基于视觉 Transformer 的鲁棒伪造语音检测算法

张 桐[1], 邓俊龙[1], 任延珍[1,2], 王丽娜[1,2]

[1] 武汉大学国家网络安全学院 武汉 中国 430072
[2] 空天信息安全与可信计算教育部重点实验室 武汉 中国 430072

摘要 在飞速发展的信息时代和数据时代, 网络攻击对个人隐私、工作生活乃至生命财产安全带来了严重威胁。而主机作为人类进行日常工作交流、生活娱乐、数据存储的重要设备, 成为网络攻击的主要目标。因此, 进行主机攻击发现技术的研究是紧迫且必要的, 而主机事件作为记录主机中一切行为的载体, 成为当今网络攻防领域的重点研究对象。攻击者在主机中的各种恶意操作会不可避免地被记录为主机事件, 但恶意事件隐藏在规模庞大的正常事件中难以察觉和筛选, 引发了如何获取主机事件、如何识别并提取恶意事件、如何还原攻击过程、如何进行安全防护等一系列问题的学术研究。本文对基于主机事件的攻击发现技术相关研究进行了广泛的调研和细致的汇总, 对其研究发展历程进行了梳理, 并将本文所研究的基于主机事件的攻击发现技术与入侵检测、数字取证两大研究方向从分析对象、分析方法、作用时间、分析目的 4 个方面进行了对比, 阐明了本文所研究问题的独特之处, 并对其下定义。随后, 本文对基于主机事件的攻击发现技术涉及的关键概念进行了解释, 提出了该领域面临的依赖关系爆炸和及时性两大问题, 并将研究按照阶段划分为主机事件采集、主机事件处理、主机事件分析三个类别, 分别介绍了三个类别围绕两大问题共计 12 个细分方向的研究成果和进展, 最后结合研究现状提出了主机事件记录的完整性和可信性、攻击发现的时效性、跨设备的攻击发现、多步骤攻击的发现、算法的运用等 5 个未来可能的研究方向。

关键词 深度伪造; 伪造语音检测; 数据失配; 鲁棒检测; 可解释性
中图法分类号 TP309 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.03.02

Robust Fake Audio Detection Algorithm based on Vision Transformer

ZHANG Tong[1], DENG Junlong[1], REN Yanzhen[1,2], WANG Lina[1,2]

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
[2] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, Wuhan 430072, China

Abstract With the rapid development of deep learning and speech synthesis technology, speech deepfake technology has made fake speech realistic in terms of naturalness and emotion, posing a great threat to social security. In order to resist the threats to security and privacy brought by these fake technologies, fake speech detection technology based on deep learning has received great attention from researchers and has achieved good performance, but there are still problems with robustness and poor interpretability. Performance degradation occurs significantly when there is a mismatch between training data and actual detection data, and there is a lack of interpretability as existing detection techniques do not provide analysis of detection results. Addressing the issue of poor performance of existing deepfake speech detection techniques under various data mismatch scenarios, this paper proposes a robust speech deepfake detection scheme based on Vision Transformer, which optimizes the entire detection algorithm from both frontend feature extraction and backend neural network aspects. In terms of feature extraction, this paper introduces a frontend feature extractor based on self-supervised learning, which fine-tunes existing generic pre-trained models using labeled data to learn better intermediate speech representations. For the backend neural network, this paper extends Vision Transformer to deepfake speech detection task, decomposing the original positional encoding into time positional encoding and frequency positional encoding. Leveraging the powerful representation capability of Transformer architecture, better feature representations are learned to capture artifacts in the speech to be detected. Experiments indicate that in various complex data mismatch scenarios, our method reduces the detection EER (Equal Error Rate) by 1% to 20% compared to existing methods, exhibiting improved robustness. Additionally, the paper utilizes the attention mechanism of Transformer model to provide interpretability analysis of the decision-making process of the deepfake speech detection model, thus possessing significant practical value.

Key words deepfake; fake audio detection; data mismatch; robust detection; interpretability

通讯作者: 任延珍, 博士, 教授, Email: renyz@whu.edu.cn。
该工作受到国家自然科学基金(NSFC)(No. 62572358, No. 62172306, No. 62372334)的支持。
收稿日期: 2024-08-26; 修改日期: 2024-12-24; 定稿日期: 2026-01-26

1 引言

随着深度学习和语音合成技术的飞速发展, 合成的语音不仅在自然度上达到了逼真水平, 而且在表达情感和细微语调上也取得了显著进展。尤其是先进的神经网络模型, 如基于 Transformer 架构的 Tacotron[1-2]系列, Fast Speech[3-4]以及 VITS[5]等方法。这些模型学习了大量的语音数据, 并能够捕捉到人类语音的微妙特征, 包括声调、节奏和语调变化, 已经能够生成与真人几乎无法区分的语音。

然而, 正如技术带来便利的同时也伴随着潜在的滥用风险, 伪造的语音被用于不法行为已经成为一个日益严重的问题。因此, 研究人员和技术开发者正在努力创建更为精确的检测工具来识别合成语音。伪造语音检测是一项旨在通过机器学习技术区分真实语音和伪造语音的任务, 已经有许多工作致力于开发更鲁棒的伪造语音检测系统。目前主流的伪造语音检测研究大致可以分为两类: 管道式(Pipeline)检测方法和端到端(End-to-end)检测方法[6]。前者通常包括一个前端特征提取器和一个后端分类器[7], 而后者在近些年受到越来越多的关注, 它采用模型通过直接对原始语音波形进行操作来联合优化特征提取和分类。

随着深度伪造方法的持续发展, 在现实场景中, 语音伪造检测技术面临着更多的困难。现有的伪造语音检测方法在面对域内数据时性能良好, 但是当待检测数据与训练数据之间存在各种不匹配时, 模型的检测性能会急剧下降。训练数据集与待检测数据集的失配主要体现在以下几个方面: 伪造语音生成算法失配、语音编码技术失配、低质量失配以及部分伪造失配等。具体来说, 伪造语音生成算法失配是指现实场景中, 待检测数据集中可能存在许多训练数据集中未出现过的生成算法来生成的语音。语音编码技术失配是指待检测数据集中的语音可能通过多种未在训练集中考虑的编码处理, 而训练数据集中的语音未经编码或使用的语音编码技术有限。低质量场景失配则是指待检测数据中可能存在噪声干扰或声场环境变化干扰, 而训练数据集中的语音则是干净的语音。部分伪造失配是指待检测数据集中的语音是经过部分伪造的, 而训练数据集中的语音是完全伪造的。

上述失配场景会给伪造语音检测技术带来极大的挑战。例如, 电话和网络通信技术会使用不同的语音编解码技术对通信音频进行压缩, 产生传输流畅但不影响人类听觉的语音数据。此过程可能会导致失真并引入类同于伪造痕迹的伪影, 严重影响伪造语音检测系统的性能。如何有效地检测这些域外语音数据和各种低质量的语音数据是目前伪造语音检测技术的研究重点。

现有的伪造语音检测算法在面对各种数据失配场景下的性能不佳, 其主要存在以下问题: 1) 鲁棒性不佳。现有的伪造语音检测算法在应用到多种失配场景下时, 其性能急剧下降, 然而现实场景多是各种失配场景的组合(例如跨数据集、语音压缩编码与噪声环境的组合)。2) 缺少可解释性。现有的伪造语音检测算法无法对检测结果进行可解释性分析。

针对上述两个问题, 本文提出了基于视觉 Transformer 的鲁棒可解释伪造语音检测算法, 并在多种数据失配场景中进行了实验验证。本文的主要贡献包括: 1) 针对现有伪造语音检测算法在应用到多种失配场景下性能急剧下降的问题, 本文提出了基于自监督学习的前端特征提取器和基于视觉 Transformer 的后端分类网络, 在多种失配场景下比现有方法的检测 EER 降低了 1%~20%; 2) 针对现有伪造语音检测技术缺少可解释性的问题, 本文方法利用自注意力机制对伪造语音的检测结果进行了具体的判决分析, 具有良好的可解释性。

2 相关工作

目前主流的伪造语音检测研究大致可以分为两类: 管道式检测方法和端到端检测方法。前者通常包括一个前端特征提取器和一个后端分类器, 且会针对不同场景使用不同的数据增强技术, 其技术框架图如图 1 所示。而后者在近些年受到越来越多的关注, 它采用模型通过直接对原始语音波形进行操作来联合优化特征提取和分类。

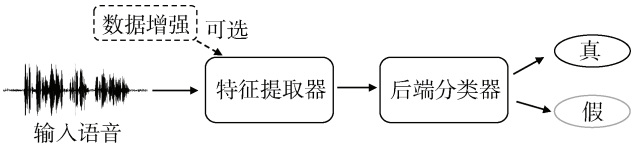

图 1 伪造语音检测技术框架图

Figure 1 Framework of Deepfake Speech Detection

从以上角度分别考虑, 可以将伪造语音检测系统分为三类, 如表 1 所示。

在众多手工提取的声学特征中, 包括恒定 Q 倒谱系数(CQCCs[16])、线性频率倒谱系数(LFCCs[11])和梅尔语谱图(MSTFT[9])等特征, 已在伪造语音检测任务中证明了它们的实用性。除了手工构造的语音特

征, 也有一些深度可学习特征被提出并应用于伪造语音检测任务, 代表的特征有 LEAF[17]、SincNet[18]和 FastAudio[19]等。后端分类器的主要功能是基于提取的特征来确定语音是伪造还是真实的, 轻量级卷积神经网络(LCNN[20])、残差网络(ResNet[21])、ECAPA-TDNN[22]以及图注意力网络(GAT[23])等架构在检测伪造语音方面展现了卓越的性能。

表 1 伪造语音检测研究工作分类汇总

Table 1 Summary of Existing Research on Deepfake Speech Detection

方法	代表研究工作	存在问题
手工提取特征+传统模式识别方法	LFCC+GMM [8] CQCCs+GMM [8]	检测方法简单, 准确率低
手工提取特征+深度神经网络	MSTFT+ResNet [9] CQT+ ResNet [10] LFCC+LCNN [11]	依赖手工提取特征的区分度, 鲁棒性不佳
端到端检测模型	CRNNspoof [12] RawNet2 [13] RawGAT-ST [14] AASIST [15]	依赖数据驱动, 在跨数据集情况下泛化性不佳, 缺少可解释性

为了充分利用深度神经网络强大的表征能力, 端到端检测模型优化了特征提取过程, 采取直接对语音原始波形进行编码, 在伪造语音检测任务中取得了具有竞争力的性能。例如, Tak 等人[13]将其应用到语音伪造检测中并进行了改进。RawNet2 是一种具有残差块的卷积神经网络, 其第一层采用了 Sinc 滤波器, 本质上与 SincNet[18]相同。RawNet2 直接在原始输入语音上进行时域卷积操作, 可以学习到使用基于知识的方法无法检测到的区分特征。Jung 等人[15]利用了 RawNet2 编码器的一个变体来提取高级表示, 然后用这些表示来构建图注意力网络(GAT)。其中, 作者引入了一个新的异质性堆叠图注意力层, 该层包括一个异质性注意力机制和一个堆栈节点, 以模拟跨不同时间和频谱间隔的人工效应。受到先前研究的启发, Tak 等人[24]用更先进的 Wav2Vec2.0[25]模型替换了 Sinc 层[18]前端, 并集成了一个基于自注意力的聚合层来优化他们的方法。同时, 为了进一步改进结果, Tak 等人采用了他们之前工作中的一种数据增强方法[26], 获得了 ASVspoof 2021 挑战赛[8]数据集上的最佳结果。

尽管上述方法有着较好的检测性能, 但普遍存在对训练数据过拟合的问题, 因此在面对多种数据失配场景时, 存在鲁棒性不佳的问题; 现有方法普遍在检测时仅输出真假二分类结果, 存在缺乏可解释性的问题。本文针对上述两个问题提出了针对性的解决方法。

3 本文方法

3.1 研究思路与总体框架

针对目前现有伪造语音检测技术在多种数据失配场景下检测性能不佳的问题, 本文提出了一种基于视觉 Transformer 的鲁棒伪造语音检测算法, 从前端特征提取和后端神经网络两个方面优化整个检测算法。

目前已经有一系列不同任务的通用神经嵌入表示[25, 27]方面的应用证明, 使用适量的标记数据对已有的通用预训练模型进行微调可以获得更好的高级特征表示。在不同领域, 使用自监督学习派生出的预训练模型可以很好地泛化许多不同的任务, 其中包括自动语音识别、说话人识别和情感识别等, 因此探索基于自监督学习的前端特征以提高伪造语音检测技术的性能具有可行性。基于以上分析, 本文提出使用基于自监督学习的前端特征提取器, 通过使用标记数据对已有的通用预训练模型进行微调, 学习到更鲁棒的中间语音特征, 这可以帮助减少过拟合, 从而提高泛化性和鲁棒性, 尤其是在面对之前未出现过的伪造算法时。

在后端神经网络的选择上, 由于 Transformer 架构已经被成功应用到计算机视觉领域, 通过从输入图像中提取小块并为每个块添加可学习的位置编码, 生成的块形成一个序列, 之后馈送到 Transformer, 这些视觉 Transformer 模型在图像分类任务上取得了最先进的性能[28]。同时, 在声音事件分类任务上, 通过使用计算机视觉的预训练模型和对音频谱图的重叠块进行微调[29], 实现了最先进的性能。利用 Transformer 架构强大的表征能力, 学习到更好的特征表示, 以捕获待检测语音中的伪影。此外, Transformer 模型的注意力机制可以为伪造语音检测模型的决策过程提供一定程度的可视化解释, 例如, 可以观察模型在处理特定语音或语言任务时关注的区域。基于以上分析, 本文提出的基于视觉 Transformer 的鲁棒可解释伪造语音检测算法的框架图如图 2 所示。

3.2 基于自监督学习的前端特征提取器

本文使用 Wav2vec 2.0 预训练模型作为基于自监督学习的前端特征提取器, Wav2vec2.0 预训练模型用于从原始输入波形 $x_{1:L}$ 中提取一系列特征表示 $o_{1:N}$, 其中 L 是样本数。如图 2(a)所示, 基于自监督学习的前端特征提取器由卷积神经网络(Convolutional neural network, CNN)、Transformer 网络组成, 前者将输入 $x_{1:L}$ 转换为隐藏特征序列 $z_{1:N}$, 而后者将 $z_{1:N}$ 转换为输出序列 $o_{1:N}$。L 和 N 之间的比率由 CNN 网络步长(20ms)决定。

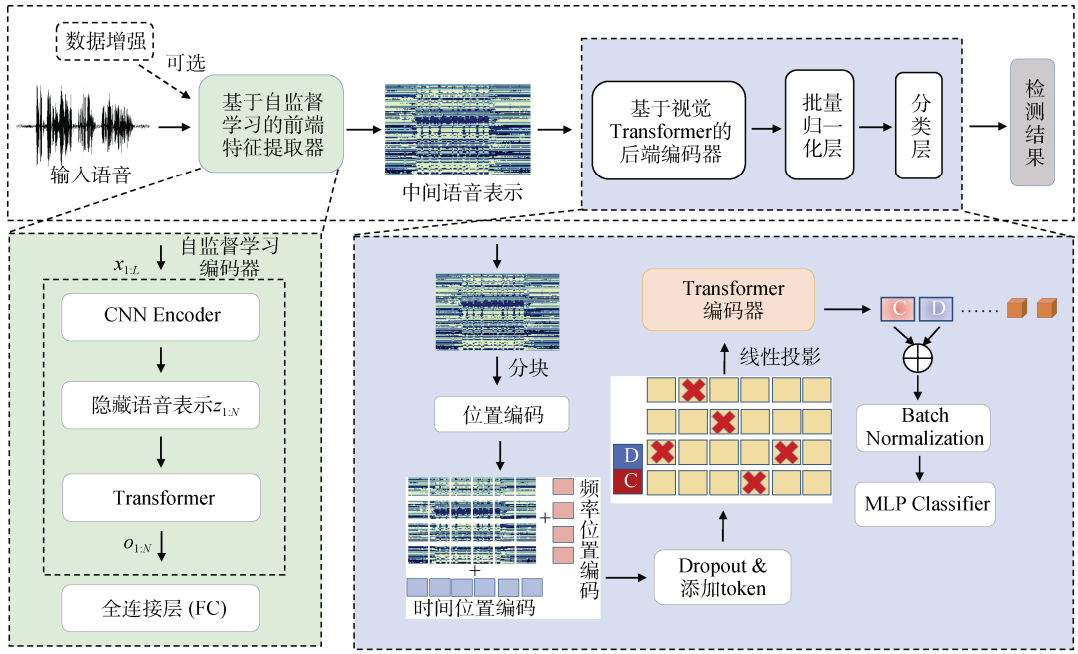

图 2　基于视觉 Transformer 的鲁棒伪造语音检测算法框架图

Figure 2　Framework of Proposed Robust Fake Audio Detection Algorithm based on Vision Transformer

3.2.1　预训练过程

Wav2vec2.0 编码器的预训练过程如图 3(a) 所示, 输入音频 $x_{1:L}$ 被 CNN 编码器转换为隐藏特征序列 $z_{1:N}$, 然后隐藏特征序列被量化为表示 $q_{1:N}$, 接着将隐藏特征序列馈送至 Transformer 网络, 其中部分隐藏特征被掩码。然后计算每个掩码时间步 n 的对比损失, 以约束在给定相应上下文向量 c_n 的情况下, 目标 q_n 在一组干扰项(即从其他被掩码的时间步 n' 中采样的 $q_{n'}$, 其中 $n' \neq n$)中能够被准确地识别。

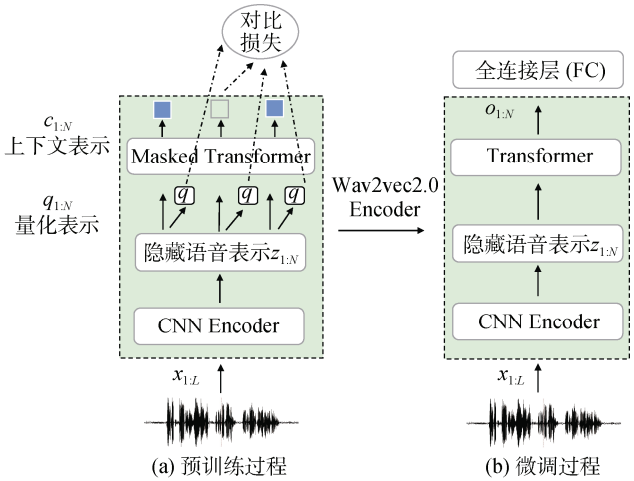

图 3　基于自监督学习的前端特征提取器预训练和微调框架图

Figure 3　Overview of Pretraining and Fine-tuning of Self-supervised Front-end Feature Extractor

3.2.2　微调过程

由于 Wav2vec2.0 模型的预训练过程仅使用真实数据, 因此本文通过使用下游伪造语音检测任务中真实的和伪造的训练数据进行微调来改进模型的性能。微调过程如图 3(b) 所示, 与预训练不同, 微调过程中不对隐藏特征 $z_{1:N}$ 进行输入掩码。并且, 在 Wav2vec 2.0 的 Transformer 编码器输出 $o_{1:N}$ 之后添加了一个全连接层, 以减小最终的特征表示维度以匹配后端分类网络的输入维度。整个微调过程中, 预训练的 Wav2vec 2.0 模型与后端分类网络一起通过反向传播使用训练数据集进行优化, 使用加权交叉熵目标函数来最小化训练损失。

3.3　基于视觉 Transformer 的后端分类网络

Vision Transformer(ViT)的工作原理是从输入图像中提取小块, 并将这些块线性投影到一系列嵌入中, 通过将可训练的位置编码添加为输入序列的偏置来增强该序列。在自注意层之后, 将一个特殊的分类嵌入(分类标记)附加到序列中, 与分类层相连, 作为最后分类的依据。除此之外, 一些工作还额外添加了另一个用于蒸馏的特殊嵌入(蒸馏标记)。本文将语音谱图看成一张通道数为 1, 宽度为频率 bins, 高度为时间帧数的图像。

视觉 Transformer 在图像分类任务中针对图像数据的空间相关性引入了一维位置编码以表征输入图像块的空间位置关系。与计算机视觉任务不同的是,

语音谱图具有频率和时间两个不同维度, 不仅音频帧之间存在时间相关性, 不同频带之间也存在一定的结构关系和相关性。基于以上观察并受[30]启发, 本文将原本的位置编码分解成频率和时间两个维度。分解的位置编码一方面可以使得在短语音长度的下游任务上对预训练模型进行微调或推断变得更简单, 因为只需裁剪时间位置编码参数, 而无需更改频率位置编码参数。另一方面, 语音谱图切分成块之后, 每一块在频率和时间上都具有不同的位置, 分解的位置编码有利于模型理解语音谱图中的时间-频带结构信息, 从而可以更好地捕获待检测语音中的伪影。

基于视觉 Transformer 的后端分类网络结构如图 2(b)所示。输入语音经过基于自监督学习的前端特征提取器后, 得到中间语音特征, 并通过一个全连接层来减小其维度以匹配后端分类网络的输入维度。中间语音特征馈送至后端分类网络后, 首先将其切分为固定大小的重叠块, 并分别在频率和时间两个维度为其添加位置编码。接着对这些块进行随机 dropout 并添加分类 token 和蒸馏 token。随机 dropout 的目的是减少序列长度对训练 Transformer 模型计算复杂度的影响, 同时, 在训练期间丢弃输入序列的部分, 可以鼓励模型使用不完整的序列执行分类, 可以有效地正则化训练过程, 从而增强模型的泛化性。接着将这些重叠块线性投影到特征向量, 输入到 Transformer 编码器。

Transformer 编码器的结构如图 4 所示, 本文仅使用其编码器结构, 一共包含 12 个基本块。每一块都由两个子层连接结构组成, 第一个子层连接结构包括一个多头自注意力层和残差归一化层, 第二个子层连接结构包括一个前馈网络层和残差归一化层。每个多头自注意力层包含 12 个自注意力头, 这些自注意力头在训练过程中会关注输入谱图中不同的块, 捕捉它们之间的关系, 从而处理长距离依赖的语音序列。经过 Transformer 编码器进一步提取的高级表示, 将会被馈送至批量归一化层(BN)。批量归一化层用于加速神经网络的训练过程, 提高训练过程的稳定性。通过对每个小批量数据进行归一化处理, 使网络中间层输入的分布更加稳定。最后将分类 token 和蒸馏 token 的平均值输入到分类层进行分类预测, 得到模型最终的输出结果。在分类网络的训练过程中, 由于 Transformer 架构往往需要大量的训练数据, 故本文使用在声音事件检测任务上的预训练模型[30], 并使用伪造语音检测任务的数据集进行微调。

4 实验及结果分析

为了评估本文所提出的方法在各种失配场景下的检测性能, 本章设计了三类实验: 域内和语音编解码失配场景检测性能分析、组合失配场景检测性能分析和部分伪造失配场景检测性能分析。在每个实验中, 分别对比了本文方法与其他现有方法的性能, 采用前文所介绍的等错误率(EER)进行检测性能的评估。通过与现有的方法进行比较, 验证了本文方法的有效性, 并在最后讨论了本文方法对于检测结果的可解释分析。

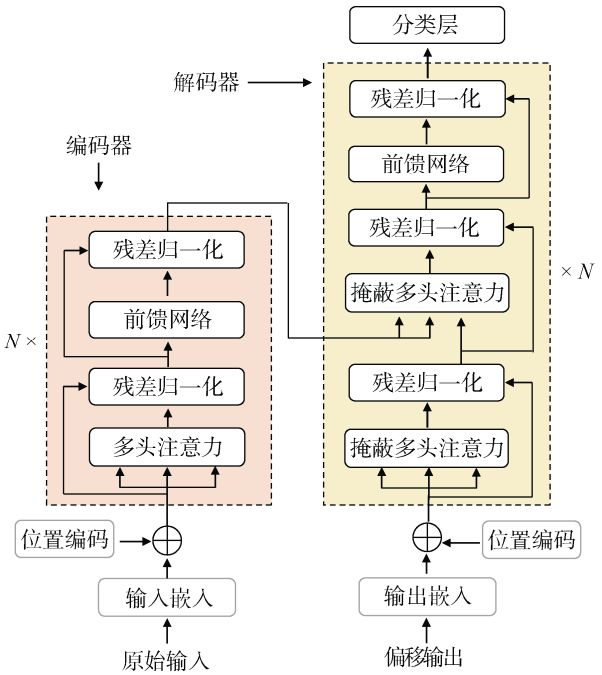

图 4 Transformer 编码器和解码器框架图

Figure 4 Overview of Transformer Encoder and Decoder

4.1 数据集与评价指标

目前伪造语音检测领域最权威的数据集是 ASVspoof 系列挑战赛发布的数据集, 本文主要以 ASVspoof 2019 挑战赛 LA 赛道的训练集和验证集作为模型训练的数据。而评估数据集一共有五个部分, 分别为 ASVspoof 2019 LA 赛道的评估数据集, ASVspoof 2021 LA 赛道的评估数据集, ASVspoof 2021 DF 赛道的评估数据集, In-the-Wild 数据集[31]以及 PartialSpoof 数据集[32]。具体来说, 各个失配场景下的数据集组成如表 2 所示。

本文所涉及的所有实验均采用等错误率(Equal Error Rate, EER)来作为性能的评价指标。它表示当错误接受率(False Acceptance Rate, FAR)和错误拒绝率(False Rejection Rate, FRR)相等时的值。伪造语音检测模型在各个场景下的等错误率越低代表该模型的检测性能越好, 针对数据集失配场景的鲁棒性越好。

表 2 各失配场景下实验数据集的具体介绍

Table 2 Introduction of Experimental Datasets in Various Mismatch Scenarios

失配场景	数据集组成		数据集具体介绍
	训练数据集	评估数据集	
域内场景	ASVspoof 2019 LA	ASVspoof 2019 LA	评估数据集存在一些训练集中未出现过的伪造算法生成的语音。可以用来测试伪造语音检测模型在域内数据集的检测性能
语音编解码失配场景	ASVspoof 2019 LA	ASVspoof 2021 LA	评估数据集中的语音不仅存在训练集中未出现过的伪造算法生成的语音, 还进一步做了语音后处理, 具体包括各种信道传输编码
	ASVspoof 2019 LA	ASVspoof 2021 DF	与 2021 LA 赛道不同, 这部分数据集包含了经过各种语音压缩编码后的真实和伪造语音的集合
组合失配场景	ASVspoof 2019 LA	In-the-Wild	模拟真实世界中的语音, 评估数据集是从各个社交媒体或直播平台上收集的, 包含许多名人和政治家。其中存在各种未见过的伪造方法生成的语音, 并且可能经过各种语音编解码和噪声处理
部分伪造失配场景	ASVspoof 2019 LA	PartialSpoof	评估数据集中包含部分伪造的语音, 其中的语音片段可以来自真实的语音片段, 也可以来自其使用 TTS 或 VC 技术生成的语音片段

4.2 实验设置

4.2.1 域内和语音编解码失配场景检测性能分析

在域内数据性能测试的实验中, 用于训练伪造语音检测模型的数据集和评估数据集之间不存在由于语音编解码和噪声干扰等引入的失配。具体来说, 训练集和验证集包含从六种语音伪造算法(A01-A06)生成的语音, 而评估集包含从 13 种算法(A7-A19)生成的语音。这部分实验可以用于评估本文所提出的算法在仅存在伪造算法失配, 而不存在其他失配场景下的检测性能。

在语音编码失配场景性能测试的实验中, 训练数据集与前面保持一致, 但评估数据集经过了各种语音编解码处理。具体来说, 评估数据集分为 ASVspoof 2021 LA 和 DF 两个部分, 都是以 ASVspoof 2019 LA 评估数据集作为基础生成的。LA 部分的语音数据经过了各种信道传输编解码处理, 包括 VoIP 和 PSTN 等。DF 部分的语音数据全部采用了深度伪造的方法生成, 并使用了不同的有损压缩编解码器进行处理, 通过对语音数据进行压缩编码和解码恢复, 引入了依赖于编解码器类型和配置的失真。除此之外, DF 部分的语音数据还包括了来自不同于 ASVspoof 2019 LA 评估数据集的源生成的语音。这部分实验可以用于评估本文所提出的算法在语音编解码失配场景下的检测性能。

在上面两部分实验中, 本文复现了在 ASVspoof 2021 挑战赛的基线模型, 并且选取了 2021 LA 和 DF 数据集上性能最好的方法进行了比较。具体来说, 除了复现 ASVspoof 2021 挑战赛的基线模型 LFCC+LCNN[11]和 RawNet2[13]外, 本文还选取了 Yang 和 Qin 等人[33]的前端特征融合方法以及 Guo 等人[34]的 WavLM+MFA 检测方法进行比较。同时, 本文针对数据增强方法以及频率位置编码进行了消融实验, 具体而言, 训练阶段在原有方案上分别去除数据增强模块和视觉 Transformer 分类器中的频率位置编码, 与原方案的检测效果进行对比。

4.2.2 组合失配场景检测性能分析

在组合失配场景性能测试中, 本文主要考虑的是在各种失配场景组合的情况下, 本文方法与其他现有方法的性能对比。In-the-Wild 数据集则模拟了现实生活中的真实场景, 该数据集中的真实语音是从主流的社交网络和视频共享平台等公开可用的来源收集的, 其中的语音来自讲英语的名人和政治家, 而伪造语音也是经过各种深度伪造的方法进行生成的。这些从现实世界中收集的数据, 不仅存在噪声干扰, 还存在信道传输编码以及语音压缩编解码等, 最贴近现实场景中的检测需求, 同时也是检测难度最大的场景。

本部分实验中, 训练集和验证集采用 ASVspoof 2019 LA 的训练集和验证集, 而评估集使用 In-the-Wild 数据集。对于对比方法, 本实验采用 LFCC+LCNN[11]和 RawNet2[13]作为对比的基线模型, 以及采用了 Muller 等人[31]在研究中所涉及的 Cqtspec+ResNet18 方法以及使用图神经网络的方法作为额外对比方法。

4.2.3 部分伪造失配场景检测性能分析

PartialSpoof 数据集是一个针对部分伪造(PS)场景构建的特定数据集, 其中的伪造语音数据均是通过部分伪造的。在部分伪造场景中, 单个语音可能包含使用多个 TTS 或 VC 方法生成的语音片段, 这给伪造检测带来了更大的挑战。本实验采用 ASVspoof 2019 LA 训练集和验证集进行训练, 而使用 PartialSpoof 评估数据集进行评估, 旨在验证本文方法在部分伪造失配场景下的性能。而对于对比方法, 本文采用 LFCC+LCNN[11]和 RawNet2[13]作为对比的基线模型, 并且使用了 Zhang 等人[32]提出的在 PartialSpoof 数据

集上性能最好的方法作为对比方法。此外, 本实验也展示了使用 PartialSpoof 的训练集训练的模型的性能, 为后续可解释性分析提供了推断模型。

4.3 实验结果

4.3.1 域内和语音编解码失配场景下的实验结果

本部分的实验结果如表 3 所示。该表对域内和语音编解码失配场景下本文方法和现有方法的检测性能进行了对比分析, 可以看出现有的伪造语音方法, 包括 ASVspoof 2021 挑战赛的基线模型在内, 在针对域内数据(2019 LA), 即仅存在伪造算法失配, 而不存在其他失配场景下失配时, 模型的 EER 都在 5%以下。本文所提出的方法在 2019 LA 和 2021 LA 数据集中分别达到了 0.19%和 0.77%的 EER, 在对比方法中均为最低。而在 2021 DF 数据集的语音编解码失配场景中, 本文提出的方法达到了 2.86%的 EER, 虽比现有最好方法的 EER 高 0.30%, 但仍明显优于其他对比方法。

表 3 域内和语音编解码失配场景下的实验结果, 评价指标: 等错误率(%)

Table 3 Experimental Results of In-domain and Codec Scenarios. Evaluation Metric: Equal Error Rate (EER, %)

方法	评估数据集		
	2019 LA	2021 LA	2021 DF
LFCC+LCNN [11]	3.86	8.90	21.35
RawNet2 [13]	4.29	8.65	23.67
SSL-Fusion [33]	2.07	-	11.78
Tak et al. [24]	0.29	4.37	6.42
Tak et al. + DA [24]	0.27	0.91	4.06
WavLM+MFA [34]	0.42	5.08	**2.56**
本文方法	**0.19**	**0.77**	2.86
本文方法 (无数据增强)	0.24	2.15	3.04
本文方法 (无频率位置编码)	0.36	1.38	15.40

在消融实验中, 当未使用数据增强方法[26]时, 本文方法在 2021 LA 和 DF 数据集上 EER 分别上升了 1.38%和 0.18%; 在未使用频率位置编码时, 本文方法在 2021 LA 和 DF 数据集上的 EER 分别上升了 0.61%和 12.54%, 在去除频率位置编码的情况下, 模型在语音编解码失配场景下检测性能出现显著下降, 证明了本文方法中频率位置编码的有效性。

基于以上分析, 本文所提出的方法在域内和语音编解码失配场景下获得了较好的检测性能, 具有良好的鲁棒性, 并且可以与现有数据增强方法结合进一步提升检测性能。

4.3.2 组合失配场景下的实验结果

本部分的实验结果如表 4 所示, ASVspoof 2021 挑战赛的基线模型在真实场景下的 In-the-Wild 数据集上的检测 EER 均在 50%左右。这表明基线模型在面对现实世界中广泛存在的失配数据时, 性能急剧下降, 其性能基本上等同于随机猜测, 无法可靠地区分真假类别。而即使是 Yang 和 Qin 等人[33]的前端特征融合方法在 In-the-Wild 数据集上检测的 EER 也高达 24%以上, 这证明了现有方法的鲁棒性不佳。本文方法在 In-the-Wild 数据集上达到了 17.78% 的 EER, 对比现有方法中的最好结果降低了 6.49% 的 EER。这一结果说明本文方法在面对现实世界中多种失配场景组合的情况下, 依旧保持了较好的鲁棒性, 优于其他对比方法。

表 4 组合失配场景下的实验结果, 评价指标: 等错误率(%)

Table 4 Experimental Results of In-the-Wild Scenario. Evaluation Metric: Equal Error Rate (EER, %)

训练数据集	方法	评估数据集
		In-the-Wild
ASVspoof 2019 LA	LFCC+LCNN [11]	53.27
	RawNet2 [13]	47.59
	Cqtspec+ResNet18 [10]	49.76
	RawGAT-ST [14]	37.15
	SSL-Fusion [33]	24.27
	本文方法	17.78

4.3.3 部分伪造失配场景下的实验结果

本部分的实验结果如表 5 所示。当使用 ASVspoof 2019 LA 的训练集作为训练数据时, ASVspoof 2021 挑战赛的基线模型在 PartialSpoof 评估数据集上的检测 EER 都达到了 20%以上。Zhang 等人[32]在论文中展示的最好的方法的检测 EER 为 14.19%, 而本文方法可以达到 13.57% 的 EER, 检测性能略有提升。这表明本文方法针对部分伪造失配场景表现出了良好的鲁棒性。

表 5 部分伪造失配场景下的实验结果, 评价指标: 等错误率(%)

Table 5 Experimental Results of Partial Spoof Scenario. Evaluation Metric: Equal Error Rate (EER, %)

训练数据集	方法	评估数据集
		PartialSpoof
ASVspoof 2019 LA	LFCC+LCNN [11]	22.67
	RawNet2 [13]	23.03
	Zhang 等[32]	14.19
	本文方法	13.57

4.4 检测结果的可解释性分析

本文提出的后端分类网络是基于视觉 Transformer 架构的, 其中的注意力机制可以为伪造语音检测模型的决策过程提供一定程度的可视化解释。例如, 可以观察模型在处理特定语音时关注的区域。具体来说, 通过可视化后端检测模型中 self-attention 层各个 token 的注意力权重, 可以获知模型在对音频进行判决时关注的时间位置, 从而定位伪造痕迹所在区域。因此在检测部分伪造语音时, 通过上述可视化结果, 本文的模型可以实现对伪造片段的定位, 具有较好的可解释性。

本文使用了在部分伪造语音数据集 PartialSpoof 训练集上训练的伪造检测模型作为分析模型, 选择了 PartialSpoof 评估集中的语音数据作为分析对象, 经过随机挑选和大量分析, 本文挑选了几份具有代表性的样本作为展示。如图 5 所示, 图中分别展示了语音 CON_E_0012267.wav 的在第 1 层、第 3 层、第 6 层和第 11 层的多头自注意力层的所有注意头(12 个)的 attention map 可视化结果, 该语音的标签为假, 模型预测值也为假。可以看到在模型中靠前的层(第 1 层, 图中左上部分)的 attention 大多只关注自身, 也即进行 self-attention 来理解自身的信息, 其特点就是 attention map 呈现出明显的对角线模式。随后, 模型开始逐渐增大感受野, 以此融合周围的信息, 展示在 attention map 中就是呈现出多条对角线的模式, 如第 3 层(图中右上部分)所示。然后随着层数的加深, 模型开始将重要的信息聚合到某些特定的 token 上, 如第 6 层(图中左下部分)所示。最后, attention 出现与 query 完全无关的情况, 在 attention map 上呈现出竖线的形状, 表明模型的注意力绝大部分集中在了竖线位置对应的 token 上, 如图中右下部分的第 11 层的 attention map。

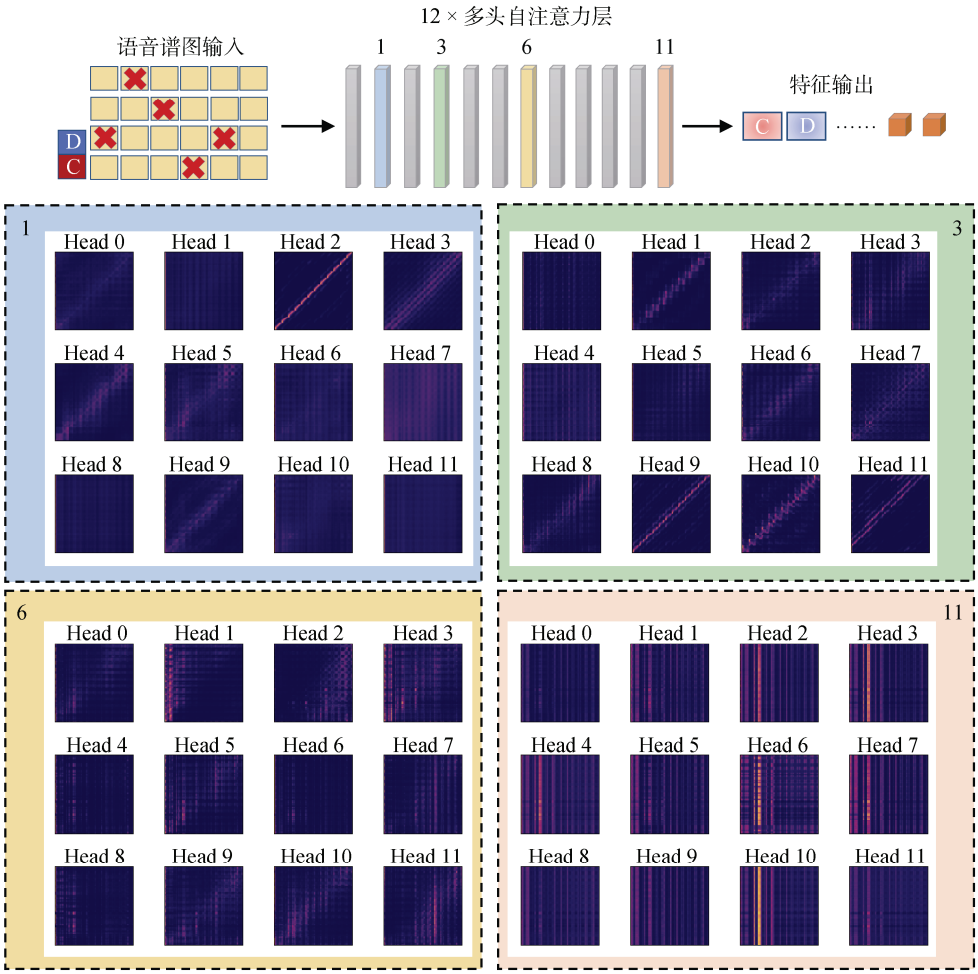

图 5 语音 CON_E_0012267.wav 的 attention map 可视化图

Figure 5 Attention Map Visualization of Speech Input CON_E_0012267.wav

除了在各个不同深度的多头自注意力层之间的 attention map 不同之外, 针对真假类别不同的语音, 同一层的不同自注意力头所关注的区域也不同。具体表现在 attention map 中如图 6 所示。本文发现, 在

针对真实语音(图 6 左侧)时, attention 集中表现在编号为 3, 5, 7 的自注意力头所关注的区域上, 而针对伪造语音(图 6 右侧)时则集中表现在编号为 6, 7, 10 的自注意力头所关注的区域上。这表明模型在进行推断时, 针对不同类别的语音会根据不同的区域来进行判决。

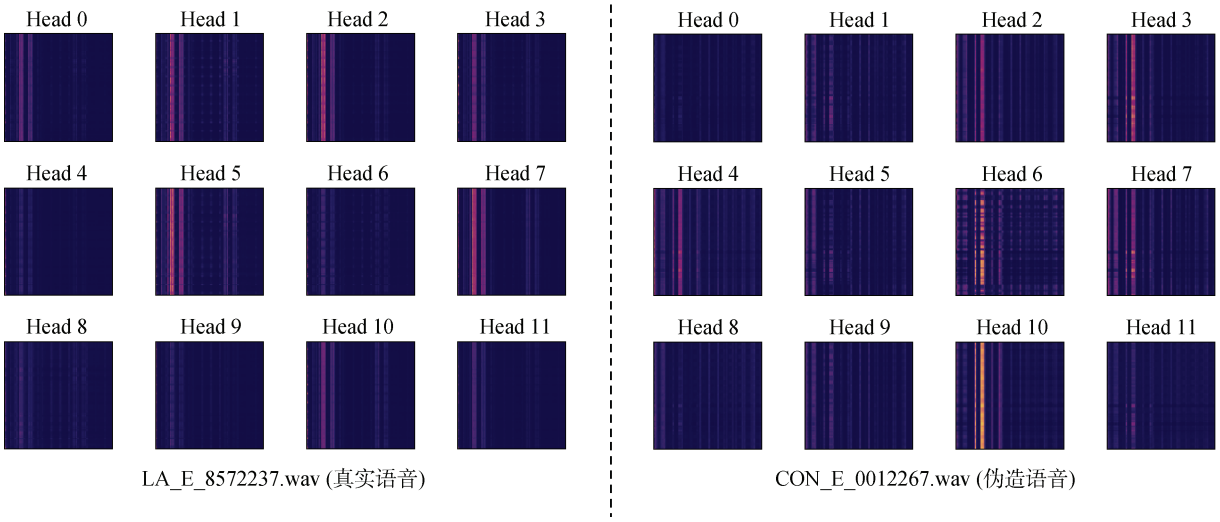

图 6 真假语音的第 11 层 attention map 对比图

Figure 6 Comparison of 11th-layer Attention Maps for Real and Fake Speech

最后, 为了验证本文的模型在检测部分伪造语音时对伪造片段的定位能力, 本文选取了两个部分伪造语音的 attention map 作为实例来进一步分析。如图 7 所示, 上层为 attention map, 中层为片段标签, 下层为梅尔语谱图特征。上层的 attention map 是使用第 11 层的第 10 个自注意力头进行映射的, 区域的颜色越深代表模型越关注该区域。而中层所展示的图例是语音的片段标签, 该标签的分辨率为 160ms, 两段语音长度均为 64600(裁剪后)个采样点, 采样率为 16000, 故这两条语音的长度约为 4037ms, 一共有 25

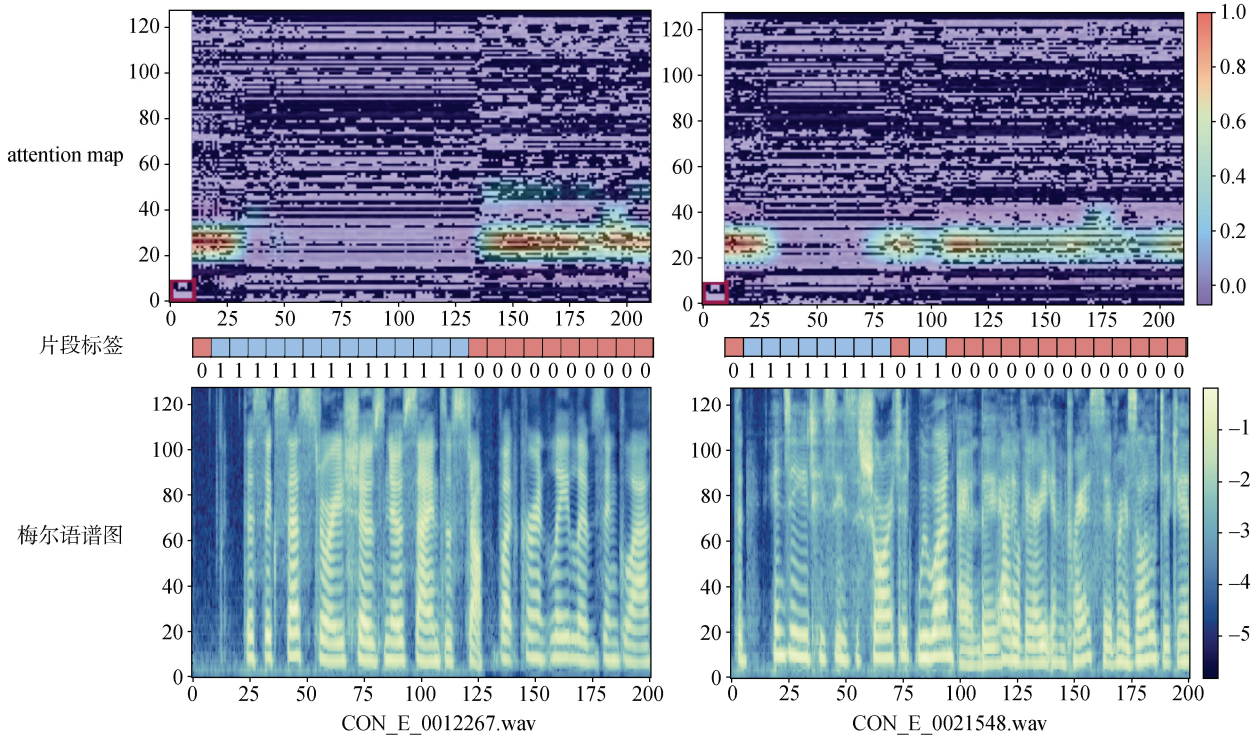

图 7 部分伪造语音的 attention map 细节对比图

Figure 7 Attention Map Details of Partial Spoof Speech

个片段级标签。其中, 标签“1”(蓝色部分)表示该片段来自真实语音, 而标签“0”(红色部分)则表示该片段来自伪造语音。通过上层和中层的对应关系, 可以看出多头自注意力层的第 10 个头更关注语音片段中的伪造片段(与本文先前的分析一致), 这为模型推断该语音是伪造的提供了主要的判决依据。需要注意的是, 图中所展示的关系并没有完全精确对应, 其主要原因是在提取特征时进行了 STFT 变换或卷积操作, 这两个操作都会导致时间帧之间有重叠(overlap), 而片段标签则是完全没有重叠。

综上所述, 本文所提出的方法可以利用自注意力机制对部分伪造语音的检测结果进行具体的判决分析, 可视化模型在对音频进行判决时关注的时间位置, 从而实现对伪造痕迹和伪造片段所在区域的定位, 具有良好的可解释性。

5 总结

本文提出了基于视觉 Transformer 的鲁棒可解释伪造语音检测方案, 采用基于自监督学习的前端特征提取器, 并通过对提取到的语音谱图的时间和频率两个维度分别进行位置编码, 输入到基于视觉 Transformer 的后端分类网络中提取高级表示, 在不依赖数据增强方法的情况下, 显著地提升了伪造语音检测算法的性能。同时也可以与各种针对性的数据增强方法结合, 进一步提升某些场景下的检测性能。实验结果表明, 本文方法可以在域内(仅存在部分未知伪造算法)场景、语音编码算法失配场景、组合失配场景以及部分伪造等多种失配场景下均能达到良好的性能, 具有良好的鲁棒性和泛化性。同时, 本文还具体介绍了如何利用注意力机制对部分伪造语音的检测结果进行具体的判决分析, 证明了本文方法具有良好的可解释性。

参考文献

[1] Wang Y X, Skerry-Ryan R J, Stanton D, et al. Tacotron: Towards End-to-End Speech Synthesis[C]. *Interspeech 2017*, 2017: 4006-4010.

[2] Skerry-Ryan R J, Battenberg E, Xiao Y, et al. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron[C]. *International Conference on Machine Learning*, 2018.

[3] Ren Y, Ruan Y, Tan X, et al. Fastspeech: Fast, Robust and Controllable Text to Speech[C]. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019: 3171-3180.

[4] Ren Y, Hu C X, Tan X, et al. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech[EB/OL]. 2020: arXiv: 2006.04558. https://arxiv.org/abs/2006.04558.

[5] Kim J, Kong J, Son J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech[EB/OL]. 2021: arXiv: 2106.06103. https://arxiv.org/abs/2106.06103.

[6] Yi J, Wang C, Tao J, et al. Audio Deepfake Detection: A Survey[EB/OL]. 2023: ArXiv Preprint ArXiv:2308.14970.

[7] Ren Y Z, Liu C Y, Liu W Y, et al. A Survey on Speech Forgery and Detection[J]. *Journal of Signal Processing*, 2021, 37(12): 2412-2439.
(任延珍, 刘晨雨, 刘武洋, 等. 语音伪造及检测技术研究综述[J]. *信号处理*, 2021, 37(12): 2412-2439.)

[8] Yamagishi J, Wang X, Todisco M, et al. ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 47-54.

[9] Tomilov A, Svishchev A, Volkova M, et al. STC Antispoofing Systems for the ASVspoof2021 Challenge[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 61-67.

[10] Li X, Li N, Weng C, et al. Replay and Synthetic Speech Detection with Res2Net Architecture[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2021: 6354-6358.

[11] Wang X, Yamagishi J. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection[C]. *Interspeech 2021*, 2021: 4259-4263.

[12] Chintha A, Thai B, Sohrawardi S J, et al. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2020, 14(5): 1024-1037.

[13] Tak H, Patino J, Todisco M, et al. End-to-End Anti-Spoofing with RawNet2[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2021: 6369-6373.

[14] 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[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 1-8.

[15] Jung J W, Heo H S, Tak H, et al. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2022: 6367-6371.

[16] 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.

[17] Zeghidour N, Teboul O, de Chaumont Quitry F, et al. LEAF: A Learnable Frontend for Audio Classification[EB/OL]. 2021: arXiv: 2101.08596. https://arxiv.org/abs/2101.08596.

[18] Ravanelli M, Bengio Y. Speaker Recognition from Raw Waveform with SincNet[C]. *2018 IEEE Spoken Language Technology Workshop*, 2019: 1021-1028.

[19] Fu Q C, Teng Z W, White J, et al. FastAudio: A Learnable Audio Front-End for Spoof Speech Detection[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2022: 3693-3697.

[20] Wu X, He R, Sun Z N, et al. A Light CNN for Deep Face Repre-

sentation with Noisy Labels[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(11): 2884-2896.

[21] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[22] Desplanques B, Thienpondt J, Demuynck K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification[C]. *Interspeech 2020*, 2020: 3830-3834.

[23] Tak H, Jung J W, Patino J, et al. Graph Attention Networks for Anti-Spoofing[C]. *Interspeech 2021*, 2021: 2356-2360.

[24] Tak H, Todisco M, Wang X, et al. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation[C]. *The Speaker and Language Recognition Workshop*, 2022: 112-119.

[25] Baevski A, Zhou H, Mohamed A, et al. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations[EB/OL]. 2020: arXiv: 2006.11477. https://arxiv.org/abs/2006.11477.

[26] Tak H, Kamble M, Patino J, et al. Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2022: 6382-6386.

[27] Babu A R, Wang C H, Tjandra A, et al. XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale[C]. *Interspeech 2022*, 2022: 2278-2282.

[28] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. 2020: arXiv: 2010.11929. https://arxiv.org/abs/2010.11929.

[29] Gong Y, Chung Y A, Glass J. AST: Audio Spectrogram Transformer[C]. *Interspeech 2021*, 2021: 571-575.

[30] Koutini K, Schlüter J, Eghbal-zadeh H, et al. Efficient Training of Audio Transformers with Patchout[C]. *Interspeech 2022*, 2022: 2753-2757.

[31] Müller N, Czempin P, Diekmann F, et al. Does Audio Deepfake Detection Generalize? [C]. *Interspeech 2022*, 2022: 2783-2787.

[32] Zhang L, Wang X, Cooper E, et al. The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance[J]. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, 2023, 31: 813-825.

[33] Yang Y J, Qin H C, Zhou H, et al. A Robust Audio Deepfake Detection System *via* Multi-View Feature[C]. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2024: 13131-13135.

[34] Guo Y L, Huang H F, Chen X, et al. Audio Deepfake Detection with Self-Supervised Wavlm and Multi-Fusion Attentive Classifier[C]. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2024: 12702-12706.

张桐 于 2023 年在华中科技大学信息安全专业获得学士学位。现在武汉大学网络空间安全专业攻读硕士学位。研究领域为多媒体内容安全。研究兴趣包括: 伪造语音检测。Email: tongzhang@whu.edu.cn

邓俊龙 于 2024 年在武汉大学网络空间安全专业获得工学硕士学位。研究领域为多媒体内容安全。研究兴趣包括: 伪造语音检测。Email: jldeng@whu.edu.cn

任延珍 于 2009 年在武汉大学通信与信息系统专业获得博士学位。现任武汉大学国家网络安全学院教授。研究领域为多媒体内容安全。研究兴趣包括: AI 交互安全、多媒体取证、多媒体伪造检测, 多媒体信息隐藏及隐写分析、多媒体特征表示学习等。Email: renyz@whu.edu.cn

王丽娜 1964 年 10 月生, 女, 博士、二级教授, 博士生导师。教育部重点实验室主任, 国务院政府特殊津贴获得者。全国信息隐藏专家委员会成员, CCF 高级会员, CCF 信息安全与保密专委会, 中国密码学会会员。主要研究方向是多媒体信息隐藏、隐写分析理论和技术、网络安全、云安全及人工智能安全。Email: lnwang@whu.edu.cn