数据集的更新，鉴伪技术能够不断适应新出现的伪造手段和挑战，从而更有效地保护个人隐私、社会信任和信息安全。

2 深度伪造音频鉴伪技术

随着深度学习技术的迅猛发展，伪造音频生成技术也在不断进步，产生了越来越逼真的深度伪造音频。这些伪造音频不仅威胁到信息安全和个人隐私，还可能被用于制造虚假新闻、诈骗和其他违法活动。因此，如何有效检测和识别深度伪造音频，已成为一个亟待解决的问题。近年来，研究人员提出了多种基于机器学习和深度学习的方法来检测伪造音频，这些方法在特征提取、模型架构和数据处理等方面不断优化，以提高检测的准确性和泛化能力。

本节系统综述近年来深度伪造音频鉴伪技术的重要研究进展。首先，探讨特征分析方法，最早的方法主要依赖于手工设计的特征提取技术，包括恒定Q倒谱系数和线性频率倒谱系数等，通过这些特征来训练分类器，如高斯混合模型等。然而，随着深度学习技术的引入，基于深度神经网络（deep neural network, DNN）、卷积神经网络（convolutional neural network, CNN）和循环神经网络（recurrent neural network, RNN）等模型的方法逐渐占据主导地位，这些模型能够自动学习和提取特征，显著提高了检测精度。特征分析方法包括声学特征分析和自监督预训练神经网络特征分析。声学特征分析方法利用音频的频谱图、相位信息等特征进行分析，而自监督预训练神经网络

特征分析方法则利用 wav2vec 2.0^[41]、HuBERT^[42]和 WavLM^[43]等自监督预训练模型，通过迁移学习来增强检测能力。其次，进一步分析基于深度学习的端到端模型方法。端到端模型方法直接处理原始波形或频谱图，不需要手工设计特征，进一步提高了检测性能。这部分详细介绍了多种模型优化的方法，包括采用不同架构的DNN，如 RawGAT-ST^[15]、RawNet2^[44]、SpecRNet^[45]和 AASIST^[46]等，同时针对泛化性的问题提出了解决方案。这些方法在处理多样化的伪造音频时表现出色，特别是在面对不同类型的伪造手段时，展现了较强的适应能力。此外，实时检测方法也是一个重要的研究方向。随着深度伪造音频技术的不断进步，实时伪造音频鉴伪技术变得尤为重要。本文讨论了近年来在实时音频流检测方面的最新进展，这些技术能够在不影响检测精度的前提下，实现高效的实时检测。

本节详细对比了具有代表性的深度伪造音频鉴伪技术的特性，展示当前主要研究方法的优缺点，通过对这些方法的综合评估，更全面地了解当前技术的发展状况，并发现未来可能的研究方向。

2.1 基于特征分析的鉴伪技术

伪造特征分析鉴伪是伪造音频检测的关键步骤。传统的伪造音频鉴伪方法依赖于手工提取特征，通过提取恒定Q倒谱系数、线性频率倒谱系数等特征来训练分类器，如高斯混合模型（Gaussian mixture model, GMM）和 X-vector。这些特征捕捉了音频信号的频域和时域信息，通过对这些信息进行分析，可以在一定程度上识别出伪造音频的特征^[13-14]。然而，随着伪造音频生成技术的发展，传统方法在面

对日益复杂的伪造手段时，表现出一定的局限性。例如，使用简单的统计模型和固定的特征提取方式，难以应对高级的GAN和其他先进的伪造技术。许多研究人员尝试手动调整特征提取器，以提高检测性能。例如，FastAudio^[47]是一种通过学习优化特征提取过程的方法。这种学习型前端能够在伪造语音检测中超过最好的手工制作特征，因为它们能够适应下游任务，动态调整特征提取过程，使其更适合特定的音频数据和检测任务。然而，这种手动构建过程费力且耗时，需要大量的试验和调整。

随着深度学习技术的发展，研究人员开始探索基于DNN的声学特征提取方法和基于自监督预训练的神经网络（neural network, NN）特征方法。这些方法具有自动学习和提取复杂特征的能力，能够捕捉到音频信号中更深层次的模式和特征。基于DNN的声学特征提取方法，如使用CNN和RNN，能够自动学习音频信号中的时频特征，显著提高伪造音频检测的准确性和鲁棒性。基于自监督预训练的NN特征方法则利用大量未标注的数据，通过自监督学习（self-supervised learning, SSL）技术进行预训练，然后将预训练模型应用于伪造音频鉴伪任务。这些预训练模型，如wav2vec 2.0、HuBERT和WavLM，能够捕捉到音频信号的丰富特征，并且具有很强的迁移学习能力。在鉴伪任务中，这些预训练模型能够显著提高鉴伪性能，同时减少对标注数据的依赖。

伪造音频鉴伪技术正在从传统的手工特征提取方法向基于深度学习和SSL的方法转变。通过引入深度学习和SSL技术，研究人员能够更有效地应对复杂的伪造手段，提高伪造音频检测的准确性和鲁棒性，为保护信息安全和个人隐私提供了重要的技术支持。

综上，对深度伪造音频鉴伪技术基于伪造特征分析的鉴伪方法进行比较，见表2，从声学特征到预训练NN特征方向分析对比不同的方法发挥的作用。本节对这两个方向展开详细的分析。

2.1.1 深度声学特征方向

在深度声学特征方向的伪造音频鉴伪领域，研究人员致力于开发更精准和鲁棒的鉴伪系统，以识别和抵御伪造音频攻击。特征提取是构建这些系统的关键步骤，涉及多种声学特征的探索，包括TTS和VC的神经声码器特征、基频（fundamental frequency, F0）特征、相位信息、韵律，以及多模态特征的融合，这些方法均旨在增强鉴伪系统的性能。

神经声码器特征正受到越来越多的关注，尤其是在SSL中的应用。Wang等^[48]利用神经声码器生成的数据，通过SSL模型进行训练，有效提升了模型识别伪造音频的能力。这一方法为伪造音频鉴伪领域带来了新的训练策略，显著提高了模型性能。然而，额外生成的声码器数据会增加训练的复杂性，并且很难泛化到未见过的声码器生成的测试数据或其他方式生成的数据的检测中。

F0特征，因其在合成语音与真实语音之间存在的显著差异，成为研究的焦点。Fan等^[49]提出的SR-LA Res2Net模型专注于F0子带，有效提高了对假语音的检测能力。Xue等^[50]则展示了结合F0和对数功率谱（log power spectrogram, LPS）特征，并利用了ECANet和SENet分类器的自蒸馏技术在提升检测性能方面的潜力。该方法通过自蒸馏框架传递深度网络知识，显著提高了鉴伪精度。尽管F0特征能有效反映伪造音频的特征，但随着伪造技术的不断发展，新的技术很快就会针对F0特征

表2 基于伪特征分析的鉴别方法比较

特征分析方向	算法	年份	特征	分类器	是否采用数据增强	是否需要额外训练数据	是否需要自监督预训练模型	是否需要人工编辑特征	是否针对解决泛化问题	备注
声学特征	Vocoded-Based ^[48]	2023	wav2vec 2.0	NN-based	✓ (RawBoost)	✓	✓	×	✓	结合 vocoded 数据和基于 SSL 的 CMs
	SR-LA Res2Net ^[49]	2023	F0 子带	Res2Net	×	×	×	×	✓	F0 子带特征和 SR-LA Res2Net 架构
	Self-Distillation ^[50]	2023	F0、LPS	ECANet、SENet	×	×	×	✓	×	自蒸馏方法
	Phase-Aware ^[51]	2023	相位	Res2Net	×	×	×	✓	×	相位网络减小幅度和相位特征的差异
	Prosodic-Based ^[52]	2023	wav2vec、韵律、发音	LCNN、Bi-LSTM	×	✓	✓	✓	✓	提出多视图特征方法
	M2S-ADD ^[53]	2023	单声道转双声道 (M2S)	Dual-branch	✓/(noise)	✓	×	×	✓	通过立体声信息增强泛化能力
	DKDSSD ^[54]	2024	结合去噪和噪声特征	SSD	✓	×	✓	×	×	噪声环境中的性能
	FairSSD ^[55]	2024	MFCC、LFCC、频谱、SSL	GMM、ResNet、CNN	×	×	✓	✓	×	对性别、年龄、口音和语障的偏见分析
	Frame-to-Utt ^[56]	2024	SDC、STC	Bi-LSTM	✓	×	×	×	×	局部频谱偏差和时序系数融合策略
	Wav2vec-Based ^[57]	2023	wav2vec 2.0	Conformer	✓ (RawBoost)	×	✓	×	×	处理可变长度的语音输入
预训练 NN 特征	WavLM-Based ^[58]	2024	WavLM	MFA	×	×	✓	×	✓	基于预训练模型
	SSL-Based ^[59]	2023	wav2vec 2.0、HuBERT	Conv-TasNet	×	×	✓	×	✓	预训练模型以及攻击者使用的增强模型
	Whisper-Based ^[60]	2023	Whisper	MesoNet	×	×	✓	×	✓	Whisper 模型应用在检测中
	Fusion Feature ^[61]	2024	XLS-R、WavLM	ResNet18	×	×	✓	×	✓	提出多视图特征结合方法

形成应对。因此，依赖单一特征在欺诈检测方面的可扩展性有限，难以应对多种伪造手段。

相位信息在提高欺诈检测精度中也起着关键作用，尤其是在对抗基于语音转换的攻击时。Kim 等^[51]引入的相位网络有效降低了相位特征的随机性，增强了鉴伪系统对欺诈语音的识别能力。然而，同样依赖单一特征在欺诈检测方面的扩展性有限。基于此，Wang 等^[52]提出新的检测方法，该方法利用注意力机制优化了 wav2vec 特征、韵律特征和发音特征的权重，从而在多模态特征融合方面解决了鉴伪问题。

此外，还有其他最新的研究，如 Liu 等^[53]探索了单声道到双声道转换在增强欺诈检测方面的潜力，他们提出的 M2S-ADD 模型采用双分支神经架构处理立体声信号，有效揭示了假音频中的伪影。文献[54]强调了结合去噪和噪声特征的重要性，通过交互式融合模块，提高了鉴伪系统的鲁棒性。文献[55]侧重于融合梅尔频率倒谱系数 (Mel-frequency cepstral coefficients, MFCC)、线性频率倒谱系数 (linear frequency cepstral coefficients, LFCC)、频谱特征以及 SSL 中获得的特征，这种多特征融合策略进一步提升了检测系统对伪造音频的识别能力。文献[56]提出了一种融合局部频谱偏差和时序系数的策略，通过检测帧级别和话语级别的异常，增强了对伪造语音的检测能力。

这些研究表明，各类深度声学特征都对鉴伪起到增强作用，创新的特征融合策略和深度学习架构能显著提升声学伪造检测的准确性和鲁棒性。未来的研究将继续探索新的声学特征和网络架构优化，以进一步提升鉴伪系统的性能，尤其是 F0 子带特征的研究，这将为深度伪造音频鉴伪领域开拓新的视野和可能性。

2.1.2 预训练神经网络特征方向

随着 GAN 等生成模型技术的发展，合成语音与真实语音的区分变得更加困难，这对深度伪造音频鉴伪技术提出了更高要求。在这一领域，基于预训练 NN 特征的方法已成为提升检测系统性能的关键技术。深度伪造音频鉴伪技术的精准度在很大程度上依赖于其底层特征的表达能力，从表 2 可以看出，基于预训练 NN 特征的方法由于具备底层特征表达能力，且预训练 NN 模型已经接触过更多不同类型的音频，因此其泛化性能力得到了显著的改善。

SSL 模型，如 wav2vec 2.0 和 WavLM，通过在大量未标记的音频数据上进行训练，学习到了能够捕捉音频信号深层特征的代表。文献[57]展示了 wav2vec 2.0 模型在伪造音频生成检测中的有效性，而文献[58]则进一步证明了 WavLM 模型在处理全栈语音任务时的优势。这些模型通过预测隐藏单元的方式来学习音频的表示，从而在伪造音频生成检测中展现出了卓越的性能。文献[59]使用 SSL 方法，训练模型识别音频中的伪造成分，这使模型能够在没有大量标注数据的情况下，学习到区分真实和伪造语音的特征，极大地提升了检测系统在面对未知音频合成系统时的泛化能力。

除了通用的 SSL 模型，特定任务训练的模型也在伪造音频鉴定中发挥了重要作用。文献[60]利用了 Whisper 模型，这是一个专门为自动语音识别 (automatic speech recognition, ASR) 任务而训练的模型，结果显示，该模型在伪造音频鉴定中也展现了优异的性能。这表明，针对特定任务优化的模型，通过专注于音频信号的某些方面，可以为伪造音频鉴定提供有价值的信息。

最新的研究趋势是将多个预训练模型的特征进行融合，以构建更全面和鲁棒的

检测系统。文献[61]提出了一种多视图特征融合方法，该方法结合来自不同模型的特征，如wav2vec 2.0、WavLM、Hubert和Whisper等，提高检测系统的性能。这种融合策略不仅提高了检测精度，还增强了系统对不同音频生成技术的适应性。

尽管当前的预训练模型在伪造音频鉴伪中已经取得了显著的成果，但仍面临诸多挑战。未来的研究将继续探索如何更有效地利用这些预训练模型，包括开发新的模型结构、训练策略以及特征融合技术。此外，随着伪造音频生成技术的不断演进，检测系统也需要不断更新，以应对新的伪造手段。

综上所述，预训练NN特征的方法在伪造音频鉴伪领域具有巨大的潜力和价值。通过结合SSL、特定任务训练模型以及多模型融合策略，研究人员构建出更精准和鲁棒的检测系统，以有效应对日益复杂的伪造音频生成挑战。然而，手工设计特征在一定程度上限制了模型的性能，为解决这一问题，端到端模型应运而生。端到端模型能够直接处理原始波形或频谱图，不需要手工设计特征，从而进一步提高了检测性能。

2.2 基于端到端模型的鉴伪技术

端到端模型在伪造音频鉴伪中扮演着至关重要的角色，它们通过直接从原始音频信号处理到最终决策输出的连贯流程，简化了检测过程并提升了效率。端到端模型能够适应各种复杂的音频环境，通过一体化学习自动提取特征并进行智能分类，从而有效识别音频中的欺诈行为。

为应对不断进步的伪造音频生成技术，对这些端到端模型进行细致的优化成为提升鉴伪性能的核心。此外，提高模型的泛化能力也是研究的重点，这涉及使用多源

数据进行训练，以及开发能够适应新场景和新类型的伪造音频生成的模型。通过持续学习和自适应学习策略，端到端模型能够不断吸收新知识，保持对新兴伪造技术的敏感性。实时检测强化技术同样重要，它确保了检测系统能够在实际应用中快速响应，对实时音频流进行高效分析。这通常要求模型在保持高精度的同时，具备较低的时延和资源消耗。

综上，对深度伪造音频鉴伪技术中基于端到端模型的鉴伪方法进行比较，结果见表3。表3详细比较了基于端到端模型的不同鉴伪方法在架构、是否采用数据增强、是否需要额外训练数据、是否需要自监督预训练模型等方面的差异。深入分析这些方法可以帮助研究人员更好地理解它们的特点和局限性。

2.2.1 模型优化

从表3可以了解到不同模型优化方法的优缺点，进而直接在时间和频域两个方向构造优化模块。研究人员设计出了能够自动学习并适应复杂欺骗攻击的模型，这些模型通过采用正交卷积和图注意力网络（graph attention network，GAT）等高级优化技术，增强了特征表达的深度和模型的辨别能力。RawNet2^[44]通过直接从原始波形提取特征，突破了传统手工特征提取的局限，展现了其在竞争力上的优势。RawGAT-ST^[15]和AASIST^[46]等模型通过图注意力机制进一步探索了不同子带和时间间隔之间的关系，以提高欺骗攻击的检测性能。

为了应对伪造音频长度不一的问题，HM-Conformer^[62]通过分层池化和多级分类标记聚合方法，有效整合了不同序列长度中的欺骗证据，提升了检测的整体性能。为了增强模型的泛化性并减少计算资源的

表3 基于端到端模型的鉴别方法比较

类别	算法	年份	特征	分类器	是否采用数据增强	是否需要额外训练数据	是否需要监督预训练模型	是否需要人工编辑特征	是否针对解决泛化问题	备注
模型优化	RawNet2 ^[44]	2020	原始波形	RawNet2	×	✓	×	×	×	从原始波形提取说话人特征
	RawGAT-ST ^[15]	2021	原始波形	GAT-ST	√(channel mask)	×	×	×	×	RawGAT-ST 模型
	AASIST ^[47]	2022	原始波形	HS-GAL、MGO	×	×	×	×	×	通过端到端方式实现对广泛欺骗攻击检测
	To-RawNet ^[63]	2023	原始波形	RawNet2、TCN	×	×	×	×	✓	通过正交卷积减少过滤器间的相关性
	light-DARTS ^[64]	2022	wav2vec 2.0	light-DARTS	×	×	✓	×	✓	降低手动参数设置的影响
	SpecRNet ^[45]	2021	频谱	SpecNet	×	×	×	✓	✓	减少内存消耗
	PS3DT ^[65]	2024	频谱	Transformer	×	×	✓	×	✓	通过处理梅尔谱的图像块来提高检测性能
	HM-Conformer ^[62]	2023	LFCC	HM-Conformer	✓	×	×	×	×	分层池化方法和多级分类标记聚合方法
	Multi-Fusion ^[66]	2024	WavLM	MFA、ASP	✓	×	✓	×	✓	通过时间池化和层间池化聚合多层表示、有效捕获音频特征的互补信息
	CD-ADD ^[67]	2024	wav2vec 2.0 Whisper	NN-based	✓	✓	✓	×	✓	跨区域评估和少样本学习提高模型泛化能力
泛化提升	DSVAE ^[68]	2023	频谱	VAE	×	×	×	×	✓	泛化到未知的语音合成器
	ONE-CLASS ^[70]	2024	wav2vec 2.0	AASIST	✓ (RawBoost)	×	✓	×	✓	one-class 分类、教师-学生框架
	RWM ^[16]	2024	wav2vec 2.0	S-CNN	✓	✓	✓	×	✓	持续学习,自适应调整梯度方向解决泛化
	RAD ^[69]	2024	WavLM	MFA	×	✓	✓	×	✓	检索相似样本,增强检测性能,解决泛化问题
	RawSpectrogram ^[17]	2023	Raw Spectrogram	LSTM、AASIST、RECAPA	×	×	×	✓	×	EmbNet 将音频转为频谱并通过 RNN 实现流式
实时检测增强	Real-Time ^[18]	2024	原始波形	TDNN、ResNet、LCNN	×	×	×	×	×	专注于开发适用于通信平台的实时系统

需求,研究人员探索了基于SSL模型的方法,如SpecRNet^[45]。这些方法利用大量未标记数据来学习深层的语音信号表示,减少了对手工设计超参数的依赖,并通过自适应的数据采样方法来提高模型对未知攻击的鲁棒性。TO-RawNet^[63]引入了正交卷积和时间卷积网络(temporal convolutional network, TCN),减少了过滤器间的相关性,并捕获了语音信号中的长期依赖性,从而在ASVspoof 2019数据集上实现了较低的等错误率(equal error rate, EER)。Light-DARTS^[64]自动学习深度语音表示,减少了手动参数设置的需求,并在ASVspoof 2019 LA数据集上取得了优异的EER。PS3DT^[65]通过处理梅尔频谱图的图像块,并使用Transformer神经网络,进一步提高了合成语音检测的性能。

此外,Multi-Fusion^[66]模型通过时间池化和层间池化聚合多层表示,有效捕获了音频特征的互补信息,提升了检测精度。CD-ADD^[67]数据集的跨域评估和少样本学习提高了模型在多样化数据集上的泛化能力,这对于确保实际应用中的有效性至关重要。这些技术的融合和发展预示着未来端到端模型将更加智能和高效,能够应对日益复杂的欺骗攻击,为伪造音频鉴伪提供更加坚实的安全保障。

2.2.2 泛化提升

在伪造音频鉴伪研究领域,提升模型对未知语音合成器的泛化能力是一个至关重要的挑战。随着语音合成技术的迅猛发展,未知合成器的多样性和复杂性日益增加,这对现有检测模型的泛化能力提出了更高的要求。为了有效应对这一挑战,研究人员已经提出了多种创新的解决方案。

首先,伪造音频鉴伪模型在面对未知

语音合成器时常常表现出泛化能力不足,这是因为在模型训练阶段,很难涵盖所有可能的伪造类型,导致模型在实际应用中难以识别新出现的伪造手段。针对这一问题,文献[68]提出了一种利用变分自编码器(variational autoencoder, VAE)的方法,该方法通过学习音频数据的潜在分布,生成新的音频样本,增强了模型对未知合成器的泛化能力。其次,文献[16]探讨了在不断变化的伪造环境中,如何通过持续学习策略使模型持续吸收新知识而不丢失旧知识,提出了一种自适应调整梯度方向的方法,这不仅使模型在接收新数据时能够进行有效学习,而且保持了对已有数据的记忆,进一步提升了模型的泛化性。再次,面对训练数据有限的挑战,文献[69]提出了一种基于检索的策略,通过检索与测试样本在特征空间中相似的样本,为模型提供了额外的信息,从而增强了模型对未知攻击的检测性能。最后,针对缺乏大量标记欺诈样本的训练问题,文献[70]介绍了一种结合One-Class分类和教师-学生框架的方法,该方法利用教师模型的先验知识来指导学生模型的学习,提高了模型对未知欺诈攻击的泛化能力。

从表3可以看出,针对泛化问题,自监督预训练NN起到了关键作用,利用自监督预训练NN的底层特征表达能力作为底座可以很好地提升模型的泛化能力。这些文献中提出的方法,不仅为解决音频欺诈检测中的泛化性问题提供了创新的解决方案,也为构建更加精准和鲁棒的检测系统奠定了理论基础,并提供了技术支持。随着技术的持续进步,这些策略在未来的音频安全领域中有望发挥更加重要的作用。

2.2.3 实时检测增强

从表3可以看出,实时伪造音频鉴伪

是通信安全领域的一个关键分支，它要求检测系统能够快速准确地分析音频流，并在通信过程中实时识别出欺诈行为。随着技术的发展，研究人员已经在此方向上取得了显著进展，特别是在实时性的特殊关注和优化方面。

文献[17]提出的RawSpectrogram方法代表了这一领域的一个创新步骤，该方法通过将原始音频信号转换为频谱图样对象，不仅减少了模型参数，还加快了在实时环境中预测概率的更新速度。RawSpectrogram的兼容性强，能够处理包括原始音频信号和经过时频变换的数据在内的多种输入类型，为实时音频处理提供了灵活性。文献[18]则专注于开发和测试适用于通信平台的实时深度伪造音频鉴伪系统，该研究基于ResNet和轻量级卷积神经网络（lightweight convolutional neural network, LCNN）架构实现的模型，在ASVspoof 2019数据集上进行了训练，并在真实的Teams会议环境中进行了测试，尽管在实时通信场景中存在性能下降的挑战，但该研究展示了深度伪造音频鉴伪模型部署的可行性，并提出了提升模型泛化能力的策略。

综上所述，实时性方向的伪造音频鉴伪是一个活跃且快速发展的领域，它不仅要求研究者关注模型的准确性，还要求关注模型在实时环境中的表现，包括适应性和响应速度。通过结合深度学习模型和实时处理技术，该领域在通信安全方面可以实现更高层次的防护，确保实时通信的完整性和安全性。

3 结束语

伪造音频生成技术的迅速发展使伪造

音频越来越难以与真实音频区分，未来的伪造技术将利用LLM和多模态生成模型，实现生成速度更快、效果更自然的伪造音频，这将使伪造音频的实时生成和动态调整成为可能，同时结合视频和文本等模态，使伪造内容更加具备欺骗性。为应对这些挑战，未来的伪造音频鉴伪系统需要不断进化，以提高先进技术的检测能力，并具备更强的泛化和自适应能力，以保持对新型伪造技术的有效性。

随着技术的发展，伪造音频鉴伪技术也必须不断进化以保持同步，这要求研究人员对伪造音频生成技术与鉴伪技术的现状和最新进展要有深入的了解和前瞻性的展望。伪造音频鉴伪技术的进展是显而易见的，但挑战也同样显著。数据集的局限性制约了检测模型的训练和评估；技术的快速迭代要求检测模型能够迅速适应新类型的伪造音频；实时性的要求则对检测系统的计算速度和准确性提出了更高标准。为了应对这些挑战，未来的研究需要在以下方向进行深入探索和创新。

首先，通过融合多模态信息，如音频、文本和视频等多种信息源，可以构建一个更全面的检测框架。这种融合不仅能够增强检测的准确性和可靠性，还能够提供更加丰富的上下文信息来辅助判断，从而在复杂环境中实现更深层次的分析 and 理解。

其次，提高检测模型的可解释性是至关重要的。这不仅有助于提升用户对检测结果的信任度，而且使模型的决策过程更加透明，有助于专家进行审核和干预。此外，特征提取方法的改进也是提高检测精度的关键。探索基于频域和时域的混合特征，结合先进的信号处理技术，能够使模型更好地捕捉伪造音频的细节和异常模式。

最后, 开发具有强泛化能力的检测模型, 能够适应不同类型和风格的伪造音频, 是提升鉴伪技术鲁棒性的关键, 通过持续学习和自适应学习等策略, 提高模型对未知伪造手段的泛化能力, 确保检测系统在面对新的伪造技术和手段时, 依然能够保持高效和准确的检测性能。

综上所述, 伪造音频生成技术的快速发展要求研究人员在鉴伪技术上进行不断的创新和改进。通过深入研究和探索多模态信息融合、提高模型可解释性、改进特征提取方法, 以及增强模型的泛化能力和实时检测性能, 研究人员可以更好地应对伪造音频生成带来的挑战, 确保深度伪造音频鉴伪技术的安全性和有效性。

参考文献:

- [1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
- [2] TANG Z, WANG D, HUANG S, et al. Full-text error correction for chinese speech recognition with large language model[C]//Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad: IEEE, 2025: 1-5.
- [3] ANASTASSIOU P, CHEN J W, CHEN J T, et al. Seed-TTS: a family of high-quality versatile speech generation models[EB]. arXiv preprint, 2024, arXiv: 2406.02430.
- [4] SIUZDAK H. Vocos: closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis[EB]. arXiv preprint, 2024, arXiv: 2306.00814.
- [5] SIGURGEIRSSON A, KING S. Controllable speaking styles using a large language model[C]//Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 10851-10855.
- [6] FENG X C, YOSHIMOTO A. Llama-VITS: enhancing TTS synthesis with semantic awareness[EB]. arXiv preprint, 2024, arXiv: 2404.06714.
- [7] KANEKO T, KAMEOKA H, TANAKA K, et al. CycleGAN-VC2: improved cycleGAN-based non-parallel voice conversion[C]//Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 6820-6824.
- [8] CAI Z X, LI M. Invertible voice conversion with parallel data[C]//Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 10041-10045.
- [9] YAO J X, YANG Y G, LEI Y, et al. PromptVC: flexible stylistic voice conversion in latent space driven by natural language prompts[C]//Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 10571-10575.
- [10] QI T H, ZHENG W M, LU C, et al. PAVITS: exploring prosody-aware VITS for end-to-end emotional voice conversion[C]//Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 12697-12701.
- [11] AL-KHAZRAJISH, SALEH H H, KHALID A

- I, et al. Impact of deepfake technology on social media: detection, misinformation and societal implications[J]. The Eurasia Proceedings of Science Technology Engineering and Mathematics, 2023, 23: 429–441.
- [12] 王健宗, 张旭龙, 姜桂林, 等. 基于分层联邦框架的音频模型生成技术研究[J]. 智能系统学报, 2024, 19(5): 1331–1339.
- WANG J Z, ZHANG X L, JIANG G L, et al. Research on audio model generation technology based on a hierarchical federated framework[J]. CAAI Transactions on Intelligent Systems, 2024, 19(5): 1331–1339.
- [13] TODISCO M, DELGADO H, EVANS N. Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification[J]. Computer Speech & Language, 2017, 45: 516–535.
- [14] SAHIDULLAH M, KINNUNEN T, HANILÇI C. A comparison of features for synthetic speech detection[C]//Proceedings of the Interspeech 2015. Dresden: ISCA, 2015: 2087–2091.
- [15] TAK H, JUNG J W, PATINO J, et al. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection[EB]. arXiv preprint, 2021, arXiv: 2107.12710.
- [16] ZHANG X H, YI J Y, WANG C L, et al. What to remember: self-adaptive continual learning for audio deepfake detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2024: 19569–19577.
- [17] GRINBERG P, SHIKHOV V. RawSpectrogram: on the way to effective streaming speech anti-spoofing[J]. IEEE Access, 2023, 11: 109928–109938.
- [18] MATHEW J J, AHSAN R, FURUKAWA S, et al. Towards the development of a real-time deepfake audio detection system in communication platforms[EB]. arXiv preprint, 2024, arXiv: 2403.11778.
- [19] 唐浩彬, 张旭龙, 王健宗, 等. 表现性语音合成综述[J]. 大数据, 2023, 9(6): 53–71.
- TANG H B, ZHANG X L, WANG J Z, et al. A survey of expressive speech synthesis[J]. Big Data Research, 2023, 9(6): 53–71.
- [20] WANG Y X, SKERRY-RYAN R J, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[C]//Proceedings of the Interspeech 2017. Stockholm: ISCA, 2017: 4006–4010.
- [21] OORD A V D, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[EB]. arXiv preprint, 2016, arXiv: 1609.03499.
- [22] LI N H, LIU S J, LIU Y Q, et al. Neural speech synthesis with transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii: AAAI, 2019: 6706–6713.
- [23] LE M, VYAS A, SHI B, et al. Voicebox: text-guided multilingual universal speech generation at scale[J]. Advances in neural information processing systems, 2023, 36: 14005–14034.
- [24] 施昊翔, 张旭龙, 王健宗, 等. 情感语音合成综述[J]. 大数据, 2024, 10(5): 56–73.
- SHI H X, ZHANG X L, WANG J Z, et al. A survey of emotional speech synthesis[J]. Big Data Research, 2024, 10(5): 56–73.
- [25] 李鹏程, 张旭龙, 王健宗, 等. 面向非平行语料的语音转换技术综述[J]. 大数据, 2024, 10(3): 65–81.
- LI P C, ZHANG X L, WANG J Z, et al. A survey of voice conversion based on non-parallel data[J]. Big Data Research, 2024, 10(3): 65–81.
- [26] KAMEOKA H, KANEKO T, TANAKA K, et al. StarGAN-VC: non-parallel many-to-many voice conversion using star

- generative adversarial networks[C]//Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE Press, 2018: 266–273.
- [27] SHAN S Y, LI Y, BANERJEE A, et al. Phoneme hallucinator: one-shot voice conversion via set expansion[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, Canada, 2024: 14910–14918.
- [28] CHOI H Y, LEE S H, LEE S W. DDDM-VC: decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2024: 17862–17870.
- [29] DELGADO H, TODISCO M, SAHIDULLAH M, et al. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements[C]//Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2018). California: ISCA, 2018: 296–303.
- [30] TODISCO M, WANG X, VESTMAN V, et al. ASVspoof 2019: future horizons in spoofed and fake audio detection[C]//Proceedings of the Interspeech 2019. Arizona: ISCA, 2019: 1008–1012.
- [31] LIU X C, 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.
- [32] MUCKENHIRN H, MAGIMAI-DOSS M, MARCEL S. Presentation attack detection using long-term spectral statistics for trustworthy speaker verification[C]//Proceedings of the 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). Piscataway: IEEE Press, 2016: 1–6.
- [33] FRANK J, SCHÖNHERR L. WaveFake: a data set to facilitate audio deepfake detection[EB]. arXiv preprint, 2021, arXiv: 2111.02813.
- [34] REIMAO R, TZERPOS V. FoR: a dataset for synthetic speech detection[C]//Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). Piscataway: IEEE Press, 2019: 1–10.
- [35] YI J Y, BAI Y, TAO J H, et al. Half-truth: a partially fake audio detection dataset[C]//Proceedings of the Interspeech 2021. Virtual: ISCA, 2021: 1654–1658.
- [36] XIE Z Y, LI B H, XU X N, et al. Fake-Sound: deepfake general audio detection [C]//Proceedings of the Interspeech 2024. Buenos Aires: ISCA, 2024: 1–5.
- [37] YI J, TAO J, DU J, et al. ADD: audio deepfake detection dataset[C]//Proceedings of the International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2021.
- [38] YI J Y, FU R B, TAO J H, et al. ADD 2022: the first audio deep synthesis detection challenge[C]//Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 9216–9220.
- [39] YI J, TAO J, FU R, et al. ADD 2023: the second audio deepfake detection challenge[C]//Proceedings of the IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis. Piscataway: IEEE Press, 2023: 125–130.
- [40] MÜLLER N M, KAWA P, CHOONG W H, et al. MLAAD: the multi-language audio anti-spoofing dataset[C]//Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2024: 1–7.

- [41] BAEVSKI A, ZHOU Y, MOHAMED A, et al. Wav2vec 2.0: a framework for self-supervised learning of speech representations[C]//Neural Information Processing Systems (NeurIPS). Piscataway: IEEE Press, 2020: 12449–12460.
- [42] HSU W N, BOLTE B, TSAI Y 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.
- [43] CHEN S Y, WANG C Y, CHEN Z Y, et al. WavLM: large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505–1518.
- [44] JUNG J W, KIM S B, SHIM H J, et al. Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms[C]//Proceedings of the Interspeech 2020. Virtual: ISCA, 2020: 1496–1500.
- [45] GUAN B C, ZHANG J N, SETHARES W A, et al. Spectral domain convolutional neural network[C]//Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 2795–2799.
- [46] JUNG J W, HEO H S, TAK H, et al. AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks[C]//Proceedings of the ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6367–6371.
- [47] FU Q C, TENG Z W, WHITE J, et al. FastAudio: a learnable audio front-end for spoof speech detection[C]//Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 3693–3697.
- [48] WANG X, YAMAGISHI J. Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? [C]//Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 10311–10315.
- [49] FAN C H, XUE J, TAO J H, et al. Spatial reconstructed local attention Res2Net with F0 subband for fake speech detection[J]. Neural Networks, 2024, 175: 106320.
- [50] XUE J, FAN C H, YI J Y, et al. Learning from yourself: a self-distillation method for fake speech detection[C]//Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1–5.
- [51] KIM J, BAN S M. Phase-aware spoof speech detection based on Res2Net with phase network[C]//Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1–5.
- [52] WANG C L, YI J Y, TAO J H, et al. Detection of cross-dataset fake audio based on prosodic and pronunciation features[C]//Proceedings of the INTERSPEECH 2023. Florida: ISCA, 2023: 3844–3848.
- [53] LIU R, ZHANG J H, GAO G L, et al. Betray oneself: a novel audio DeepFake detection model via mono-to-stereo conversion[C]//Proceedings of the INTERSPEECH 2023. Florida: ISCA,

- USA, 2023: 3999–4003.
- [54] FAN C H, DING M M, TAO J H, et al. Dual-branch knowledge distillation for noise-robust synthetic speech detection[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, 32: 2453–2466.
- [55] SINGH YADAV A K, BHAGTANI K, SALVI D, et al. FairSSD: understanding bias in synthetic speech detectors[C]// *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Piscataway: IEEE Press, 2024: 4418–4428.
- [56] KHAN A, MALIK K M, NAWAZ S. Frame-to-utterance convergence: a spectra-temporal approach for unified spoofing detection[C]// *Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE Press, 2024: 10761–10765.
- [57] ROSELLO E, GOMEZ-ALANIS A, GOMEZ A M, et al. A conformer-based classifier for variable-length utterance processing in anti-spoofing[C]// *Proceedings of the INTERSPEECH 2023*. Florida: ISCA, 2023: 5281–5285.
- [58] GUO Y L, HUANG H F, CHEN X, et al. Audio deepfake detection with self-supervised WavLM and multi-fusion attentive classifier[C]// *Proceedings of the ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE Press, 2024: 12702–12706.
- [59] ITO A, HORIGUCHI S. Spoofing attacker also benefits from self-supervised pre-trained model[C]// *Proceedings of the INTERSPEECH 2023*. Florida: ISCA, 2023: 5346–5350.
- [60] KAWA P, PLATA M, CZUBA M, et al. Improved DeepFake detection using whisper features[C]// *Proceedings of the INTERSPEECH 2023*. Florida: ISCA, 2023: 4009–4013.
- [61] YANG Y J, QIN H C, ZHOU H, et al. A robust audio deepfake detection system via multi-view feature[C]// *Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE Press, 2024: 13131–13135.
- [62] SHIN H S, HEO J, KIM J H, et al. HM-CONFORMER: a conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods[C]// *Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE Press, 2024: 10581–10585.
- [63] WANG C L, YI J Y, TAO J H, et al. TO-RawNet: improving RawNet with TCN and orthogonal regularization for fake audio detection[C]// *Proceedings of the INTERSPEECH 2023*. Florida: ISCA, 2023: 3137–3141.
- [64] WANG C L, YI J Y, TAO J H, et al. Fully automated end-to-end fake audio detection[C]// *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. New York: ACM, 2022: 27–33.
- [65] SINGH YADAV A K, XIANG Z Y, BHAGTANI K, et al. PS3DT: synthetic speech detection using patched spectrogram transformer[C]// *Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA)*. Piscataway: IEEE Press, 2023: 496–503.

[66] GUO Y L, HUANG H F, CHEN X, et al. Audio deepfake detection with self-supervised WavLM and multi-fusion attentive classifier[C]//Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 12702-12706.

[67] LI Y A, ZHANG M, REN M X, et al. Cross-domain audio deepfake detection: dataset and analysis[EB]. arXiv preprint, 2024, arXiv: 2404.04904. .

[68] SINGH YADAV A K, BHAGTANI K, XIANG Z Y, et al. DSVAE: disentangled representation learning for synthetic speech detection[C]//Proceedings of the 2023 International Conference on Ma-

chine Learning and Applications (ICMLA). Piscataway: IEEE Press, 2023: 472-479.

[69] KANG Z H, HE Y Y, ZHAO B T, et al. Retrieval-augmented audio deepfake detection[C]//Proceedings of the 2024 International Conference on Multimedia Retrieval. New York: ACM, 2024: 376-384.

[70] LU J Z, ZHANG Y X, WANG W C, et al. One-Class knowledge distillation for spoofing speech detection[C]//Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 11251-11255.

作者简介



曾志平（1989-），男，平安科技（深圳）有限公司高级算法工程师，主要研究方向为语音识别、语音合成、音频欺诈检测以及机器学习和深度学习方法在人工智能领域的应用。



张旭龙（1988-），男，博士，平安科技（深圳）有限公司高级算法研究员，IEEE、中国自动化学会以及中国计算机学会会员，联邦数据与联邦智能专委会委员，主要研究方向为语音合成、语音转换、音频驱动虚拟人生成、音乐信息检索以及机器学习和深度学习方法在人工智能领域的应用。



瞿晓阳（1988-），男，博士，平安科技（深圳）有限公司前沿机器学习算法分组负责人，主要研究方向为机器学习、大数据、体系结构、语音语义分析、零样本和小样本学习、高性能计算与存储等。



肖春光（1976–），男，深圳市宝安区教育事业发展中心教育信息技术副研究员，主要研究方向为教育数字化、网络安全、信息化教学创新。



王健宗（1983–），男，博士，平安科技（深圳）有限公司副总工程师、中国计算机学会（CCF）杰出会员，主要研究方向为大模型、联邦学习和深度学习等。

收稿日期: 2024-10-21
通信作者: 王健宗, jzwang@188.com
基金项目: 广东省重点领域研发计划“新一代人工智能”重大专项(No.2021B0101400003)
Foundation Item: Guangdong Province Key Areas of Research and Development Plan “New Generation of Artificial Intelligence” Major Project (No.2021B0101400003)