ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 347－370
DOI:10. 16337/j. 1004-9037. 2026. 02. 005

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

语音深度伪造溯源技术研究现状及展望

张雄伟[1]，张　强[1]，孙　蒙[1]，杨吉斌[1]，李毅豪[1]，葛晓义[2]

(1. 陆军工程大学,南京 210007;2. 信息支援部队工程大学,武汉 430000)

摘　要：随着生成式人工智能技术的快速发展,语音深度伪造技术日益精进,其生成的语音在听感上已难辨真假,给信息安全、司法取证和社会互信带来严峻挑战。传统的语音伪造检测重点在于解决语音“真/假”的二元分类问题。然而,在复杂的安全对抗与取证场景中,仅判定语音的真或假已无法满足追根溯源、厘清责任的需求。本文聚焦“语音伪造溯源”这一前沿课题,系统综述了国内外当前的研究进展。首先,构建了一个层级化的语音伪造溯源任务体系,明确界定了伪造方法溯源、源说话人溯源和模型逆向这3个子任务的内涵。然后,从生成模型的基本原理、语音信号的声学特性等角度,阐述了各子任务可行的核心机理;区分体系架构、训练策略等不同维度,系统地梳理了各子任务的研究现状、主流方法及技术演进路径。最后,总结了当前研究面临的开放世界溯源、复杂信道条件下溯源等关键挑战,展望了面向语音深度伪造反制的主动溯源等未来的发展方向,旨在为构建更完善的语音安全防御体系提供参考。

关键词：语音深度伪造;语音伪造方法溯源;源说话人溯源;模型逆向;开放集识别

中图分类号：TN912　　**文献标志码：**A

引用格式：张雄伟,张强,孙蒙,等.语音深度伪造溯源技术研究现状及展望[J].数据采集与处理,2026,41(2):347-370. ZHANG Xiongwei,ZHANG Qiang,SUN Meng, et al. Speech deepfake attribution: The state of the art and prospects[J]. Journal of Data Acquisition and Processing,2026,41(2):347-370.

引　　言

随着人工智能技术的迅猛发展,语音伪造已成为数字时代最具破坏力的威胁之一[1]。从2017年WaveNet[2]的出现,到2024年VALL-E[3]等端到端语音克隆模型的普及,伪造者只需几秒钟录音即可生成高度逼真的假语音,用于诈骗、选举操纵和假新闻传播。根据DeepMedia的统计,2025年社交平台上共享的深度伪造语音片段已超过50万条,造成全球经济损失超百亿美元,到2026年,这一数字预计将翻倍[4]。例如,在2024年香港一桩伪造语音诈骗案中,攻击者利用深度伪造音视频在电话会议中模仿Arup公司高管,欺骗一名Arup的香港员工将约2亿港元转入欺诈账户[5];2025年一名受害者收到冒充财务经理的人工智能克隆语音信息,指示其转移加密货币,造成约1 850万美元损失[6]。此外,深度伪造语音还被用于捏造名人言论,扰乱社会秩序。近年来,针对国家政要等的伪造语音被广泛传播,损害了重要人物的声誉,对选举和外交产生一定程度的负面影响。这些事件不仅侵蚀公众信任,还放大社会不稳定风险。由此可见,语音深度伪造技术的滥用将对个人、企业和国家造成严重的影响。

为应对深度伪造语音带来的挑战,国内外研究者着重研究语音伪造检测等针对伪造语音的反制技术。然而,语音伪造检测主要聚焦于二元分类,即判断语音为真还是假。早期方法主要以梅尔频率倒

基金项目：国家自然科学基金(62371469,62071484)。
收稿日期：2026-01-10;**修订日期：**2026-02-27

谱系数等特征为输入,以融合高斯混合模型和隐马尔可夫模型的混合模型等机器学习框架作为检测器,实现基本筛查[7]。近年来,深度学习驱动的检测器,如改进版轻量化卷积神经网络(Improved lightweight convolutional neural network, ILCNN)[8]和RawNet[9],在ASVspoof挑战赛中将等错误率降至5%以下,并通过与视频等其他模态检测的融合进一步提升了鲁棒性[10]。但是,在复杂对抗环境中,这种"看门人"式的二元检测已显不足,攻击者可通过对抗攻击技术轻松绕过检测器[11]。同时,当一条伪造语音导致了实际危害(如诈骗得手),司法机关或安全团队不仅需要确认它是假的,更迫切需要回答以下问题:这条语音是使用哪种技术生成的?攻击者是谁,原始声音来源于哪里?是否能锁定具体的作案工具,比如模型实例?这些问题构成了语音伪造溯源的核心范畴,也凸显了从检测向溯源的范式转变:溯源不仅是"事后审计",更是多媒体取证的核心,是从"检测伪造"跃迁到"解构伪造"的关键[12]。因此,以伪造语音溯源为重点的语音伪造反制技术逐渐成为当下的研究热点。

语音伪造溯源旨在逆向追踪伪造链条[13],包括伪造方法溯源、源说话人溯源,以及模型逆向。伪造方法溯源多依赖伪造痕迹提取,源说话人溯源大多借鉴说话人验证框架,而语音伪造模型逆向研究尚未见公开学术报道。这一空白源于语音生成模型的高维非线性逆向的计算挑战和潜在隐私泄露的伦理风险。近年来,国内外研究机构也逐步开展了一些关于伪造方法溯源、源说话人溯源的挑战赛。在国外,由IEEE信号处理学会语音与语言技术委员会主办的源说话人溯源挑战赛,聚焦判定两段经语音转换处理后的语音是否来自同一个源说话人,即对"源说话人"做同/异说话人验证。此外,该赛事也将语音转换方法识别作为一个相关任务引入,用于识别具体的语音转换方法,并采用多任务学习范式同时完成源说话人验证与方法识别。这是目前唯一明确以"源说话人溯源"为核心任务的公开国际挑战赛,吸引了国内外大量队伍参赛。在国内,目前没有直接包含伪造语音溯源任务的赛事,唯一相近的是由厦门市数据管理局和鲸智实验室主办的第六届中国人工智能大赛。该赛事开展面向特定说话人确认场景的伪造语音检测任务,既要判断给定语音是否为假,还要判定是否是针对特定说话人的伪造语音,相当于"针对特定集合的源说话人溯源"。

语音伪造溯源不仅是检测的延伸,更是建立数字责任归属机制的关键。它要求从防御思维转向取证思维,从粗粒度的真伪判别转向细粒度的特征归因。尽管语音伪造检测领域已有全面综述,但伪造语音溯源领域的相关研究仍然碎片化,没有系统的综述文献。本文聚焦伪造方法溯源、源说话人溯源及模型逆向这3个语音伪造溯源的核心任务,系统理顺其与二元检测的层级逻辑关系,即检测作为"入口"提供伪造信号,溯源则追踪伪造链条,推断伪造方法、源说话人、模型实例等伪造环节,最终形成端到端框架。在此基础上,综述语音伪造溯源技术的研究现状,并指出语音伪造溯源的挑战和未来方向,旨在为智能语音处理及其安全应用研究提供路线图,推动语音鉴伪从"检测"向"溯源"的跃升。

1 语音伪造溯源概述

与语音伪造检测只关注语音真/假不同,语音伪造溯源的目标是在确认语音为伪造的前提下,进一步挖掘隐藏在伪造语音中的生成机制、身份信息或模型痕迹。为更好地阐述语音伪造溯源,本节着重介绍其研究基础,包括作为溯源对象的语音伪造、语音伪造溯源的任务层级及其与语音伪造检测的逻辑关系,以及常用数据集和评价指标。

1.1 语音伪造简介

作为溯源的对象,伪造语音是由相应的伪造方法生成的。生成过程可以抽象为

$$y=f(x,\theta) \tag{1}$$

式中:x为源说话人语音或者文本,$f(\cdot)$为语音伪造方法,θ为语音伪造方法采用的模型权重参数,y为生成的伪造语音。语音伪造方法主要分为语音转换(Voice conversion, VC)和文本到语音合成(Text to

speech synthesis,TTS)两大类:(1)当$f(\cdot)$属于TTS时,对应的x是文本,TTS依据输入文本生成特定说话人语音。不同流派的TTS架构具有不同的生成机理。比如,自回归模型Tacotron 2[14]采用逐帧预测的方法生成伪造语音,各帧伪造语音容易产生时序不连续性;流模型Parallel WaveGAN[15]采用并行技术高效生成语音,但谱图噪声明显;扩散模型Diff-TTS[16]采用逐步去噪的方法,伪造痕迹更隐蔽;基于变分自编码器(Variational auto-encoder,VAE)的变体VITS[17],融合了变分自编码器,音色转换更加灵活。这些独特的伪造痕迹构成了TTS溯源的"签名"。(2)当$f(\cdot)$属于VC时,对应的x是源说话人语音,语音转换负责将源说话人的语音转换为目标说话人的音色,同时保留语言内容。典型方法基于自编码器框架或生成对抗网络(Generative adversarial networks,GAN)架构。比如,采用自编码器框架的AutoVC[18],通过解耦内容与说话人特征实现VC。基于GAN框架的StarGAN-VC[19],通过生成对抗训练,隐藏源说话人身份,实现VC。不同的VC不仅会在语音中留下独特的伪造痕迹,成为方法溯源的签名,而且转换过程不彻底,会残留源说话人特征,为源说话人溯源提供了线索。同时,依据计算机视觉领域的相关研究,生成模型的权重参数可以通过模型逆向技术[20]进行估计,这也为语音伪造模型的参数逆向分析提供了方法参考。

1.2 溯源任务层级架构与逻辑关系

基于式(1)描述的伪造语音生成过程,伪造语音溯源可以抽象为

$$g(y)\in\{x,\theta,f(\cdot)\} \tag{2}$$

式中:$g(\cdot)$表示溯源管道,x、$f(\cdot)$、θ为需要溯源的对象。如图1所示,依据溯源对象,语音伪造溯源任务可以分为以下3类:

(1)对$f(\cdot)$的溯源,即语音伪造方法溯源,学术界也称之为语音伪造方法识别。此任务追溯"作案手法",是最基础的溯源层级。其目标是识别生成语音所使用的具体方法族、模型架构甚至具体的声码器类型,例如判断伪造语音是由WaveNet还是HiFi-GAN[21]生成的。如2.1节所述,不同方法在生成波形时会留下独特的频谱伪影或相位不连续性,构成了方法的"指纹",为语音伪造方法溯源提供了依据。

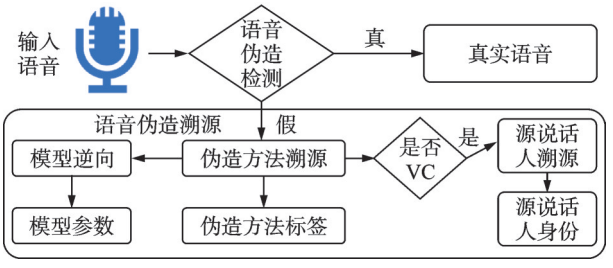

图1 语音伪造溯源层级架构及逻辑关系

Fig.1 Hierarchical architecture and logical relationships for speech forgery attribution

(2)对x的溯源,如果x是文本,对x的溯源相当于语音识别任务,不在本文讨论的语音伪造溯源研究范围内;如果x是源说话人语音,对x的溯源对应源说话人溯源任务。源说话人溯源任务追溯"源说话人身份",主要针对经过VC生成的伪造语音。VC旨在隐藏源说话人身份,但完美的解耦极其困难,VC语音中会残留源说话人的韵律习惯、特定发音方式等。源说话人溯源的目标是穿透VC的伪装,从转换后的语音中恢复或识别出源说话人的身份特征,甚至还原源语音,如使用X-vector嵌入矢量[22]来匹配韵律习惯。只有当$f(\cdot)$的溯源结果属于VC攻击时,才启动该任务。

(3)对θ溯源,相当于求解伪造方法使用的模型参数,追溯"具体实例",在计算机视觉领域,该技术被称为模型逆向工程[20]。考虑到语音数据的高维特性,语音伪造模型一般也是高维非线性的,对θ溯源也是溯源任务中难度最大的。这是最深层次,也是最具挑战性的溯源。其目标不仅是识别方法类别,而是试图通过分析伪造语音,逆向推断出生成该语音的具体模型参数。这通常需要更强的假设条件(如灰盒访问),在知识产权保护和深度取证中具有重要意义,同时也面临高伦理风险。

理解语音伪造溯源的关键在于理顺这3个子任务的层级性及其与真假二元检测的关系。由上述分析可以看出,语音伪造溯源标志着从真假二元检测向深度取证的演进[1],溯源不是孤立模块,而是检测后的递进式分层多粒度调查,旨在逆向解构伪造链条,实现从“是否伪造”到“如何、何人、何参数”的全面审计。如图1所示,以二元检测为“守门员”,只有当检测为伪造语音时,才触发多粒度溯源分支,按需开展从宽泛方法到深层逆向的递进式溯源:当语音伪造方法溯源结果属于VC时,才激活源说话人溯源任务;同时,了解使用了哪种VC模型,有助于更有针对性地进行源说话人特征的解耦分析。语音伪造方法溯源结果为模型逆向提供了模型结构信息,使模型逆向更有针对性。这一设计避免了对真实语音的无效溯源,同时也将检测模块提取的伪造嵌入或置信分数等信息作为初始特征馈入溯源模块,确保了计算资源的高效利用。若无真假检测作为前置筛选,溯源模型会在真实语音上产生错误标签,造成噪声放大与误导性预测。语音伪造方法溯源任务的中间表示能增强检测模型对不同伪造模式的判别力,从而提高其面对未知伪造方法时的鲁棒性,这可视为一种弱耦合的闭环或多任务协同机制。

1.3 数据集及评价指标

高质量的公开数据集与标准化的评价指标是评估语音伪造方法溯源及源说话人溯源技术性能的基石。由于未见关于伪造语音模型逆向的公开学术报道,本节主要聚焦语音伪造方法溯源及源说话人溯源这两个任务展开。不同于传统的二分类真伪检测,语音伪造方法溯源更侧重于多分类识别,源说话人溯源更侧重于跨域身份验证,即在伪造语音中检索源身份,因此其数据构成与评估体系具有独特性。

1.3.1 常用数据集

现有研究主要依赖两类数据集:一类是包含多种生成算法标签的综合性伪造数据集,适用于语音伪造方法溯源;另一类是包含成对源-目标语音的语音转换数据集,适用于源说话人溯源。

语音伪造方法溯源数据集主要包括ASVspoof系列数据集[23-26]、WaveFake数据集[27]、IEEE信号处理杯(Signal processing cup, SPC)挑战赛数据集[28]、中文伪造音频检测(Chinese fake audio detection, CFAD)数据集[29]。其中,ASVspoof系列数据集是该领域最权威的基准数据集,不仅用于反欺诈检测,还被广泛用于伪造方法溯源研究。例如,ASVspoof2019的逻辑访问(Logical access, LA)子集包含19种不同的TTS和VC算法,涵盖了从统计参数合成到端到端深度学习生成的多种范式,是评估闭集与开集溯源算法性能的首选数据。WaveFake数据集明确标注了具体的声码器架构,特别适用于研究基于“声码器指纹”的细粒度溯源任务。SPC挑战赛数据集来源于IEEE信号处理学会举办的“合成语音溯源”挑战赛。该数据集不仅关注区分具体的生成架构,还着重评估模型在面对经过压缩、滤波等不同强度后处理后的溯源鲁棒性。CFAD数据集为评估语音伪造溯源算法在中文语境下的泛化能力提供了关键的数据支持,并模拟了真实应用场景中的多种信道干扰。

源说话人溯源数据集主要包括语音转换挑战赛(Voice conversion challenge, VCC)系列数据集[30-32]、IEEE源说话人溯源挑战赛(Source speaker tracing challenge, SSTC)数据集[33]、第六届中国人工智能大赛数据集。其中,VCC系列数据集提供了“源语音-转换语音”配对样本,可以用来评估模型从转换语音中恢复或识别源身份的能力。SSTC数据集是目前国际上唯一明确以“源说话人溯源”为核心任务的数据集。该数据集聚焦于判定两段经VC处理后的语音是否源自同一说话人(即源说话人验证任务),同时还引入了语音转换方法识别作为关联任务,鼓励采用多任务学习策略同时解决身份溯源与算法溯源问题。

1.3.2 常用评价指标

鉴于语音伪造方法溯源(多分类/开集识别)与源说话人溯源(验证/检索)在任务范式上的显著差异,两者采用不同的评价指标体系。

语音伪造方法溯源任务旨在区分不同的生成算法或伪造工具。在实验评估中,除了关注模型在已知伪造方法上的分类精度,更需重点考察其在开集场景下对未知伪造方法的防御(拒识)能力,以及在开放世界中无监督条件下发现和区分新兴伪造类别的能力。常用的评价指标包括准确率、F1度量、开集分类率(Open-set classification rate, OSCR)[34]、聚类准确率、归一化互信息、调整兰德系数(Adjusted Rand index, ARI)等。在开集场景下,通常采用"未知类作为第$K+1$类"的策略计算F1度量,从而综合衡量模型对已知类的分类性能与对未知类的检测能力。OSCR通过绘制不同阈值下"已知类正确分类率"与"未知类误报率"的权衡曲线并计算曲线下面积,量化评估模型在保持对已知伪造方法高识别率的同时,有效拒识未知伪造方法的鲁棒性。聚类准确率、归一化互信息、ARI被用于在无标签场景下,衡量对不同未知类的聚类性能。

源说话人溯源任务本质上是在强信道干扰(语音转换)下的说话人确认或辨认问题,其核心在于评估模型穿透语音转换的伪装,并检索源身份的能力。常用指标包括等错误率(Equal error rate, EER)、最小检测代价函数(Minimum detection cost function, minDCF)[35]、Top-N准确率、余弦相似度等。相较于EER,minDCF引入了先验概率与错误代价权重。它更能反映系统在实际安全应用场景(如司法取证、访问控制)中的综合风险期望,是衡量溯源系统实用性的关键标准。

由于语音伪造领域的模型逆向研究尚处于空白阶段,接下来区分伪造方法溯源、源说话人溯源两个方面,阐述其可行的核心机理,总结研究现状以及局限性,刻画从宽泛审计到精准取证的发展路径,推动"解构伪造"范式。

2 语音伪造方法溯源

语音伪造方法溯源是伪造语音溯源的入口层,该任务假设不同生成机制会引入具有区分性的伪造痕迹。通过模式识别方法识别生成语音的算法族、架构或声码器,利用模型固有"指纹"实现伪造方法的逆向归因,即伪造方法溯源。本节依次介绍伪造方法的指纹形成机制、语音伪造方法溯源的主要算法。

2.1 伪造方法的指纹形成机理

语音伪造方法之所以能够被有效溯源,其根本原因在于:不同的语音生成算法在构建声学特征、拟合统计分布以及处理频谱残差的过程中,会不可避免地引入具有系统性差异的特征模式,即生成模型固有的"生成指纹"[36]。这些指纹并非随机噪声,而是由模型架构设计、训练目标函数以及推理采样机制的固有差异所决定的必然产物。因此,这些痕迹在信号处理与特征空间中表现出显著的可检测性与可分类性[37],为从二元检测向细粒度溯源的演进提供了理论支撑。下文从模型结构、训练策略、推理机制及参数空间等维度出发,系统阐述不同方法生成语音的物理与统计差异,揭示不同语音伪造方法指纹的形成机理。

2.1.1 模型结构差异引发的频谱与时域特征偏差

实证研究表明,不同生成范式会在频域、相位域或时域残差中形成独特的统计特征偏差,这些偏差构成了可被识别的"架构指纹"。以Tacotron 2、WaveNet为代表的自回归生成模型采用逐帧逐点预测的方式生成语音。由于其序列依赖性,模型存在逐帧预测的累积误差,这种误差倾向于在静音段或弱音区导致噪声堆积及谱图抖动[38]。此外,长时依赖建模的局限性使得模型在长语音生成中容易出现韵律不连贯或局部失真。以MelGAN[39]等为代表的生成对抗网络声码器,其反卷积或上采样操作常引发谱图不连续性,形成周期性的"棋盘格效应"[40]。同时,对抗训练机制往往以牺牲高频段保真度为代价来实现整体感知质量提升,导致高频谐波畸变等显著异常[15]。这些由生成器-判别器博弈机制引入的结构性伪影在频域能量分布中表现为模式固定的异常峰值或缺失。以DiffWave[41]等为代表的扩散模

型，尽管其生成的语音特征的整体统计分布表现较为平滑，但其依赖多步去噪的迭代采样过程容易引入残留相位噪声，导致生成的声学信号相干性降低[42]。此外，去噪步数的选择和噪声调度策略会直接影响生成语音的细粒度纹理特征，使得不同扩散模型在微观尺度上呈现出差异化的噪声残差模式。这些结构性差异在频域、相位域或时域残差中表现为模式固定的异常，为语音伪造方法溯源提供了可靠的物理线索。

2.1.2 训练策略导致的统计分布偏移

训练目标函数直接决定了生成语音的分布特性，不同的优化目标会在特征空间中留下独特的“训练痕迹”。GAN模型使用对抗损失作为训练目标函数，虽强迫生成分布逼近真实分布，但易在高频细节产生过拟合，致使高频伪影明显[39]。判别器的有限能力导致生成器学习到“欺骗性”而非“真实性”的特征，使得生成语音在某些频段（尤其是高频）呈现出非自然的规则性模式。流模型受限于严格的可逆架构设计，且通常假设隐变量服从高斯先验分布，这种强约束常导致生成的语音在低频能量分布上表现出过于规则化的人工痕迹。由于其必须保持雅可比矩阵的可逆性，模型倾向于生成统计上“安全”但缺乏自然变异性的低频成分。扩散模型基于噪声反演的策略，使其易在输出中保留微弱的过渡噪声纹理[41]，这种纹理会在长时累积下形成可识别的“去噪签名”。这些由“训练目标偏差”形成的特征构成了可被捕捉的“方法标签”的一部分，为基于统计学习的溯源方法提供了判别依据。

2.1.3 推理机制与参数空间的系统性差异

推理机制的差异进一步将微观痕迹固化为可重现的指纹模式。自回归生成表现为累积误差模式；Parallel WaveGAN[15]等采用的并行生成模式伴随相位连续性不足[43]；不同神经声码器因激励模型差异在高频段留下独特的差异性模式[37]。此外，训练集构成、参数规模及特征编码方式的不一致引入了参数空间系统性的统计偏差，具体表现为：不同模型对说话人身份的编码方式不同，导致说话人嵌入空间的几何结构差异；从声学特征到波形的逆变换过程中不可避免的信息损失模式；声码器在合成基频时引入的规则性伪影，以及由于相位信息难以精确建模而产生的系统性偏差。这些偏差在统计上具有良好的类间可分性，为基于深度学习的溯源模型提供了丰富的判别信号。

上述由生成机制引入的多维度差异构成了语音伪造方法溯源的物理基础。深度神经网络能够有效利用这些证据，在频谱、相位及能量包络等高维空间中挖掘隐含指纹[44]。通过学习棋盘格效应、谱图不连续性等局部伪影，频谱倾斜、能量分布偏移等长期统计差异，残留噪声、相位不连续等扰动模式，高频截断、谐波缺失等高频能量衰减规律等特征，溯源模型能够将伪造方法的指纹映射为可分类的嵌入向量，从而实现细粒度的伪造方法归属判定。基于生成指纹的模式识别研究不仅证实了不同伪造方法之间的可区分性，也为构建高精度的溯源取证系统提供了切实可行的技术路径。接下来介绍语音伪造方法溯源的主要算法。

2.2 语音伪造方法溯源的主要算法

语音伪造方法溯源，以模式识别技术为基础，以语音为输入，最终给出语音伪造方法类别的判断。定义X为用于语音伪造方法溯源的语音数据集，Y为对应的标签集。其一般流程如图2所示，即：在训练集上，对语音训练样本x进行特征提取，得到相应的样本特征$b(x)$；将各类别的样本特征和对应的真实标签y送入语音伪造方法溯源模型F进行训练，学习各类别的分布规律，得到训练良好的模型F^*，即

$$F^* = \underset{F}{\operatorname{argmin}} \{L(F(b(x)), y)\} \quad (3)$$

式中：$L()$为训练模型F使用的损失

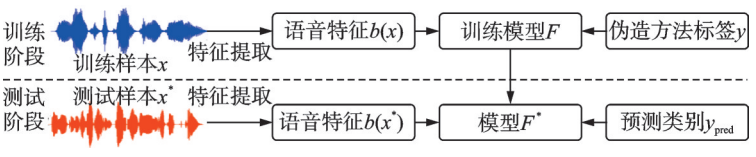

图2 语音伪造方法溯源的一般流程

Fig.2 General workflow of speech forgery method attribution

函数,$y\in[1,2,\cdots,C]$,C为训练集中的语音伪造方法类别数。在测试集上,将语音测试样本x^*的相应特征$b(x^*)$送入语音伪造方法识别模型F^*,模型输出关于测试样本的伪造方法类别预测标签y_{pred},即

$$y_{\text{pred}}=F^*(b(x^*)) \tag{4}$$

一个训练好的语音伪造方法溯源模型F^*,会将$x^*\in X$判断为其真实类别标签y^*,即$y_{\text{pred}}=y^*$。如果对于测试集中任意样本,其y^*满足:$y^*\in[1,2,\cdots,C]$,即任意测试样本的类别均在训练集中出现过,则称为闭集识别。如果$y^*\notin[1,2,\cdots,C]$,即除了训练集中见过的类别,测试集中还包含训练集中未见过的类别,则称为开集识别。

人工智能技术尤其是模式识别领域的突破,为语音伪造方法溯源提供了坚实的理论与技术支撑。由上述关于溯源机理的分析可知,不同的语音合成与转换算法会在生成语音中残留独特的“伪造指纹”,这些指纹在特征空间中具有可分性。依据模式识别器构建机理的不同,现有的语音伪造方法溯源研究可划分为基于传统机器学习的方法与基于深度学习的方法两大类。各类方法的逻辑演进与技术分支如图3所示。

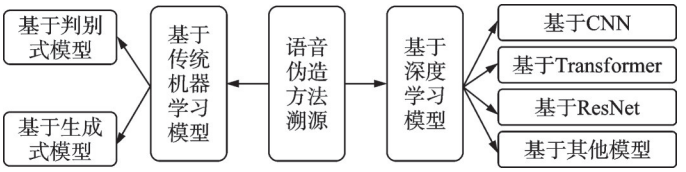

图3 语音伪造方法溯源技术分类框架

Fig.3 Taxonomy of speech forgery method attribution techniques

2.2.1 基于传统机器学习模型的语音伪造方法溯源

此类方法主要通过手工设计的声学特征,结合参数化的统计模型来挖掘伪造方法的类别属性。根据统计学习策略的差异,可进一步细分为基于判别式模型和基于生成式模型的溯源算法。

(1) 基于判别式模型的语音伪造方法溯源算法

该类算法使用的经典判别式模型包括随机森林(Random forest,RF)和支持向量机(Support vector machine,SVM)。判别式模型假设样本所在的特征表示空间被超平面分割成不同的区域,不同类样本占据不同的区域。由样本特征表示所在的区域能够确定样本所属的类别。Borrelli等[12]首次探究了线性SVM、径向基函数SVM和RF在语音伪造方法溯源领域中的性能表现。作者基于对信号时间演变的短时和长时分析(Short-term and long-term analysis,STLT),提出了一套音频描述符作为输入。以样本的STLT音频描述符为输入,训练RF或SVM,直接利用RF或SVM输出伪造方法类别识别结果。进一步地,作者探究了在原始语音上训练的SVM模型对经微信传输的Opus编解码语音的伪造方法识别鲁棒性。实验结果显示,在原始语音上训练的SVM模型对经Opus编解码的语音的伪造方法识别性能大幅下降。此外,为了应对开集条件下的语音伪造方法识别,作者提出了一种基于分布假设的开集识别方法,即使用已知类分布建模未知类分布。该方法假设未知类数据的分布规律可以从已知类数据中学习到,通过将部分已知类数据作为未知类数据的替代品,将原本包含C个已知类和1个未知类的开集识别问题转化为$C+1$类闭集识别问题,从而实现识别不同的已知类和未知类。

(2) 基于生成式模型的语音伪造方法溯源算法

该类算法认为特征表示是由统计分布生成的,在训练阶段依据训练数据的统计量优化分布参数,在评估时,通过定义决策规则,来推断最可能生成评估样本的分布。使用的经典生成式模型包括高斯混合模型(Gaussian mixture model,GMM)。Salvi等[45]使用线性频率倒谱系数(Linear frequency cepstral coefficients,LFCC)[46]特征描述声音信号的频谱包络的局部特性。而后对每类LFCC特征,单独训练一个GMM,以建模该类特征的统计分布。评估时,基于样本属于各个GMM的对数似然概率,输出识别结果。此外,作者提出了两种方法来处理开集条件下的语音伪造方法溯源。一种是统一置信度概

率比值阈值,即如果样本的最大类置信度概率与第二大概率之比大于2,则该样本属于置信度概率最大的类别。否则,则属于未知类。另一种是一类支持向量机(One class SVM, OC-SVM)[47]。具体地,作者在模型输出的类置信度概率空间中,为每个类别单独训练一个OC-SVM来识别该类别。测试时,如果样本不属于任何类,将其识别为未知类。

可以看出,该类方法大多依赖手工特征为输入,伪造指纹在手工特征中体现较为明显,可解释性较强。同时该类语音伪造方法溯源算法对算力、数据集规模要求不高,但其在处理高维非线性问题时表现不够理想,泛化性较差。随着人工智能技术的进步,深度学习算法展现出强大的高维非线性建模能力,基于深度学习的语音伪造方法溯源逐渐成为主流。

2.2.2 基于深度学习模型的语音伪造方法溯源

近年来,随着深度学习的快速发展,算力、数据集规模的不断增长,深度神经网络在语音欺骗检测[48]、语音增强[49]、说话人验证系统攻击[50]和语音隐写[51]等领域得到了广泛的运用。同时,能够区分复杂非线性特征的深度神经网络层出不穷,基于深度学习的语音伪造方法溯源算法也日趋成熟。该类算法主要以神经网络为基本框架,常用的神经网络框架包括卷积神经网络(Convolutional neural network,CNN)、残差网络(Residual network,ResNet)[52]和转换器(Transformer)[53]等。

(1) 基于CNN的语音伪造方法溯源算法

CNN通常由一系列的卷积层和池化层构成。卷积层中,不同的特征区域共享卷积核参数,有利于控制模型规模。池化层中,通过下采样操作,减小特征图尺寸,有利于进一步减小模型大小。这些特性使得CNN训练不容易过拟合,网络结构可以设计得更深更宽,更有利于挖掘数据中的深层抽象信息。同时,CNN具有归纳偏置特性,如平移不变性、尺度不变性和定位性[54],有利于学习到高质量的特征表示。

由于卷积结构可以保留特征图的结构信息,在语音伪造方法溯源领域,包含较多结构信息的时频特征谱图在基于CNN的语音伪造方法溯源算法中被广泛使用。比如,Choi等[55]以语音样本的梅尔谱图(Mel-spectrogram,Mel-Spec)特征作为输入,训练CNN模型。评估时,模型输出Mel-Spec特征属于各个类的概率,基于概率推断样本的伪造方法类别。Neri等[56]提出了一种名为ParalMGC的方法,其示意图如图4所示。该方法包含1个双分支结构特征提取器,以及1个后端分类器。每个分支结构均为1个CNN模型,分别以梅尔倒谱系数(Mel frequency cepstrum coefficient, MFCC)特征和伽马音调倒频谱系数(Gamma tone cepstral coefficients, GTCC)特征作为输入,提取样本的深度特征表示。基于CNN的后端分类器以提取的深度特征表示为输入,输出样本属于各个类别的概率,推断其类别。Rahman等[57]提出了一种基于多CNN集成的语音伪造方法溯源算法。作者以Mel-Spec为输入,将不同CNN输出的类置信度概率进行平均,得到最终的类置信度概率,以推断伪造方法类别。此外,作者利用额外的无标签数据作为未知类数据的替代品,在训练时,使用集成模型输出的预测类别作为无标签数据的伪标签,引导网络训练,增强对未知类的泛化性。测试时,作者将样本判断为置信度概率最大的类别。

(2) 基于Transformer的语音伪造方法溯源算法

CNN中的卷积操作,有利于挖掘语音时频特征谱图中的局部结构信息,但它并不擅长捕捉全局信息。语音作为时序信号,不同时间帧之间的关联性,尤其是远距离帧之间的关联性等全局信息,也是识别伪造方法的

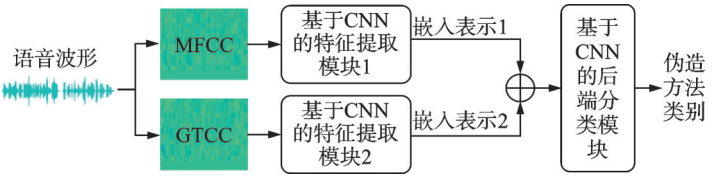

图4 基于CNN的ParalMGC语音伪造方法溯源算法

Fig.4 CNN-based ParalMGC speech forgery method attribution algorithm

关键信息。因此,不少研究者开始利用善于挖掘全局信息的Transformer进行语音伪造方法识别。

Bhagtani等[58]以Mel-Spec为输入,将在Audio Set数据集[59]上预训练的块掩蔽音频Transformer[60]适配到语音伪造方法溯源任务中进行微调,以提取样本的深度特征表示,并训练一个多层感知机,来预测伪造方法类别。类似地,Yadav等[61]以Mel-Spec为输入,将在Audio Set数据集[59]和Librispeech数据集[62]上预训练的自监督音频谱图变换器(Self-supervised audio spectrogram transformer,SSAST)[63]适配到语音伪造方法溯源任务中,以提取样本的特征表示,并训练一个全连接层,来预测伪造方法类别。Zhang等[64]提出了一种基于重叠子频对数梅尔声谱图块Transformer的溯源架构。该方法充分利用自注意力机制挖掘语音信号中的全局上下文信息,配合重叠子带切片技术,能够更敏锐地捕捉隐藏在局部频段与全局时序中的细微伪造痕迹,为基于全局特征的伪造方法溯源提供了高效的解决方案。此外,该作者团队还探究了基于联合数据集预训练Transformer在跨数据集上的泛化效果[65]。实验表明,基于聚类簇引导下的标签优化方案,有利于溯源模型学习到判别性更强的特征表示,提升溯源性能。上述方法使用的预训练Transformer模型参数量大,并且使用的输入需匹配预训练Transformer对输入的要求,限制了语音伪造方法溯源性能的进一步提高。考虑到Transformer缺乏CNN的归纳偏置特性,从头开始训练Transformer来学习高质量的深度特征表示,需要极大的数据集。而语音伪造方法溯源领域缺乏类似Audio Set[59]这样的大规模数据集。因此,有研究者探索将CNN和Transformer相结合,充分发挥二者的优势,直接依托伪造语音数据集,从头开始训练语音伪造方法溯源模型。比如,Li等[33]采用了结合CNN局部建模能力与Transformer全局建模能力的MFA-Conformer架构作为基线系统。实验表明,Conformer结构能够有效捕捉不同语音转换与合成系统在时频域留下的特定指纹,能够实现高精度的伪造方法分类。Bartusiak等[66]首先利用卷积模块从归一化谱图中提取深度特征图,以学习归纳偏置特性。而后使用紧凑Transformer[67]从深度特征图中提取样本的嵌入表示向量,并通过后接Softmax函数的全连接层输出样本属于各个类别的概率,来预测伪造方法类别。与Borrelli等[12]类似,Bhagtani等[58]也探究了使用已知类分布建模未知类分布的方法在开集识别中的性能。此外,Yadav等[61]还测试了在原始语音上训练的SSAST模型在经高级音频编码(Advanced audio coding,AAC)[68]处理的语音上的语音伪造方法溯源性能。实验结果表明,编码比特率越低,识别性能越差,当比特率为16 kb/s时,准确率下降到21.2%。

(3) 基于ResNet的语音伪造方法溯源算法

除了Transformer模型之外,ResNet模型通过引入残差连接,很好地克服了神经网络训练中的梯度消失或爆炸问题,使得不同深度的网络层之间能够高效传递梯度信息,也有利于模型捕捉到更大的感受野区域的信息,甚至全局信息。此外,ResNet模型继承了CNN的卷积结构,具有归纳偏置特性。因此,ResNet模型不仅擅长捕捉局部信息,也有利于建模语音信号远距离帧之间的关联性等全局信息,在语音伪造方法溯源领域受到不少研究者的青睐。

Yan等[69]探究了ResNet模型[52]与不同特征相结合的性能表现。采用的特征包括LFCC、MFCC和常数Q倒频谱系数(Constant Q cepstral coefficient, CQCC)。实验结果表明,基于ResNet模型的语音伪造方法溯源性能优于基于轻量级卷积神经网络(Lightweight CNN, LCNN)[70]等模型的语音伪造方法溯源性能表现。Deng等[37]设计了一种基于深度残差网络的溯源框架,专门处理64×64的对数梅尔时频图块。在模型优化方面,作者并未单纯依赖传统的分类损失,而是引入了有监督对比损失作为辅助约束,通过双重损失函数共同监督特征空间的学习。作者还测试了该算法在经AAC编解码转换的伪造语音上的伪造方法识别性能。实验结果表明,当压缩比特率高达64 kb/s时,识别准确率下降至80%左右。为进一步提高算法的溯源准确率,Zhang等[71]受深度学习训练末期"神经崩塌"现象的启发,提出了一种基于欧氏距离优化的Softmax损失函数。实验表明,该损失函数通过固定基于等角紧框架的分

类器权重并直接优化特征与类中心的欧氏距离，能够实现决策边界的自适应调整，使模型学习到判别性更强的嵌入表示，进一步提高了溯源准确率。为进一步挖掘无标签数据中的未知伪造算法信息，Zhang等[72]提出了软对比伪学习算法，该算法通过基于相似性的软过滤策略，筛选并匹配属于同一伪造类别的样本对，并利用基于相似度的自适应权重来增强对比学习的效果，从而显著提升特征的类内紧凑性。同时，针对未标记样本，结合类标签平滑技术和基于余弦相似度的置信度加权机制，避免传统硬伪标签可能引入的噪声误差，使模型在监督训练中更加稳健。针对开放世界场景中未见类别数量未知的情况，还提出了一种基于软对比伪学习的迭代估计算法，能够自动预测未见伪造类别的数量。此外，为防止攻击者通过对抗样本技术逃避溯源，张强等[73]以ResNet50模型为特征提取器，提出了基于鲁棒对抗防御边界的语音伪造方法溯源算法。如图5所示，该算法使用可微随机变换模块对输入进行多种组合随机变换，通过增加输入的随机性，来增加鲁棒性；同时，在计算损失时，除了基于交叉熵的Softmax分类损失之外，该算法通过约束模型学习到的嵌入表示与决策边界的距离，来增加攻击者实施成功攻击需要的扰动幅度，以进一步增强对抗鲁棒性。

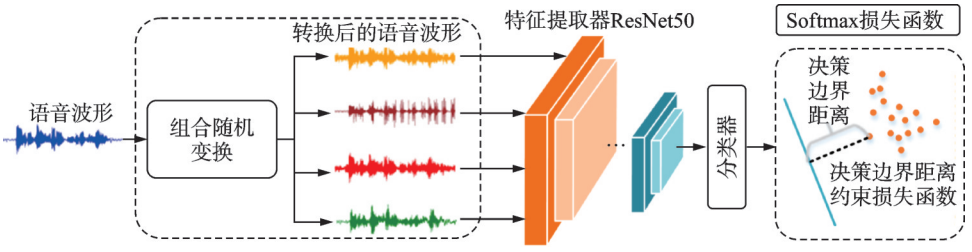

图5 基于鲁棒对抗防御边界的语音伪造方法溯源算法

Fig.5 Speech forgery method attribution based on robust adversarial defense boundaries

（4）基于其他模型的语音伪造方法溯源算法

除了上述主流架构，研究者还探索了多样化的建模视角。序列建模：Müller等[38]以语音波形为输入，使用基于长短时记忆网络层的循环神经网络（Recurrent neural network，RNN）直接从波形中提取768维的深度特征表示，捕捉时序上的微观伪造痕迹，而后通过多层感知机输出样本属于各伪造方法类别的概率。此外，Zhu等[74]使用RawNet2[75]提取语音信号的深度特征表示，并探索了TTS或VC方法的不同模块的可区分性。图神经网络：Shi等[76]对DAC[77]等主流神经编解码器进行了系统性评估，指出这些编解码器会在音频中引入独特的非线性压缩伪影。这些伪影不同于传统语音伪造痕迹，需要专门的建模分析。Xie等[78]将神经音频编解码方法视为一种新型语音伪造方法，并探索了图神经网络与Mel-Spec、深度特征表示相结合的语音伪造方法溯源性能。这里的深度特征表示是指直接使用预训练wav2vec2-XLS-R模型[79]提取的1 024维的嵌入表示向量。类似地，Klein等[80]将伪造算法的不同模块建模为不同属性，并探索了图神经网络识别不同属性的性能。具体地，作者使用预训练Whisper模型[81]提取768维的嵌入表示，并使用这些嵌入表示训练LCNN模型作为分类器，以识别不同的伪造方法属性。此外，作者还利用预训练Wav2vec 2.0模型[82]提取160维的嵌入表示，并使用这些嵌入表示来训练基于图神经网络的分类模型AASIST[83]，来识别不同的伪造方法属性。

2.2.3 *各类语音伪造方法溯源技术对比总结*

为了更直观地分析各类溯源技术的特点，表1对上述方法的优缺点进行了系统对比。首先，在传统机器学习方法中，判别式模型和生成式模型模型结构相对简单，计算开销小且具有较强的可解释性，但极其依赖手工特征的设计质量，难以有效捕捉现代神经声码器产生的高维非线性伪造痕迹，且抗信道

压缩能力较差。其次,基于CNN和ResNet的深度学习方法具有强大的局部特征提取能力和优良的归纳偏置特性,训练相对稳定,但在捕捉语音信号长距离全局依赖信息方面存在局限。最后,基于Transformer、序列建模、图神经网络等方法凭借自注意力机制在建模全局上下文方面表现优异,特征表征能力极强,但存在计算复杂度高、对大规模预训练数据依赖性强的问题。因此,在具体使用中应当充分考虑算力资源、数据规模和识别精度等实际情况,可以将卷积结构与注意力机制相结合,从而在保留局部细节的同时捕捉全局信息,提升溯源性能,但其本质仍主要依赖数据驱动的特征映射。因此,仍存在以下问题:(1)传统方法需要通过人工的方式设计并筛选特征,难以适应快速迭代且生成机理各异的新型伪造算法;(2)深度学习方法虽然在闭集测试下表现优越,但在面对未知伪造方法及音频编解码压缩时,模型的泛化能力和鲁棒性仍面临严峻挑战;(3)现有的溯源研究大多侧重于判别模型在特征层面的区分度,缺乏对生成模型固有指纹(如声码器相位伪影、扩散模型迭代噪声)的深度解耦与可解释性分析。

综上所述,语音伪造方法溯源技术正经历从“基于手工特征的浅层统计学习”向“基于数据驱动的深层表示学习”演变的趋势。尽管基于深度学习的方法在闭集测试中取得了显著成效,但在面对未知伪造方法、高信道压缩及跨数据集泛化等方面仍存在诸多挑战,需在未来的研究中进一步探索特征解耦、自监督学习及对抗鲁棒性增强等技术路径。

表1 语音伪造方法溯源技术对比

Table 1 Comparative analysis of speech forgery method attribution techniques

技术类别	细分方法	方法原理	优点	缺点/局限性
传统机器学习	基于判别式模型	寻找特征空间的最优分类超平面	模型简单,可解释性强,适合小样本	高维非线性拟合能力弱,抗编解码鲁棒性差
	基于生成式模型	建模特征的统计分布,基于最大似然判决	适合建模特征分布,开集识别逻辑清晰	依赖手工特征质量,难以捕捉深层抽象特征
深度学习	基于CNN	卷积与池化,提取局部时频结构特征	具备平移不变性,局部特征提取能力强,训练稳定	难以捕捉长距离全局依赖信息
	基于Transformer	自注意力机制,建模全局上下文依赖	擅长捕捉全局信息,特征表征能力强	缺乏归纳偏置,需大量数据或预训练,计算开销大
	基于ResNet	残差连接,允许训练极深网络	解决梯度消失,兼顾深层语义与浅层细节	对低比特率压缩的鲁棒性仍有待提升
	其他模型(RNN/GNN等)	时序记忆或图结构关系建模	适合处理变长序列或复杂的属性关联	训练复杂度高,属性识别未必能直接对应单一伪造方法

3 源说话人溯源

当确定伪造类型属于VC时,激活源说话人溯源任务。该任务负责从转换后语音中,挖掘转换过程残留的韵律、音色等源说话人身份线索,重建与源说话人身份相匹配的声纹特征,甚至还原源语音。VC在保留语音内容信息的同时,通过改变源说话人的音色来模仿指定目标说话人的声音。源说话人的声学特性在VC过程中发生了很大的变化,从转换后的语音中再次识别源说话人具有挑战性。但由于VC解耦的不完美以及技术实现中的缺陷,源说话人的某些信息仍然会留在转换语音中,源说话人溯源任务仍具有可行性。下文首先介绍源说话人溯源的身份残留机理,而后介绍源说话人溯源的主要算法。

3.1 语音转换中的源说话人身份残留机理

源说话人溯源之所以可行,其根本原因在于当前VC技术在“解耦-重构”的全流程中,不可避免地存在源身份信息的泄露与特征残留。尽管理想的VC旨在实现语言内容与说话人身份的完全解耦,但在实际操作中,源说话人的生物学特征(声纹)与行为学特征(韵律、习惯)往往通过“非完全解耦的隐变量”和“为了保持语义自然度而保留的韵律结构”渗透至转换语音中。基于物理声学特性与深度生成模型的运作机制,可将源说话人溯源的身份残留机理归纳为以下5个关键维度。

3.1.1 内容-音色解耦的不完全性

语音转换的核心假设是“内容”与“音色”可以被独立提取。然而,在物理声学层面,这两者是高度耦合的,导致算法层面的“硬解耦”难以彻底实现,这是源说话人溯源可行的根本原因。主要体现在如下两个方面:一是生理结构耦合。音色主要由声道的物理结构(如声道长度、口腔几何形状)决定,但这些物理属性同时也塑造了共振峰的精细结构,直接决定了每一个音素(内容单位)的声学表现。二是纠缠的特征表示。当VC模型尝试提取内容特征时,无法完全剥离这些与内容紧密纠缠的微观声学属性[84]。例如,不同人发出同一个元音“啊”,其共振峰的中心频率、带宽和能量分布是由其特定声道决定的。模型在提取“这是‘啊’音”这一抽象内容时,难免会携带“这是由源说话人特定声道发出的”残留信息,导致源身份信息潜伏在内容编码中被传递到解码端。

3.1.2 频谱细节与物理特征残留

现有的内容编码器和声码器在处理信号时,往往会保留源语音的微观频谱细节与物理特征,这是源说话人溯源的声学线索。主要体现在如下两个方面:一是高频细节泄露。现有的内容编码器往往是一个有损压缩过程,为了保证重建语音的清晰度,编码器倾向于保留源语音的高频频谱包络。这些高频分量往往包含个体独特的音质色彩。二是非语言残差信息保留。声门闭合模式、呼吸声等非语言残差信息往往被视为“风格”的一部分被保留[85]。例如,若源说话人具有显著的“气声”或“沙哑”特质,这种物理发声状态很难被完全抹除,即使转换了基频和共振峰,生成的语音仍会带有源说话人的“声质感”。

3.1.3 韵律模式的显式保留

为了保证转换语音的自然度、情感表达及语义连贯性,大多数VC模型(尤其是One-shot VC和基于帧的VC)采用了“韵律模式的显式保留”策略,这是源说话人溯源的时序线索。主要体现在如下两个方面:一是时长与停顿复制。模型通常直接复制源语音的音素时长和停顿模式。这意味着转换语音严格遵循了源说话人的语速变化和断句逻辑。二是基频动态轮廓保留。虽然VC会调整基频的均值以匹配目标说话人的音高范围(如男转女),但源语音的基频动态变化模式(即语调起伏、微颤抖动)往往被保留[86]。这种动态轮廓是说话人情感表达和语调习惯的直接载体,具有极强的身份指纹属性。

3.1.4 行为习惯与个人风格残留

除了上述物理声学特征线索,源说话人后天养成的行为习惯也是溯源的重要依据,这类特征通常被称为“行为生物特征”,即源说话人溯源的生物特征线索。主要体现在如下两个方面:一是独特的说话节奏。源说话人习惯性的重音位置、特定词汇的拖长处理,以及独有的语调节奏,构成了个体独特的“行为指纹”[87]。二是口音与发音方式。尽管目标是改变音色,但源说话人的方言口音、咬字力度以及特定的发音缺陷(如齿音过重)往往会穿透转换模型,残留在生成的语音中。这些行为层面的特征极难通过简单的声学变换被完全覆盖。

3.1.5 目标映射的统计平滑效应

这是生成模型自身机制缺陷导致的源说话人溯源线索。VC模型在推断目标说话人音色时,通常依赖于目标说话人的统计平均表征,如说话人嵌入向量。这种统计平均表征会在如下两个方面为源说

话人溯源提供线索:一是平均化导致的细节缺失。目标嵌入向量捕捉的是该说话人最典型、最常见的声学特征(均值)。然而,这个平均表示无法涵盖目标说话人在所有音素、所有情感状态下的极端细节。二是音色回退。当目标嵌入向量无法覆盖源语音中的某些罕见发音、复杂语境或极端情感时,解码器因为缺乏目标域的对应先验,往往会发生“泛化失败”,进而“回退”并重构出源语音的声学纹理。这种现象导致转换语音中出现“音色混合”,直接暴露了源说话人的身份底色[33]。

3.2 源说话人溯源的主要算法

源说话人溯源旨在从经过VC的语音中,提取并确认原始说话人的身份信息。定义X_{vc}为转换后的伪造语音,S_{source}为源说话人身份标签。源说话人溯源的一般流程为通过构建溯源模型F_{source},逆向推导语音生成过程中的身份映射关系,从而输出预测的源说话人身份$\hat{y}$,甚至还原源语音。

$$\hat{y}=F_{\mathrm{trace}}(b(X_{\mathrm{vc}})) \tag{5}$$

式中$b(\cdot)$代表特征提取操作。根据溯源机理和实现路径的不同,现有的源说话人溯源算法主要分为基于语音还原的溯源算法和基于识别的溯源算法两大类。各类方法的技术分支如图6所示。

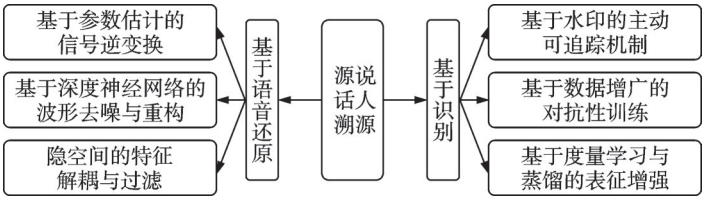

图6 源说话人溯源技术分类框架

Fig.6 Taxonomy of source speaker tracing techniques

3.2.1 基于语音还原的源说话人溯源

此类算法的核心思想是“逆变换”,即假设伪造过程是对源语音施加了某种变换函数F,溯源的目标则是寻找逆映射F^{-1}或消除变换带来的干扰,从而恢复出源说话人的波形或声学特征,最后利用现有的自动说话人验证(Automatic speaker verification, ASV)进行身份判定。根据还原机制与操作域的不同,现有方法可分为以下3类:

(1) 基于参数估计的信号逆变换

早期的语音变换常采用变音调、变速度或声道长度归一化(Vocal tract length normalization, VTLN)等基于数字信号处理的手段。此类变换通常具有明确的数学形式,因此溯源的关键在于估计变换参数并执行逆操作。Wang等[88]将基频线性变声的还原问题建模为伪装因子估计问题,利用基频比反推伪装参数,并设计了改进的MFCC提取算法,成功从变声语音中恢复了原始倒谱特征。为了应对更复杂的非线性参数变换,Zheng等[89]引入了“分析-合成”的优化思路,提出了一种基于最小化说话人验证分数的反函数方法。该方法将参数估计转化为优化问题,利用X-vector的噪声鲁棒性来指导参数搜索,在处理基频缩放和VTLN时取得了较好的还原效果。Li等[90]进一步将此思路推广至盲还原场景,实现了在无嫌疑人参考语音情况下的参数反演。然而,此类方法主要针对白盒或灰盒条件下具有明确物理定义的参数化变换,并假设变换模型已知,难以应对基于黑盒神经网络的非线性生成式伪造。

(2) 基于深度神经网络的波形去噪与重构

随着深度学习的发展,针对参数未知或难以用简单公式描述的复杂伪装,研究者开始利用神经网络强大的非线性拟合能力来端到端地学习逆变换映射。王永全等[91]提出了一种基于扩大因果卷积神经网络(Dilated casual CNN, DC-CNN)的还原模型。该模型将变音调、变节拍等伪装语音的历史采样点信息映射回原始状态,通过跳跃连接优化深层传递并结合压扩转换,实现了对伪装语音的高质量波形重构。此外,针对对抗性伪装场景,Guo等[92]近期探讨了对抗扰动的可逆性,提出了用于语音隐私保护的说话人对抗扰动添加及去除算法,其示意图如图7所示。作者发现,加了防御性扰动的源说话人语音可以欺骗ASV系统,同时通过训练专门的“扰动去除网络”,可以剥离用于掩盖身份的对抗噪声,从而

“净化”被保护语音并恢复源声纹信息。此类方法本质上将还原视为一个“去噪”或“超分”过程，适用于处理非结构化的噪声干扰或复杂的信号伪装。

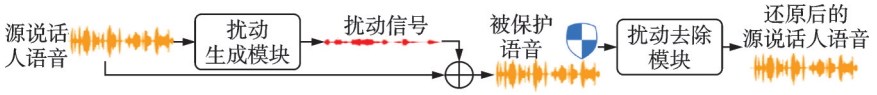

图7 用于语音隐私保护的说话人对抗扰动添加及去除算法

Fig.7 Adversarial perturbation injection and removal algorithms for speech privacy protection

(3) 隐空间的特征解耦与过滤

对于基于GAN或Diffusion的现代VC系统，其映射函数高度非线性且通常不可逆，直接在波形层面还原源语音极具挑战。因此，前沿研究转向在特征空间进行操作，旨在解耦并滤除目标说话人的特征，保留源说话人的残留信息。Deng等[93]开发了配备差分校正算法的表示学习模型REVELIO。在目标说话人参考语音的引导下，该模型试图在特征空间中减去与目标声纹平行的分量，从而消除VC引入的目标身份影响。然而，简单的减法操作容易导致源特征的过度损失。针对这一局限，Zhang等[94]提出了一种基于目标说话人特征过滤的创新框架，其示意图如图8所示。该方法不再寻求对波形的直接还原，而是利用目标参考语音构建“过滤器”，通过正交分解显式建模目标特征与源特征的差异，并预测软掩码以精准滤除转换语音中属于目标说话人的分量。实验表明，该方法在白盒和黑盒条件下均能有效提取残留的源说话人嵌入，代表了从“粗糙还原”向“精细化特征解耦”演进的重要方向。

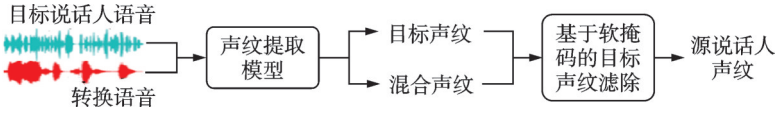

图8 基于目标说话人特征过滤的源说话人溯源

Fig.8 Source speaker tracing based on target speaker feature filtering

综上所述，基于还原的溯源方法经历了如下的技术范式演变：针对参数化线性/非线性变换后的伪造语音，基于物理模型的参数估计方法已较为成熟[88-90]；针对电子伪装和对抗扰动伪装后的语音，基于神经网络的波形重构提供了有效的去噪手段[91-92]；然而，面对黑盒条件下的深度生成式VC攻击，现有的波形还原方法显得力不从心。以REVELIO[93]和特征过滤[94]为代表的隐空间特征解耦技术打破了必须重构波形的限制，成为解决复杂非线性伪造溯源问题的主流趋势。

3.2.2 基于识别的源说话人溯源

此类算法不追求显式地还原语音波形，而是侧重于特征空间的鲁棒性构建。其核心思想是将源说话人溯源视为一个跨域或对抗环境下的声纹识别问题，致力于训练能够“穿透”VC伪装层、直接提取源说话人身份不变特征的深度模型。根据特征学习范式的不同，主要分为以下3类：

(1) 基于数据增广的对抗性训练

针对VC系统对声纹特征的非线性破坏，最直观的防御策略是将攻击样本纳入训练过程，通过强制模型学习伪造数据的分布来重划分类边界。Mohan等[95]系统评估了基于VC的语音匿名化系统的安全性，指出若攻击者掌握VC系统的先验知识(如训练数据或模型结构)，可显著提升溯源成功率。基于此发现，Cai等[84]提出了一种基于对抗性数据增广的溯源策略，其示意图如图9所示。该算法将带有源说话人标签的转换语音加入说话人验证模型的训练集，通过相似性损失函数鼓励模型将源自同一个源说话人的转换语音判为来自同一个源说话人，将来自不同源说话人的转换语音判断为来自不同的源说话

人,达到身份溯源目标。实验证明,在近似白盒条件(训练与测试使用同一VC模型)下,该方法能有效提升模型对特定VC模式的穿透能力。尽管数据增广在白盒或灰盒场景下表现尚可,但在黑盒场景下仍面临严峻挑战。Liu等[96]在评估语音隐私保护有效性时揭示了其中的理论瓶颈:数据增广本质上是试图通过扩充样本覆盖伪造空间,但在黑盒条件下,复杂的非线性变换可能导致不同说话人的特征映射到同一区域,出现“特征挤压”现象。因此,单纯依赖数据覆盖难以有效应对未知且高度非线性的复杂变换。

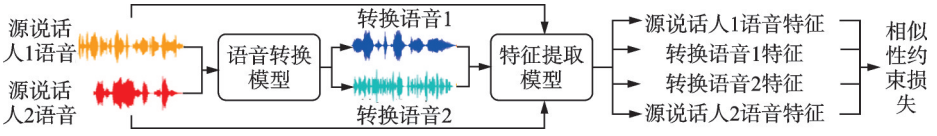

图9 基于对抗性数据增广的源说话人溯源策略

Fig.9 Source speaker tracing strategy based on adversarial data augmentation

(2) 基于度量学习与蒸馏的表征增强

为了解决黑盒场景下源特征提取难、特征空间混淆严重的问题,最新的研究转向探索更精细的损失函数设计与蒸馏学习范式,旨在挖掘深层身份表征。针对转换语音中源身份信息微弱的问题,Wang等[97]提出了一种基于说话人对比学习的溯源框架。该方法在嵌入提取器中引入对比损失机制,显式地拉近同一源说话人(无论是否经过VC转换)的嵌入距离,同时推远干扰说话人的嵌入。这种强监督的度量学习策略迫使模型挖掘那些不易被VC算法抹去的深层身份不变特征,在SSTC测试集上实现了最低的等错误率。考虑到现有的ASV模型在真实语音上已具备强大的特征提取能力,Ma等[98]提出利用知识蒸馏进行特征增强。该方法将预训练的鲁棒ASV模型作为“教师网络”,指导针对源说话人验证的“学生网络”学习。通过让学生网络模仿教师网络对真实语音的响应,模型学会了如何从含噪或转换语音中恢复出稳定的身份表征,为利用大规模预训练模型辅助溯源任务提供了新思路。Li等[33]在构建SSTC基准时指出,由于伪造过程中源信息丢失严重,单纯的源说话人验证往往陷入局部最优。作者提出引入转换方法识别作为辅助任务。通过多任务学习框架,模型在识别伪造手段的同时,能够辅助定位该手段可能残留的特定指纹,从而反向提升源说话人溯源的准确性。

(3) 基于水印的主动可追踪机制

与上述“被动分析”不同,另一类研究主张通过“主动防御”来保证可追踪性。Ren等[99]设计了一种可追踪的VC框架,将包含源说话人身份信息的不可感知水印集成到VC生成过程中。在溯源阶段,通过特定的解码器即可从转换语音中准确提取并验证源说话人身份。这种方法虽然规避了特征提取的难题,实现了高精度的身份认证,但其前提是需要拥有对VC生成过程的控制权(即要求服务提供商配合植入水印),因此更适用于合规监管下的溯源场景,而非针对恶意攻击的盲取证。

3.2.3 各类源说话人溯源技术对比总结

为了深入剖析现有源说话人溯源技术的演进脉络与应用边界,表2对上述主流方法的原理、优势及局限进行了系统性对比。首先,基于语音还原的方法遵循“逆变换”的物理直觉,具有较强的可解释性。其中,传统的信号参数估计(如基频修复[88])计算高效,但受限于线性假设,无法应对深度生成模型的非线性映射;基于神经网络的波形重构(如DC-CNN[91])虽然提升了对非结构化伪装(如电子变声)的还原能力,但在面对复杂的VC伪影时往往难以完美重构波形。值得注意的是,最新的隐空间特征解耦策略(如REVELIO[93]、掩码过滤[94])打破了“必须还原波形”的限制,通过在特征域滤除目标说话人干扰,显著提升了对非线性VC攻击的溯源精度,代表了该类别的高阶演进方向。其次,基于识别的方法侧重于

构建端到端的鲁棒判别空间。对抗性数据增广策略简单直接,能有效降低白盒/灰盒条件下的等错误率,但面临黑盒场景下的“特征挤压”问题,泛化瓶颈明显;表征增强学习(如对比学习[97]、知识蒸馏[98])通过挖掘深层身份不变特征,有效缓解了未知攻击下的性能衰减,是当前SSTC挑战赛的主流趋势;而主动水印虽然实现了近乎完美的溯源准确率,但依赖于生成源头的介入与控制,不适用于针对恶意第三方攻击的盲取证场景。

综上所述,源说话人溯源研究正经历从“针对简单参数变换的信号反演”向“对抗复杂深度生成的特征解耦与鲁棒表征”的范式转移。尽管现有方法在白盒条件下已取得显著进展,但在黑盒跨模型场景下的鲁棒性仍面临挑战。未来的研究需进一步探索如何结合声码器残留指纹、韵律行为特征等深层线索,突破未知VC系统下的溯源瓶颈。

表2 源说话人溯源技术对比

Table 2 Comparative analysis of source speaker tracing techniques

技术类别	细分方法	方法原理	优点	缺点/局限性
基于语音还原	基于参数估计的信号逆变换	假设伪装是线性或参数化的,通过数学推导估计变换参数并执行逆操作	(1) 计算复杂度低 (2) 物理可解释性强 (3) 少量样本即可实现	(1) 仅适用于线性/简单非线性变换 (2) 无法应对黑盒神经网络生成的复杂伪造
	基于深度神经网络的波形去噪与重构	利用CNN/U-Net等结构学习“伪造→原始”的端到端非线性映射,直接重构波形	(1) 对电子伪装、噪声掩盖等非结构化干扰效果好 (2) 不需要显式参数假设	(1) 难以完美还原VC丢失的高频细节 (2) 模型训练依赖成对数据,跨数据集泛化差
	隐空间的特征解耦与过滤	不还原波形,而是在特征域利用正交分解或减法,滤除目标说话人特征,保留源身份残留	(1) 能够应对高度非线性的VC攻击 (2) 在跨模型测试中鲁棒性较强	(1) 依赖目标说话人的参考语音作为引导 (2) 极端转换下源特征残留过少时失效
基于识别	基于数据增广的对抗性训练	将多种转换/伪造语音加入训练集,强迫模型学习覆盖伪造空间的分类边界	(1) 实现简单,易于集成现有ASV系统 (2) 白盒/灰盒条件下性能提升显著	(1) 黑盒泛化差,难以覆盖未知的生成算法 (2) 易出现“特征挤压”,导致不同说话人混淆
	基于度量学习与蒸馏的表征增强	利用对比学习、知识蒸馏或多任务学习,挖掘不易被VC抹去的深层身份不变特征	(1) 黑盒鲁棒性强:不依赖特定攻击类型的先验 (2) 能有效利用预训练大模型的先验知识	(1) 训练策略复杂,难收敛 (2) 对计算资源要求较高
	基于水印的主动可追踪机制	在VC生成过程中嵌入不可感知的源身份水印,溯源时通过解码器提取	(1) 溯源准确率极高 (2) 即使经过有损传输也能保持可追踪性	(1) 应用受限,需控制生成端,属于“主动防御” (2) 无法用于检测第三方恶意攻击

4 挑战与未来研究方向

尽管语音伪造溯源技术在近年来取得了显著进展,但随着生成式人工智能的迭代速度呈指数级增长,溯源研究仍面临着从理想化实验室环境走向复杂现实场景的严峻挑战。

4.1 开放世界与持续性溯源

现有的溯源模型大多基于“闭集假设”训练,难以适应动态变化的现实环境。现实世界中,新型算法层出不穷,从GAN到Diffusion,再到最新的Flow Matching,伪造机理的突变导致数据分布发生剧烈漂移。虽然近期研究者提出的软对比伪学习[72]等方法在一定程度上提升了模型对未知伪造数据的适应性,初步缓解了闭集模型的“过度自信”问题。但在面对跨度极大的异构伪造攻击时,现有方法仍难以实现真正意义上的广义零样本识别。未来的溯源系统应具备类似生物免疫系统的能力,即持续感知与终身学习。研究重点应转向元学习与提示学习,使模型能够利用极少量样本快速适应从未见过的伪造架构,实现对未知威胁的敏捷响应。

4.2 复杂信道与神经编解码下的鲁棒特征表示

在司法取证与社交媒体监控等实际应用中,截获的伪造语音往往经过了复杂的信道传输与后处理。一方面,有损压缩(如MP3, AAC, Opus)、网络丢包、环境噪声叠加以及恶意对抗后处理(如滤波、重采样)会严重破坏语音中的微观伪造指纹(如高频伪影、相位不连续性),导致溯源性能急剧下降。另一方面,EnCodec、SoundStream等新兴的神经音频编解码器会引入非线性量化残差与伪影,这不仅破坏了原始伪造指纹,更可能引入干扰性的“编解码指纹”,导致溯源模型误判。因此,发展复杂信道条件下鲁棒的溯源技术是当务之急。一方面,可以通过数据增强模拟各类信道失真及神经音频编解码处理,提高模型的泛化能力;另一方面,应利用自监督学习在大规模含噪数据上预训练溯源编码器,探索提取具有信道不变性的深层声学特征,使其能够从严重失真的信号中提取稳定的身份与方法痕迹。

4.3 深度伪造模型逆向工程与实例溯源

语音伪造溯源领域,目前的研究多止步于“方法分类”或“源说话人识别”,对于更深层次的“模型实例溯源”,还没有公开的学术成果出现。但在计算机视觉与通用机器学习安全领域,针对黑盒模型的逆向与窃取技术,已经展开了初步的探索。Shi等[100]提出了基于探索性查询的模型窃取攻击,证明了在黑盒条件下通过训练替代模型来复制目标分类器功能的可能性。更为深入的是,Asnani等[20]在图像生成领域首创了“模型解析”任务,提出了一种从生成的伪造图像中直接反向推断生成模型超参数(如网络架构深度、损失函数类型)的逆向工程框架,证明了利用指纹估计网络从输出中恢复模型微观配置的可行性。然而,针对语音伪造模型(如TTS和VC)的同类研究目前仍是一片空白。语音信号的高维时变特性与复杂的声码器相位重构机制,使得从生成的语音样本逆向推导具体的模型参数或超参数配置成为一个极具挑战的病态问题。因此,模型逆向将是高阶溯源的重要方向。未来的研究可借鉴图像领域的成功经验,探索利用生成模型的微观参数差异(如不同随机种子初始化的同架构模型、不同训练步数产生的细微指纹差异),实现对具体模型实例的唯一性归因。这将填补从“类溯源”到“例溯源”的缺口,有助于在版权保护与网络犯罪调查中,精准锁定具体的侵权模型或作案工具。

4.4 源说话人溯源的非线性反演与攻防博弈

源说话人溯源本质上是在与不断进化的VC解耦技术进行博弈。源说话人溯源旨在从伪造语音中恢复攻击者的真实身份,但在深度学习时代,这一任务面临着理论与实践的双重壁垒。在理论层面,面临着深度神经网络的非线性与不可逆性难题:早期的语音变换多基于音高缩放等线性操作,其逆变换相对容易。然而,现代VC系统依赖深层神经网络,其包含大量ReLU等非线性激活函数与降采样操作,导致信息在正向传播中发生有损压缩。从数学角度看,深度神经网络的逆变换是一个典型的病态不适定问题。如何构建有效的反变换近似解,在缺少单侧抑制信息的情况下重构输入特征,是溯源算法扩展至深层架构的前提。在实践层面,面临着极致解耦与特征残留的“矛盾之争”:当前的VC研究正致力于通过更精细的瓶颈层设计、纯净文本监督以及对抗性训练,来彻底“清洗”源说话人的音色残留,

以提升转换的逼真度与解耦能力。这使得溯源任务变得愈发困难,攻击方试图抹去指纹,防御方则试图挖掘残留。未来的研究重心需从简单的特征减法转向行为生物特征挖掘。即便VC模型极力去除音色等源信息,源说话人的发声习惯、韵律节奏、呼吸模式等“行为特征”往往难以被完全重构。未来的溯源算法应聚焦于挖掘这些在高维空间中具有持久性的深层身份痕迹,实现在多重转换或有损压缩干扰等黑盒条件下,依然保持对源身份特征空间的可区分性。

4.5 主动防御与被动取证的协同共治

单一的技术手段无法应对复杂的安全威胁,构建分层防御体系势在必行。虽然数字水印技术等主动防御技术能为合规的商业大模型提供确权手段,但对于蓄意犯罪的不法分子,他们往往会使用开源模型并剥离水印模块,或训练私有的无水印生成器来规避追踪。在这种非合作场景下,主动防御将完全失效。此时,“事后诸葛亮”式的溯源,作为被动防御手段,将是反制不法分子的有效手段。因此,未来的防御体系不应是单一维度的,而应构建“主动防御确权,被动取证兜底”的协同机制。对于合规应用,强制推行鲁棒水印与区块链存证,降低监管成本。对于恶意攻击,将被动取证技术作为不可替代的“最后一道防线”。未来的被动取证研究需要更加关注对微弱痕迹的盲检测能力,确保在攻击者刻意抹除水印或进行反取证操作时,依然能够从声学信号的底层物理特性中锁定伪造来源。

4.6 伦理、法律与隐私考量

技术的落地必须伴随着法律与伦理的约束。随着语音溯源技术的深入应用,技术之外的伦理与法律边界问题日益凸显,这直接关系到技术落地的可行性。首先,说话人溯源技术的落地需要平衡隐私和安全。源说话人溯源技术虽然有助于打击电信诈骗,但也可能被滥用于去匿名化攻击,侵犯合法用户的隐私(如匿名举报者或配音演员的身份保护)。因此,如何在可溯源性与隐私保护之间寻找平衡,例如设计仅授权机构可见的“可控溯源”机制,是未来系统设计的核心伦理准则。其次,取证的法律效力对溯源技术提出了更高的要求。在司法实践中,溯源结果的可解释性至关重要。基于深度学习的“黑盒”判决在法庭上缺乏解释力,难以作为直接定罪证据。未来的研究必须强化溯源证据的可解释性,即不仅要给出“是谁是什么伪造的”结论,还要提供可视化的声学证据(如频谱篡改区域热力图、相位异常点),以满足司法取证对证据链完整性和客观性的严苛要求。

5 结束语

随着生成式人工智能技术日益精进,社会已步入了一个真假难辨的“后真相”时代。传统的“真/假”二元检测虽然筑起了防御的第一道防线,但在面对复杂的司法取证、责任认定及版权保护需求时,仅回答“是否伪造”已显不足,回答“由谁伪造(方法溯源)”、“伪造了谁(源说话人溯源)”,以及“用什么伪造(模型逆向)”成为构建数字语音安全纵深防御体系的关键。

本文立足于取证与防御的现实需求,系统构建了涵盖语音伪造方法溯源、源说话人溯源以及模型逆向工程的层级化溯源技术体系,并梳理了该领域的演进脉络:在语音伪造方法溯源方面,研究经历了从“声学特征+统计模型”向“深度表征学习”的跨越。当前的算法不仅利用CNN、ResNet和Transformer/Conformer等混合架构精准捕捉伪造算法在时频域引入的细微指纹,更开始结合软对比学习与分布外检测技术,积极探索开放世界场景下的未知伪造方法识别与新类发现问题,试图解决从闭集分类走向开放识别的泛化难题。在源说话人溯源方面,技术路线正经历深刻的范式转移。早期的研究主要针对线性变声或参数化伪装,采用信号参数估计与反演的思路;而面对现代深度生成模型带来的非线性映射与信息丢失,前沿研究已转向基于神经网络的波形重构与隐空间特征解耦。通过引入对抗性数据增广、知识蒸馏及正交分解机制,现有方法正逐步突破黑盒条件下的“特征挤压”瓶颈,实现在复杂伪造攻击下的源身份鲁棒提取。

展望未来,构建可信数字语音生态需要技术与治理的协同共进。尽管数字水印等主动防御技术为合规监管下的溯源提供了高确定性方案,但被动取证技术作为对抗恶意第三方攻击的基石,其重要性愈发凸显。尽管面临开放世界泛化、信道鲁棒性及模型可解释性等严峻挑战,但随着深度学习理论的突破与跨学科研究的深入,构建一个可信、可溯源的数字语音生态环境将成为可能。未来的研究需要在提升算法泛化性、鲁棒性及可解释性的同时,深化主动与被动取证的融合,构建一个涵盖“事前预防、事中检测、事后溯源”的全流程闭环治理体系,为维护国家信息安全、打击网络犯罪与重塑社会数字互信提供坚实的技术支撑。

参考文献:

[1] ALMUTAIRI M. A comprehensive survey of audio forgery detection: Challenges and novel trends[J]. Journal of Electrical Systems and Information Technology, 2025, 12(1): 30.

[2] OORD A V D, DIELEMAN S, ZEN H, et al. WaveNet: A generative model for raw audio[C]//Proceedings of Speech Synthesis Workshop. [S.l.]: ISCA, 2016: 1-15.

[3] CHEN S, LIU S, ZHOU L, et al. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers[J]. arXiv preprint arXiv: 2406.05370, 2024.

[4] FRASER M. Deepfake statistical data (2023-2025)[EB/OL]. (2025-05-27). https://views4you.com/deepfake-database/.

[5] LENG C, HO-HIM C. Arup lost $25 million in Hong Kong deepfake video conference scam[EB/OL]. Financial Times, https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea.

[6] COLMAN B. $18.5 million lost in a crypto scam: AI voice fraud's growing threat to financial security[EB/OL]. Reality Defender, https://www.realitydefender.com/insights/millions-lost-in-hong-kong-crypto-scam.

[7] LI M, AHMADIADLI Y, ZHANG X P. A survey on speech deepfake detection[J]. ACM Computing Surveys, 2024, 57(7): 165.

[8] MA X, LIANG T, ZHANG S, et al. Improved lightCNN with attention modules for ASV spoofing detection[C]//Proceedings of IEEE International Conference on Multimedia and Expo. [S.l.]: IEEE Computer Society, 2021: 1-16.

[9] TAK H, PATINO J, TODISCO M, et al. End-to-end anti-spoofing with RawNet2[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 6369-6373.

[10] ZHANG B, CUI H, NGUYEN V, et al. Audio deepfake detection: What has been achieved and what lies ahead[J]. Sensors (Basel, Switzerland), 2025, 25(7): 1989.

[11] RABHI M, BAKIRAS S, DI PIETRO R. Audio-deepfake detection: Adversarial attacks and countermeasures[J]. Expert Systems with Applications, 2024, 250: 123941.

[12] BORRELLI C, BESTAGINI P, ANTONACCI F, et al. Synthetic speech detection through short-term and long-term prediction traces[J]. EURASIP Journal on Information Security, 2021, 2021(1): 1-14.

[13] LI X, CHEN P Y, WEI W. Where are we in audio deepfake detection? A systematic analysis over generative and detection models[J]. ACM Transactions on Internet Technology, 2025, 25(3): 20.

[14] SHEN J, PANG R, WEISS R J, et al. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2018: 4779-4783.

[15] YAMAMOTO R, SONG E, KIM J M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2020: 6199-6203.

[16] JEONG M, KIM H, CHEON S J, et al. Diff-TTS: A denoising diffusion model for text-to-speech[J]. arXiv preprint arXiv: 2104.01409, 2021.

[17] KIM J, KONG SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[C]//Proceedings of International Conference on Machine Learning. [S.l.]: ACM, 2021: 5530-5540.

[18] QIAN K, ZHANG Y, CHANG S, et al. AutoVC: Zero-shot voice style transfer with only autoencoder loss[C]//Proceedings of International Conference on Machine Learning. [S.l.]: ACM, 2019: 5210-5219.

[19] 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 IEEE Spoken Language Technology Workshop. [S. l.]: IEEE, 2019: 266-273.

[20] ASNANI V, YIN X, HASSNER T, et al. Reverse engineering of generative models: Inferring model hyperparameters from generated images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15477-15493.

[21] KONG J, KIM J, BAE J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. Advances in Neural Information Processing Systems, 2020, 33: 17022-17033.

[22] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust DNN embeddings for speaker recognition[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018: 5329-5333.

[23] WU Z, KINNUNEN T, EVANS N, et al. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge[C]//Proceedings of Annual Conference of the International Speech Communication Association. [S. l.]: [s.n.], 2015: 2037-2041.

[24] WANG X, YAMAGISHI J, TODISCO M, et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech[J]. Computer Speech and Language, 2020, 64: 101114.

[25] 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.

[26] WANG X, DELGADO H, TAK H, et al. ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale[C]//Proceedings of The Automatic Speaker Verification Spoofing Countermeasures Workshop. [S.l.]: [s.n], 2024: 1-8.

[27] FRANK J, SCHÖNHERR L. WaveFake: A data set to facilitate audio deepfake detection[C]//Proceedings of Conference on Neural Information Processing Systems Datasets and Benchmarks Track. [S.l.]: [s.n], 2021: 1-17.

[28] SALVI D, BORRELLI C, BESTAGINI P, et al. Synthetic speech attribution: Highlights from the IEEE signal processing cup 2022 student competition[J]. IEEE Signal Processing Magazine, 2023, 40(6): 92-98.

[29] MA H, MEMBER S, YI J, et al. CFAD: A Chinese dataset for fake audio detection[J]. Speech Communication, 2024, 164: 10312.

[30] SISMAN B, YAMAGISHI J, KING S, et al. An overview of voice conversion and its challenges: From statistical modeling to deep learning[C]//Proceedings of IEEE/ACM Transactions on Audio Speech and Language Processing. [S.l.]: IEEE, 2021, 29: 132-157.

[31] LORENZO-TRUEBA J, YAMAGISHI J, TODA T, et al. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods[C]//Proceedings of Speaker and Language Recognition Workshop. Les Sables: [s.n.], 2018: 195-202.

[32] ZHAO Y, HUANG W C, TIAN X, et al. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion[C]//Proceedings of Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge. [S.l.]: [s.n.], 2020: 80-98.

[33] LI Z, LIN Y, YAO T, et al. The database and benchmark for the source speaker tracing challenge 2024[C]//Proceedings of IEEE Spoken Language Technology Workshop. [S.l.]: IEEE, 2024: 1254-1261.

[34] DHAMIJA A R, MANU EL G, BOULT T E. Reducing network agnostophobia[C]//Proceedings of Advances in Neural Information Processing Systems. Montreal, Canada: [s.n.], 2018, 31: 1-12.

[35] GREENBERG C S, MASON L P, OMID S. Two decades of speaker recognition evaluation at the national institute of standards and technology[J]. Computer Speech & Language, 2020, 60: 101032.

[36] LASZKIEWICZ M, KOLOSSA D, FISCHER A. Single-model attribution for spoofed speech via vocoder fingerprints in an open-world setting[J]. arXiv preprint arXiv: 2306. 06210, 2024.

[37] DENG J, REN Y, ZHANG T, et al. VFD-Net: Vocoder fingerprints detection for fake audio[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul: IEEE, 2024: 12151-12155.

[38] MÜLLER N M, DIECKMANN F, WILLIAMS J. Attacker attribution of audio deepfakes[C]//Proceedings of Annual Conference of the International Speech Communication Association. Incheon: [s.n.], 2022: 2788-2792.

[39] KUMAR K, GESTIN L, COURVILLE A. MelGAN: Generative adversarial networks for conditional waveform synthesis[J]. Advances in Neural information Processing systems, 2019, 32: 1-12.

[40] ODENA A, DUMOULIN V, OLAH C. Deconvolution and checkerboard artifacts[J]. Distill, 2016, 1(10): 1-10.

[41] KONG Z, PING W, HUANG J, et al. DiffWave: A versatile diffusion model for audio synthesis[C]//Proceedings of

International Conference on Learning Representations. [S.l.]: [s.n.], 2021: 1-17.
[42] HUANG R, HUANG J, YANG D, et al. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models [C]//Proceedings of International Conference on Machine Learning. [S.l.]: [s.n.], 2023: 13916-13932.
[43] KANEKO T, KAMEOKA H, TANAKA K, et al. CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion[C]//Proceedings of Interspeech. [S.l.]: [s.n.], 2020: 2017-2021.
[44] ABDULHAMIED R M, NAIEM S, NASR M M, et al. Deepfake audio detection using feature-based and deep learning approaches[J]. International Journal of Advanced Computer Science & Applications, 2025, 16(6): 1-14.
[45] SALVI D, BESTAGINI P, TUBARO S, et al. Exploring the synthetic speech attribution problem through data-driven detectors[C]//Proceedings of International Workshop on Information Forensics and Security. [S.l.]: IEEE, 2022: 1-6.
[46] LI X, LI N, WENG C, et al. Replay and synthetic speech detection with Res2Net architecture[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2021: 6354-6358.
[47] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
ZHANG Xuegong. Introduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1): 32-42.
[48] 张雄伟, 李嘉康, 孙蒙, 等. 语音欺骗检测方法的研究现状及展望[J]. 数据采集与处理, 2020, 35(5): 807-823.
ZHANG Xiongwei, LI Jiakang, SUN Meng, et al. Speech anti-spoofing: The state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2020, 35(5): 807-823.
[49] LI Y, ZHANG X, SUN M. A unified speech enhancement approach to mitigate both background noises and adversarial perturbations[J]. Information Fusion, 2023, 95: 372-383.
[50] 张雄伟, 张星昱, 孙蒙, 等. 说话人验证系统攻击方法的研究现状及展望[J]. 数据采集与处理, 2021, 36(5): 831-849.
ZHANG Xiongwei, ZHANG Xingyu, SUN Meng, et al. Attack methods in speaker verification system: The state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2021, 36(5): 831-849.
[51] 张雄伟, 葛晓义, 孙蒙, 等. 音频隐写方法综述: 从传统到深度学习[J]. 数据采集与处理, 2023, 38(5): 995-1016.
ZHANG Xiongwei, GE Xiaoyi, SUN Meng, et al. An overview of audio steganography methods: From tradition to deep learning[J]. Journal of Data Acquisition and Processing, 2023, 38(5): 995-1016.
[52] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE Computer Society, 2016: 770-778.
[53] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Annual Conference on Neural Information Processing Systems. Long Beach: [s.n.], 2017: 5998-6008.
[54] ZHAI X, UNTERTHINER T, DEHGHANI M, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//Proceedings of International Conference on Learning Representations. Virtual Event: [s.n.], 2021: 22-43.
[55] CHOI J, IM S. CNN and MKDE-based classification of synthetic speech attribution[C]//Proceedings of International Technical Conference on Circuits/Systems, Computers, and Communications. [S.l.]: IEEE, 2023: 1-5.
[56] NERI M, FERRAROTTI A, LUISA L D , et al. ParalMGC: Multiple audio representations for synthetic human speech attribution[C]//Proceedings of European Workshop on Visual Information Processing. Lisbon: IEEE, 2022: 1-6.
[57] RAHMAN M A, PAUL B, SARKER N H, et al. Syn-Att: Synthetic speech attribution via semi-supervised unknown multi-class ensemble of CNNs[EB/OL]. (2023-09-15). https://doi.org/10.48550/arXiv.2309.08146.
[58] BHAGTANI K, YADAV A K S, XIANG Z, et al. FGSSAT: Unsupervised fine-grain attribution of unknown speech synthesizers using transformer networks[C]//Proceedings of Asilomar Conference on Signals, Systems, and Computers. [S.l.]: IEEE, 2023: 1135-1140.
[59] GEMMEKE J F, ELLIS D P W, FREEDMAN D, et al. Audio Set: An ontology and human-labeled dataset for audio events [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans: IEEE, 2017: 776-780.
[60] KOUTINI K, SCHL J, EGHBAL-ZADEH H, et al. Efficient training of audio transformers with patchout[C]//Proceedings of Annual Conference of the International Speech Communication Association. [S.l.]: ISCA, 2022: 2753-2757.
[61] YADAV A K S, BARTUSIAK E R, BHAGTANI K, et al. Synthetic speech attribution using self supervised audio spectrogram transformer[J]. Electronic Imaging, 2023, 35: 372.
[62] PANAYOTOV V, CHEN G, POVEY D, et al. Librispeech: An ASR corpus based on public domain audio books[C]//

Proceedings of International Conference on Acoustics, Speech and Signal Processing. South Brisbane: IEEE, 2015: 5206-5210.

[63] GONG Y, LAI C J, CHUNG Y, et al. SSAST: Self-supervised audio spectrogram transformer[C]//Proceedings of 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI, 2022: 10699-10709.

[64] ZHANG Q, ZHANG X, SUN M, et al. A transformer-based deep learning approach for recognition of forgery methods in spoofing speech attribution[J]. Applied Soft Computing, 2025, 171: 112798.

[65] ZHANG Q, SUN M, YANG J, et al. Unknown forgery method recognition: Beyond Classification[C]//Proceedings of International Conference on Communication, Image and Signal Processing. [S.l.]: IEEE, 2023: 125-132.

[66] BARTUSIAK E R, DELP E J. Transformer-based speech synthesizer attribution in an open set scenario[C]//Proceedings of 21st International Conference on Machine Learning and Applications. Nassau: IEEE, 2022: 329-336.

[67] HASSANI A, WALTON S, SHAH N, et al. Escaping the big data paradigm with compact transformers[J]. arXiv preprint arXiv: 2104.05704, 2021.

[68] GRILL B. The MPEG-4 general audio coder[C]//International Conference of High-Quality Audio Coding. [S.l.]: Audio Engineering Society, 1999: 25-48.

[69] YAN X, YI J, TAO J, et al. An initial investigation for detecting vocoder fingerprints of fake audio[C]//Proceedings of International Workshop on Deepfake Detection for Audio Multimedia. [S.l.]: [s.n.], 2022: 61-65.

[70] BAGHERINEZHAD H, RASTEGARI M. LCNN: Lookup-based convolutional neural network[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Computer Society, 2017: 7120-7129.

[71] ZHANG Q, ZHANG X, YANG J, et al. Introducing euclidean distance optimization into softmax loss under neural collapse [J]. Pattern Recognition, 2025, 162: 111400.

[72] ZHANG Q, ZHANG X, SUN M, et al. A soft-contrastive pseudo learning approach toward open-world forged speech attribution[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 1135-1148.

[73] 张强, 张雄伟, 孙蒙, 等. 基于鲁棒对抗防御边界的语音伪造方法识别[J]. 电子学报, 2025, 53(6): 2022-2037.
ZHANG Qiang, ZHANG Xiongwei, SUN Meng, et al. Robust adversarial defense boundary-based speech forgery method recognition[J]. Acta Electronica Sinica, 2025, 53(6): 2022-2037.

[74] ZHU T, WANG X, QIN X, et al. Source tracing: Detecting voice spoofing[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. [S.l.]: [s.n.], 2022: 216-220.

[75] 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 Annual Conference of the International Speech Communication Association. [S.l.]: [s.n.], 2020: 1496-1500.

[76] SHI J, TIAN J, WU Y, et al. ESPnet-Codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech[C]//Proceedings of IEEE Spoken Language Technology Workshop. [S.l.]: IEEE, 2024: 562-569.

[77] KUMAR R, SEETHARAMAN P, LUEBS A, et al. High-fidelity audio compression with improved RVQGAN[J]. Advances in Neural Information Processing Systems, 2023, 36: 27980-27993.

[78] XIE Y, LU Y, FU R, et al. The codecfake dataset and countermeasures for the universally detection of deepfake audio[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025(33): 386-400.

[79] BABU A, WANG C, TJANDRA A, et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale[C]//Proceedings of Annual Conference of the International Speech Communication Association. Incheon: ISCA, 2022: 2278-2282.

[80] KLEIN N, CHEN T, TAK H, et al. Source tracing of audio deepfake systems[C]//Proceedings of Conference of the International Speech Communication Association. [S.l.]: [s.n.], 2024: 1100-1104.

[81] KAWA P, PLATA M, CZUBA M, et al. Improved deepake Detection using whisper features[EB/OL]. (2023-06-22). https://doi.org/10.48550./arXiv 2306.01428.

[82] BAEVSKI A. Wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in Neural Information Processing Systems, 2020, 33: 12449-12460.

[83] JUNG J W, HEO H S, TAK H, et al. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022: 2405-2409.

[84] CAI D, CAI Z, LI M. Identifying source speakers for voice conversion based spoofing attacks on speaker verification systems [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2023: 1-5.

[85] XU L, YI J, WANG T, et al. Residual speaker representation for one-shot voice conversion[C]/Proceedings of Interspeech. [S.l.]: [s.n.], 2024: 2760-2764.
[86] WELLINGTON S, LIU X, YAMAGISHI J. Quantifying source speaker leakage in one-to-one voice conversion[C]// Proceedings of International Conference of the Biometrics Special Interest Group. [S.l.]: IEEE, 2024: 1-6.
[87] YUAN R, WU Y, LI J, et al. DeID-VC: Speaker de-identification via zero-shot pseudo voice conversion[C]//Proceedings of Interspeech. [S.l.]: [s.n.], 2022: 2593-2597.
[88] WANG Y, WU H, HUANG J. Verification of hidden speaker behind transformation disguised voices[J]. Digital Signal Processing, 2015, 45: 84-95.
[89] ZHENG Linling, LI Jiakang, SUN Meng, et al. When automatic voice disguise meets automatic speaker verification[J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 824-837.
[90] LI Y, LIN X, YANG J. PSVRF: Learning to restore pitch-shifted Voice without reference[J]. arXiv preprint arXiv: 2210.02731, 2022.
[91] 王永全, 施正昱, 张晓. 基于DC-CNN的电子伪装语音还原研究[J]. 计算机科学, 2019, 46(8): 183-188.
WANG Yongquan, SHI Zhengyu, ZHANG Xiao. Study on restoration of electronic disguised voice based on DC-CNN[J]. Computer Science, 2019, 46(8): 183-188.
[92] GUO C, CHEN L, LI Z, et al. On the generation and removal of speaker adversarial perturbation for voice-privacy protection [C]//Proceedings of IEEE Spoken Language Technology Workshop. [S.l.]: IEEE, 2024: 1-6.
[93] DENG J, CHEN Y, ZHONG Y, et al. Catch you and I can: Revealing source voiceprint against voice conversion[C]// Proceedings of USENIX Security Symposium. [S.l.]: [s.n.], 2023: 5163-5180.
[94] ZHANG J, ZHANG X, SUN M, et al. Target speaker filtration by mask estimation for source speaker traceability in voice conversion[J]. Engineering Applications of Artificial Intelligence, 2024, 136: 109071.
[95] MOHAN B, SRIVASTAVA L, VAUQUIER N, et al. Evaluating voice conversion-based privacy protection against informed attackers[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2020: 2802-2806.
[96] LIU W, LI J, WEI C, et al. A novel method to evaluate the privacy protection in speaker anonymization[C]//International Conference on Artificial Intelligence and Security. [S.l.]: [s.n.], 2022: 627-636.
[97] WANG Q, GUO H, KANG J, et al. Speaker contrastive learning for source speaker tracing[C]//Proceedings of IEEE Spoken Language Technology Workshop. [S.l.]: IEEE, 2024: 1247-1253.
[98] MA X, LU W, ZHANG R, et al. Distillation-based feature extraction algorithm for source speaker verification[C]// Proceedings of IEEE Spoken Language Technology Workshop. [S.l.]: IEEE, 2024: 1240-1246.
[99] REN Y, ZHU H, ZHAI L, et al. Who is speaking actually? Robust and versatile speaker traceability for voice conversion[C]// Proceedings of ACM International Conference on Multimedia. [S.l.]: ACM, 2023: 8674-8685.
[100] SHI Y, SAGDUYU Y, GRUSHIN A. How to steal a machine learning classifier with deep learning[C]//Proceedings of IEEE International Symposium on Technologies for Homeland Security. [S.l.]: IEEE, 2017: 1-5.

作者简介:

张雄伟(1965-),男,教授,研究方向:智能语音处理和信息内容安全,E-mail:xwzhang9898@163.com。

张强(1991-),**通信作者**,男,博士后,研究方向:智能语音处理和伪造语音反制,E-mail:zq308297543@126.com。

孙蒙(1984-),男,副教授,研究方向:智能语音处理和机器学习,E-mail:sunmeng@aeu.edu.cn。

杨吉斌(1978-),男,副教授,研究方向:智能语音处理和声源定位,E-mail:yjbice@sina.com。

李毅豪(1996-),男,讲师,研究方向:智能语音处理和语音增强,E-mail:liyihao@aeu.edu.cn。

葛晓义(1995-),男,讲师,研究方向:智能信息处理与信息隐藏,E-mail:lgd_gxy@163.com。

(编辑:夏道家)

Speech Deepfake Attribution: The State of the Art and Prospects

ZHANG Xiongwei[1], ZHANG Qiang[1*], SUN Meng[1], YANG Jibin[1], LI Yihao[1], GE Xiaoyi[2]

(1. Army Engineering University of PLA, Nanjing 210007, China; 2. Information Support Force Engineering University, Wuhan 430000, China)

Abstract: With the rapid evolution of generative artificial intelligence, speech deepfake technologies have achieved unprecedented realism, enabling the synthesis of highly natural and speaker-specific speech from only a few seconds of reference audio. While traditional countermeasures have primarily focused on binary detection—such approaches are insufficient for forensic investigation, legal accountability, and security governance. In real-world adversarial scenarios, it is not enough to determine whether speech is fake; it is equally critical to identify how it was generated, whose voice characteristics were exploited, and which specific model instance may have been involved. This paradigm shift from "detection" to "attribution" marks a fundamental transformation in speech security research. This paper presents a comprehensive survey of speech deepfake attribution, systematically organizing the field into a hierarchical forensic framework that includes three progressive tasks: forgery method attribution, source speaker attribution, and model inversion. Forgery method attribution aims to identify the generative architecture or vocoder family responsible for producing the fake speech by exploiting intrinsic "model fingerprints" embedded in spectral, temporal, and phase domains. Source speaker tracing focuses on recovering or verifying the identity of the original speaker whose voice was converted, leveraging residual prosodic, behavioral, and physiological cues that survive imperfect disentanglement in voice conversion systems. Model inversion represents a deeper forensic objective, attempting to infer specific model parameters or configurations from generated speech, thereby bridging the gap between class-level attribution and instance-level accountability. From both the perspectives of generative model mechanisms and physical acoustic characteristics of speech signals, the feasible core principles for each subtask are elaborated. Different dimensions, such as architectural frameworks and training strategies, are distinguished to systematically organize the research status, mainstream methodologies, and technological evolution paths of each subtask. Furthermore, benchmark datasets and evaluation metrics for both closed-set and open-set scenarios are systematically summarized. Finally, the paper discusses emerging challenges such as open-world generalization, robustness under complex channel distortions and neural codecs, adversarial attacks, and ethical constraints related to privacy and legal admissibility. Future directions are outlined toward proactive traceability, model-level reverse engineering, robust feature disentanglement, and the integration of active watermarking with passive forensic techniques. The survey aims to provide a structured roadmap for advancing speech deepfake attribution and fostering a trustworthy digital speech ecosystem.

Highlights:

1. A hierarchical framework for speech deepfake attribution is systematically established, unifying forgery method attribution, source speaker tracing, and model inversion into a progressive forensic paradigm beyond binary real/fake detection.
2. The intrinsic mechanisms of attribution are analyzed from generative model fingerprints and acoustic signal characteristics, revealing how architectural design, training strategies, and inference processes leave distinguishable trace patterns.
3. Open-world robustness, complex channel conditions, and model instance reverse engineering are identified as key challenges, with future directions proposed toward proactive traceability and a comprehensive speech security defense ecosystem.

Key words: speech deepfake; speech forgery method attribution; source speaker attribution; model inversion; open-set recognition

Foundation items: National Natural Science Foundation of China (Nos.62371469, 62071484).

Received: 2026-01-10; **Revised**: 2026-02-27

***Corresponding author**, **E-mail**: zq308297543@126.com.