基于 WavLM 特征解相关的深度伪造语音检测方法

王帅斌[1,2], 易小伟[1,2], 刘长军[1,2], 苏小苏[1,2], 曹 纭[1,2], 吕美杨[1,2]

[1]中国科学院信息工程研究所 北京 中国 100085
[2]中国科学院大学网络空间安全学院 北京 中国 100049

摘要 随着语音合成与转换技术的成熟及应用, 高质量的伪造语音足以欺骗人类听觉感知和说话人验证系统, 深度伪造语音技术的恶意利用对个人财产安全和社会稳定产生了严重的威胁。近年来, 深度伪造语音检测研究受到了广泛关注, 并且在特定数据集上获得了很好的检测效果。然而, 已有检测方法在跨域的通用伪造特征提取方面存在局限性, 以及语音特征之间存在统计相关性会误导模型学习到与语音检测任务无关的特征, 导致模型在跨域场景下的性能严重下降。本文提出了一种基于 WavLM 特征解相关的深度伪造语音检测方法, 该方法首先提出了一个基于自监督预训练 WavLM 模型和图注意力网络结合的 WavLMAST 模型, 利用 WavLM 模型提取语音的声学层、内容层和语义层特征, 再结合基于图注意力的后端网络进一步建模语音的自适应时频域特征, 这种设计增强了模型对深度伪造语音中微妙伪影的表示能力。然后, 通过动态调整训练样本的特征相关度权重对 WavLMAST 模型提取的多层特征解相关, 使模型更关注与伪造语音检测任务相关的特征, 从而提高其在跨域检测场景下的泛化能力。实验结果表明, 本文方法在 ASVspoof 2019 logical access (LA)和 ASVspoof 2021 LA 数据集上比最先进的 Mixture of Experts 方法的等错误率分别降低了 40.5%和 36.8%。

关键词 深度伪造语音; 语音合成; 伪造语音检测; 泛化性; ASVspoof 数据集
中图法分类号 TP37 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2025.11.15

A Deepfake Speech Detection Method Based on WavLM Feature Decorrelation

WANG Shuaibin[1,2], YI Xiaowei[1,2], LIU Changjun[1,2], SU Xiaosu[1,2], CAO Yun[1,2], LYU Meiyang[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract With the maturation and increasing application of speech synthesis and voice conversion technologies, high-quality deepfake speech has become capable of deceiving both human auditory perception and automated speaker verification systems. The malicious use of deepfake speech technology therefore poses a substantial threat to personal financial security as well as social stability. In recent years, deepfake speech detection has attracted significant research attention and has achieved strong performance on certain benchmark datasets. However, existing detection approaches still exhibit limitations in extracting universal and domain-invariant forgery features. Furthermore, statistical correlations among multi-level speech features may mislead models into learning patterns that are irrelevant to the detection task, resulting in severe performance degradation when deployed in cross-domain scenarios. To address these challenges, this paper proposes a deepfake speech detection method based on WavLM feature decorrelation. First, we introduce a hybrid architecture termed WavLMAST, which integrates a self-supervised pre-trained WavLM model with a graph attention network. The WavLM component is utilized to extract acoustic-level, content-level, and semantic-level speech representations. These representations are then fed into a graph-attention-based backend, which adaptively models the time–frequency domain characteristics of speech. This design enhances the model's capability to represent subtle artifacts associated with deepfake speech. Subsequently, a dynamic feature decorrelation strategy is applied to the multi-layer representations extracted by WavLMAST. By adjusting the feature-correlation weights of training samples, the proposed method reduces statistical dependencies among features and encourages the model to focus on attributes that are directly relevant to the deepfake speech detection task. This mechanism significantly improves the model's generalization ability in cross-domain detection scenarios. Experimental results demonstrate the effectiveness of the proposed approach. On the ASVspoof 2019 Logical Access (LA) and ASVspoof 2021 LA datasets, the proposed method reduces the Equal Error Rate by 40.5% and 36.8%, respectively, compared with the state-of-the-art Mixture of Experts approach.

Key words deepfake speech; speech synthesis; deepfake speech detection; generalization; ASVspoof dataset

通信作者: 刘长军, 硕士, 高级工程师, Email: liuchangjun@iie.ac.cn。
本课题得到国家自然科学基金项目(No. 62272456)资助。
收稿日期: 2025-06-30; 修改日期: 2025-11-15; 定稿日期: 2025-11-17

1 引言

随着语音合成(Text To Speech, TTS)与语音转换(Voice Conversion, VC)技术的快速发展, 深度伪造语音的质量已经几乎达到人类真实语音的水准。目前, 此类技术已广泛应用于智能助理、虚拟主播、媒体配音等领域, 为音视频的制作带来便利。然而, 技术的迅速进步也带来了安全隐患, 不法分子开始利用其实施诈骗、诬陷他人、冒用身份等违法行为。近期有研究表明[1], 人类在区分真实语音与部分伪造语音方面准确率极低, 该研究对 148 名参与者进行了实验, 结果显示, 受试者对于陌生声音的伪造语音识别准确率仅为 16%, 对熟悉声音的伪造语音识别率也仅为 17.5%。另有研究显示[2], 2023 年基于人工智能的深度伪造欺诈行为暴增 30 倍。由此可见, 语音生成技术在被广泛应用的同时, 也对个人的隐私和财产安全甚至公共安全造成了严重威胁。

鉴于语音深度伪造攻击带来的持续威胁, 学术界与工业界已提出多种检测技术路线作为应对措施。早期的检测方法大多依赖于手工设计的传统声学特征作为输入, 如梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient, MFCC)[3]、线性频率倒谱系数(Linear Frequency Cepstral Coefficients, LFCC)[4]、梅尔频谱[5]、短时傅里叶变换频谱[6]等, 搭配卷积神经网络、循环神经网络等结构进行分类。这些特征虽然在一定程度上能够捕捉虚假语音中的伪造痕迹, 但存在多方面的局限性。首先, 手工特征对特定攻击模式具有一定偏向性, 缺乏对攻击多样性和复杂性的适应能力, 容易在面对未知攻击、跨设备或跨环境失真时失效。其次, 传统特征通常以帧级或低维的形式表示语音信号, 无法充分表达时序上下文和长距离依赖关系, 限制了模型对伪造语音中潜在细粒度伪迹的建模能力。此外, 这些特征设计之初针对自动语音识别或说话人识别等任务, 并未考虑伪造语音的结构特性, 因此缺乏与伪造方式相关的判别能力。即使一些工作尝试了多特征、多模型融合[7]等提高模型表现的策略, 但依然没有解决核心的特征表示不足的问题。因此, 如何设计和提取更具表达力、更稳定、更泛化的表示特征, 成为提升深度伪造语音检测(Deepfake Speech Detection, DSD)任务性能的关键问题。

在此背景下, 随着自监督学习在自然语言处理、语音识别、自动说话人验证(Automatic Speaker Verification, ASV)等领域取得的重大进步, 为这一问题提供了新的研究思路, 出现了一些如 Wav2Vec2.0[8]、WavLM[9]和 HuBERT[10]等的预训练语音模型, 这类模型通过在大规模无标签语音数据上进行掩码预测或对比学习, 能够自动提取声音的多层次语义、时序和声学信息, 从而获得比传统手工特征更丰富、更细微的深层特征表示。已经有研究把这些模型用作 DSD 任务的特征提取器。Xie 等人[11]提出了一种利用 Wav2Vec2.0 模型提取的特征来训练基于孪生网络的伪造语音检测系统。Wang 等人[12]通过注意力机制将 Wav2Vec 模型[13]提取到的特征、韵律特征和发音特征融合在一起用于伪造语音检测。

尽管自监督预训练模型的引入显著增强了系统的特征挖掘和伪造检测能力, 但仍不足以应对现实世界中复杂语音环境的干扰。真实应用环境中的语音容易受到多种噪声和失真的影响, 而预训练模型往往在干净语音上训练, 难以覆盖多变的噪声模式, 导致在识别带噪语音时准确率下降。为缓解上述问题, 数据增强成为改善模型鲁棒性的常用策略。通过针对性引入背景噪声、混响、压缩编码等多种扰动, 可以使训练数据更贴近真实使用场景, 从而显著减少模型对干净环境的依赖, 提升其处理未知噪声与信道失真的能力。例如, Tak 等人[14]提出了一种直接对原始语音波形进行数据增强的 RawBoost 方法, 在 ASVspoof 2021 LA 数据集[15]上比基线系统的性能相对提高了 27%。

然而, 除了编码失真、录制设备和噪声环境的干扰, DSD 任务还面临未知伪造方法下的分布偏移挑战, 数据增强可以在一定程度上扩大训练数据的分布范围, 但对于在帮助模型学习到更通用的特征上贡献有限。在深度模型训练中, 不同特征之间存在统计相关性, 尤其是与类别无关的伪相关特征与类别相关特征之间的相关性。这种相关性容易导致模型学习到错误的判别边界, 使其在训练分布内表现良好, 而在分布外严重退化。例如, 若训练集中某一个说话人的真实语音数量较多, 其说话人特征被模型学习为真伪语音的判别边界, 则该说话人特征即为伪相关特征。针对机器学习中的分布偏移问题, Zhang 等人[16]提出了深度稳定学习框架, 通过样本权重学习(Sample Weight Learning, SWL)的训练机制, 去除特征间的虚假相关, 从而使模型聚焦于标签与输入样本间的因果稳定特征而非偶然的统计关联, 进而提高模型的泛化性。在 DSD 任务中, 该模块可以在一定程度上解耦语音的语义内容、声学信息、背景噪声等特征, 通过加权抑制导致伪相关特征的样本的影响, 使模型学习到对伪造语音更具判别力的特征。Wang 等人[17]的工作已经展现了 SWL 模块

在 DSD 任务中的有效性。

现有的 DSD 方法存在手工提取的声学特征计算模式固定, 往往会对某些种类的特征表示具有一定的偏向性, 在面对多样化的攻击模式时存在局限性; 基于卷积神经网络的检测模型由于其池化操作会导致时间信息的丢失[18]; 语音的多种特征之间存在相关性的问题, 与 DSD 任务伪相关的特征可能会对模型学习造成过多干扰。因此, 本文针对 DSD 任务提出了更具泛化性和鲁棒性的通用特征提取模型与伪造识别特征解相关方法, 本文的主要贡献如下。

(1) 针对当前伪造检测特征存在跨域跨数据分布表征能力差、检测特征与伪造方法紧耦合等问题, 本文提出了 WavLMAST 模型。首先, 该模型引入通过自监督学习、掩码预测和去噪机制联合训练的 WavLM 模型, 提取说话人一致性、语音韵律边界和语音扰动等丰富的通用特征表示, 获得语音的声学层、内容层和语义层特征。其次, 针对卷积神经网络模型存在丢失时间信息的问题, 引入图注意力网络分别提取语音的时域和频域特征, 并联合建模跨域的时频交叉关系。WavLMAST 模型通过融合通用预训练语音特征与时频域交叉特征, 提高了对伪造痕迹通用特征的提取与增强。

(2) 针对语音特征之间的统计相关性导致模型学习到与 DSD 任务伪相关特征的问题, 本文引入特征解相关模块, 通过调整训练数据的重要性权重, 减小导致模型学习到虚假相关特征的样本的影响, 使模型更关注那些能够提供更“纯净”相关特征的样本, 学习到与 DSD 任务真正因果相关的本质特征, 提高模型伪造特征空间的不变性和表征一致性。

(3) 针对现有方法在检测带噪环境下语音表现下降的问题, 本文提出了多种不同模式的噪声和失真的影响, 对训练数据加入对抗性噪声, 有效缓解了模型对语音质量的依赖问题。实验结果表明本文方法同时在 ASVspoof 系列比赛的 3 个常用数据集上均达到了很好的检测结果, 尤其在检测未知环境、未见伪造方法时具有较好的鲁棒性与泛化性。

2 相关工作

现有方法可分为 3 类: 第一类是基于传统声学特征的方法, 通常以 MFCC、LFCC 等频谱变换特征作为输入, 并结合经典分类器或浅层神经网络进行建模; 第二类方法依赖于预训练语音模型提取的深层次语音表示, 如 Wav2Vec、HuBERT、WavLM 等, 通过迁移学习提升特征的判别能力; 第三类方法则从数据增强方法、模型结构设计和训练机制等方面进行了探索。

2.1 基于传统声学特征的方法

文献[3]总结了 ASVspoof 2015 比赛中 16 个主要提交的比赛方案, 其中 9 个方案都用到了 MFCC 作为输入特征, 大部分方法使用高斯混合模型[19]、支持向量机[20]或它们的融合变种作为分类器, 也有方法使用多层感知机(Multilayer Perceptron, MLP)[21]和深度神经网络[22]进行分类, 超过半数的方法使用基于特征或者后端分数的融合策略。Kang 等人[4]使用 LFCC 作为特征, SE-ResNet-18 网络[23]作为特征提取器, 探索了不同激活函数对模型整体性能的影响。Yadav 等人[5]将语音转换为梅尔频谱作为特征, 使用预训练的 Transformer 模型[24]学习特征并用于分类。Tomilov 等人[6]提出了一种融合包含倒谱系数和基于短时傅里叶变换在内的多种频谱特征, 和 ResNet18[25]、RawNet2[26]、轻量级卷积神经网络(Light Convolutional Neural Network, LCNN)[27]等多种深度神经网络模型, 以及多种数据增强方法的混合架构。Cáceres 等人[7]提出了一种轻量级的延时神经网络[28], 融合 MFCC、对数滤波器组能量、恒 Q 倒谱系数[29]等特征, 以及高斯线性分类器和数据增强方法的检测系统。由于传统声学特征在表现伪造特征上的局限性, 基于这类特征的方法大多只能采取融合多种特征、多个模型的方法来提高系统的整体表现。

2.2 基于预训练语音模型特征的方法

Xie 等人[11]提出了一种基于孪生神经网络架构的检测系统, 其使用 Wav2Vec 模型作为特征提取器, 之后将特征输入 LCNN 或 ResNet[25]学习特征, 最后使用 MLP 进行分类。Wang 等人[12]提出了一种多特征多模型融合的方法, 其使用预训练的 HuBERT 模型提取音素时长特征, 使用基于 Conformer[30]结构的模型提取发音特征, 使用预训练的 Wav2Vec2.0 模型提取自监督语音表示特征, 然后使用 Transformer 模型融合这些特征, 最后送入由长短期记忆递归神经网络(Long Short Term Memory, LSTM)[31]、LCNN、全局平均池化[32]层和全连接层[21]组成的后端结构进行分类。Yang 等人[33]评估了包括梅尔频谱、MFCC、LFCC、Wav2Vec2 XLS-R[34]、HuBERT、WavLM 在内的 14 种手工或自监督学习特征的泛化能力, 并使用 Transformer 模型融合这些特征和 ResNet18 模型作为分类器。

2.3 基于数据增强和模型结构优化的方法

Tak 等人[14]提出了一种无须外部数据, 直接对原始波形音频进行数据增强的 RawBoost 方法, 该方法基于线性与非线性卷积噪声、脉冲信号相关加

性噪声和平稳信号无关加性噪声的组合, 对编码、传输、麦克风和放大器等引入的线性和非线性失真进行了建模。Jung 等人[35]提出了一种基于图注意力网络的 AASIST 模型, 它基于 RawGAT-ST[36]网络, 提出了新的异构堆叠图注意力层, 通过异构注意力机制和堆栈节点对时域和频域特征进行建模, 并增加了最大图操作和新的读出机制来选择融合后的特征和分类。Wang 等人[17]提出了一种基于样本加权的训练方案, 该方案将深度稳定学习[16]中的样本权重学习引入 DSD 任务, 把模型提取到的特征映射到高维空间[43]量化独立性, 并通过优化样本权重最小化特征之间的相关性, 使模型在不同分布的数据上表现稳定。

3 本文方法

3.1 总体框架

本文提出的方法由数据增强模块、WavLMAST 模型、特征解相关模块构成, 其总体框架如图 1 所示。在训练阶段, 首先语音数据被裁剪或重复处理为 4 秒的固定长度, 并进入数据增强模块引入噪声, 之后送入 WavLMAST 模型获取隐藏层特征和预测分数。其中预测分数与该训练批次样本的标签通过加权交叉熵获得损失, 隐藏层特征传入特征解相关模块获取当前批次训练样本的权重, 最后将权重与当前损失值逐元素相乘, 获得最终的加权损失。在推理阶段, 语音数据不经过数据增强模块, 直接输入 WavLMAST 模型并获得预测分数, 然后使用 Softmax 函数计算预测分数的概率分布, 并取概率最高结果的下标作为预测标签。

3.2 数据增强模块

通过前期的文献调研[14, 37-42]和实验尝试, 本文对训练语音加入房间脉冲响应、各向同性和点声源噪声[38]来进行数据增强。

房间脉冲响应描述了一个封闭空间中声音从声源传播到接收麦克风过程中的全部响应, 包括直达声、早期反射和混响尾音。房间脉冲响应可以模拟语音在真实房间中的传播效果, 从而增加模型对空间混响和远场环境的鲁棒性。

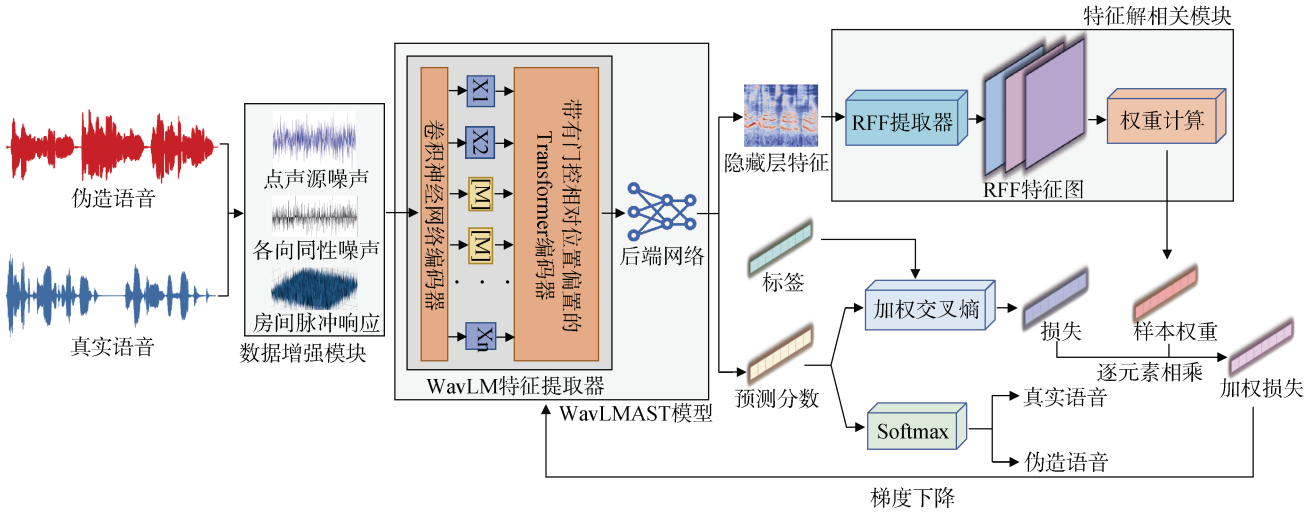

图 1 本文方法的总体框架图

Figure 1 Overall framework of the proposed method

各向同性噪声是一种在空间中均匀分布、来自多个方向的背景噪声, 例如人群嘈杂声、雨声或风声等。这种噪声不集中于某一个声源点, 而是模拟在真实环境中被多个物体、墙面反射后的随机背景噪声。引入各向同性噪声可以模拟拥挤公共场所或多声源干扰环境下的语音, 增强模型在复杂背景下的抗干扰能力。

点声源噪声是指来自空间中特定位置的单一声源, 例如某人说话、机器运转、铃声等。与各向同性噪声相比, 点声源噪声具有更明确的方向性和空间位置。点声源噪声可以模拟现实中语音信号受到局部干扰的情况, 有助于训练模型区分目标语音与局部噪声干扰, 提高其在实际应用中的鲁棒性。

3.3 WavLMAST 模型结构

3.3.1 WavLM 特征提取器

经过前期对 Wav2Vec2.0、HuBERT 和 WavLM 三个预训练语音大模型的实验探索和对比, 本文发现 WavLM 模型在 DSD 任务上表现更好, 因此, 为了提取更具判别力的语音特征, 本研究利用 WavLM 作为前端特征提取器。WavLM 是一种基于自监督学

习的语音预训练模型, 其设计在结构上延续了 Wav2Vec2.0 模型的架构, 主要由一个卷积特征编码器和一个 Transformer 编码器构成。

在 WavLM 模型中, 输入为原始语音波形序列 a。首先, 该输入通过前端卷积特征提取器 f_{conv} 进行处理, 得到初始的低阶声学特征可表示为

$$X=\{x_1, x_2, \cdots, x_N\}=f_{\text{conv}}(a),\quad x_i\in\mathbb{R}^d \tag{1}$$

其中, $X\in\mathbb{R}^{N\times d}$ 表示经由多层卷积网络提取出的局部时间窗口内的表示, N 为帧数, d 为特征维度。若 $d\neq D$ (其中 D 为 Transformer 编码器的输入维度), 则通过线性投影层 $W_{\text{proj}}\in\mathbb{R}^{d\times D}$ 将其映射到统一维度, 得到

$$H^{(0)}=X\cdot W_{\text{proj}}+b\,. \tag{2}$$

然后, 特征序列 $H^{(0)}$ 被输入至包含 L 层的 Transformer 编码器中, 每一层包括带有门控位置偏置的多头自注意力模块和前馈神经网络。在第 l 层中, 其输出记作 $H^{(l)}$, 并通过残差连接与层归一化完成信息更新, 即

$$H^{(l)}=\text{TransformerLayer}^{(l)}\left(H^{(l-1)}\right),\quad l=1,2,\cdots,L \tag{3}$$

最后, Transformer 编码器的输出 $H^{(L)}\in\mathbb{R}^{N\times D}$ 即为包含全局上下文的语音表征, 然后将其送入后端网络进一步提取和融合语音的时频域特征。

3.3.2 基于 AASIST 模型的后端网络

WavLMAST 模型的后端网络基于 AASIST 模型实现。本文去除了原 AASIST 模型中基于 RawNet2 模型的编码器, 将 WavLM 模型提取到的特征输入到后续的图注意力网络中。后端网络由图结合模块、异构堆栈图注意力层、最大图操作与读出模块组成。其网络结构如图 2 所示。

图结合模块: 首先分别构建时间图 $G_t\in\mathbb{R}^{N_t\times D_t}$ 与频谱图 $G_s\in\mathbb{R}^{N_s\times D_s}$, 其中 N_t 与 N_s 分别表示结点数量, D_t 与 D_s 分别表示对应的结点特征维度。结点特征来自编码器输出特征 $F\in\mathbb{R}^{C\times S\times T}$, 其中 C 为通道数量, 其通过时间与频率维度最大池化获得初始输入, 即

$$\begin{aligned}G_t&=\text{Graph\_module}\left(\max_s|F|\right)\\ G_s&=\text{Graph\_module}\left(\max_t|F|\right)\end{aligned} \tag{4}$$

其中, Graph_module 为由图注意力层与图池化层组合而成的子模块。图注意力层通过结点间自适应注意力计算边权重, 图池化操作则依据结点注意力得分选取前 k 个重要结点以降低图规模, 增强结构表达能力。

然后, 为进一步融合时间与频谱域信息, 将 G_t 与 G_s 两个子图中的所有结点两两相连构建为一个异构图 G_{st}。该联合图包含 N_t+N_s 个结点, 通过异构边建立跨域关联。虽然结点被投射至统一空间, 但不同来源的结点仍处于不同语义子空间, 使得 G_{st} 仍具备异构特性。

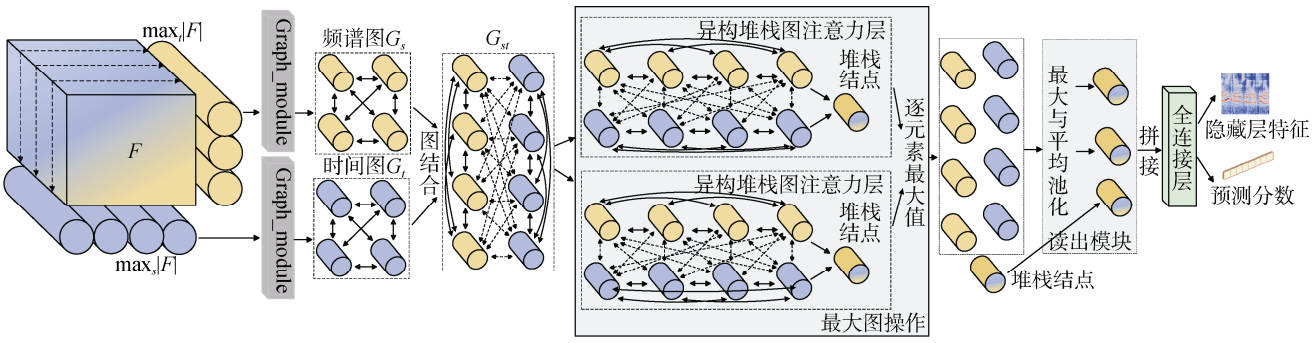

图 2 后端网络结构图

Figure 2 Backend network structure diagram

异构堆栈图注意力层: 用于建模异构图中的时频交叉关系, 该模块首先通过全连接变换将两个子图结点投射至公共特征空间

$$\tilde{G}_t=G_tW_t+b_t,\ \tilde{G}_s=G_sW_s+b_s \tag{5}$$

其中, $W_t\in\mathbb{R}^{D_t\times D_{st}}$、$W_s\in\mathbb{R}^{D_s\times D_{st}}$ 为学习到的映射矩阵, D_{st} 为共享维度。为计算不同边类型上的注意力权重, 该模块引入三组独立的投影向量 p_t, p_s, p_{ts}, 分别用于时间内结点、频率内结点以及时频交叉结点之间边的注意力权重计算, 其结构如图 2 中的 G_{st} 所示。以两个结点 h_i,h_j 为例, 其边权重定义为

$$a_{ij}=\frac{\exp\left(\sigma\left(p_{e_{ij}}^{\mathrm{T}}(h_i\odot h_j)\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\sigma\left(p_{e_{ik}}^{\mathrm{T}}(h_i\odot h_k)\right)\right)} \tag{6}$$

其中, $\odot$表示逐元素乘法, σ 为激活函数, $\mathcal{N}(i)$ 为第 i 个节点的邻居集合, $p_{e_{ij}}$ 表示边类型对应的投影向量, e_{ij} 表示图中从结点 i 指向结点 j 的边的类型, 包括时间图内边($i,j\in G_t$)、频谱图内边($i,j\in G_s$)和时间频率跨域边($i\in G_t, j\in G_s$ 或反之)三种类型。

该模块还引入一个堆栈结点, 作用是积累异质信息, 即谱域和时域之间的信息或关系。堆栈结点连接到所有结点的集合。使用从所有其他结点到堆栈结点的单向边有助于保留 G_t 和 G_s 中的信息。此外, 当顺序使用多个异构堆栈图注意力层时, 可以将前一层的堆栈节点传递到下一层。

最大图操作与读出模块: 包含两个结构对称的子分支, 每条分支由两层异构堆栈图注意力层与图池化层组成, 其输出通过逐元素最大值操作进行合并, 同时, 对两个分支中的堆栈结点也执行最大操作, 保留更具判别能力的全局表示。

最后, 通过对时间图与频谱图的结点分别施加最大与平均池化, 并将四个池化结点与堆栈结点拼接, 形成最终输出表示

$$z=\mathrm{Concat}\left(\max(G_t),\ \mathrm{avg}(G_t),\ \max(G_s),\ \mathrm{avg}(G_s),\ \mathrm{stack}\right) \tag{7}$$

z 即为隐藏层特征, 被送入特征解相关模块学习当前批次训练数据的权重, 同时被送入全连接层获得预测分数。

3.4 特征解相关模块

特征解相关模块通过 SWL 机制全局加权训练数据来解相关每个输入样本的特征, 从而降低 DSD 任务中无关特征对模型的影响。3.4.1 节介绍使用随机傅里叶特征(Random Fourier Features, RFF)度量训练样本的特征间独立性的理论依据与公式, 3.4.2 节介绍降低计算和存储开销的迭代更新样本权重机制。

3.4.1 基于 RFF 的特征独立性度量

设神经网络学习到的特征为 $Z\in\mathbb{R}^{n\times m}$, 其中 n 为样本数量, m 为特征空间的维度, 特征空间中的第 i 个变量表示为 $Z_{:,i}$。为了消除特征空间中任意两个特征 $Z_{:,i}$ 和 $Z_{:,j}$ 的依赖关系, 首先需要度量特征之间的独立性。

从两个随机特征的分布 A 和 B 中采样出样本 $(A_1, A_2,\cdots, A_n)$ 和 $(B_1, B_2,\cdots, B_n)$, 设定义在 A 的域上的一个可测量正定核函数为 k_A, 其对应的再生核希尔伯特空间(Reproducing Kernel Hilbert Space, RKHS)为 $\mathcal{H}_A$, 类似地定义 k_B 和 $\mathcal{H}_B$, 定义 $\mathcal{H}_B$ 与 $\mathcal{H}_A$ 的互协方差矩阵为 Σ_{AB}。

由希尔伯特-施密特独立性准则(Hilbert-Schmidt Independence Criterion, HSIC)可知 Σ_{AB} 的 Hilbert-Schmidt 范数可被用作监督特征解相关[44], 但其计算开销在大规模数据集上较大, 由于在欧几里得空间中 Frobenius 范数与 Hilbert-Schmidt 范数[45]等价, 因此本研究使用 Frobenius 范数来度量特征之间的独立性。设部分互协方差矩阵为

$$\hat{\Sigma}_{AB}=\frac{1}{n-1}\sum_{i=1}^{n}\left(u(A_i)-\frac{1}{n}\sum_{j=1}^{n}u(A_j)\right)^{\mathrm{T}}\cdot\left(v(B_i)-\frac{1}{n}\sum_{j=1}^{n}v(B_j)\right) \tag{8}$$

其中,

$$\begin{Bmatrix}u(A)=(u_1(A),\ u_2(A),\ \ldots,\ u_{nA}(A)),\ u_j(A)\in\mathcal{H}_{\mathrm{RFF}}\\ u(B)=(u_1(B),\ u_2(B),\ \ldots,\ u_{nB}(B)),\ v_j(B)\in\mathcal{H}_{\mathrm{RFF}}\end{Bmatrix} \tag{9}$$

$\mathcal{H}_{\mathrm{RFF}}$ 表示 RFF 所构成的函数空间, 其形式为

$$\mathcal{H}_{\mathrm{RFF}}=\left\{\sqrt{2}\cos(\omega x+\phi)\,|\,\omega\sim\mathcal{N}(0,1),\ \phi\sim\mathrm{Uniform}(0,2\pi)\right\} \tag{10}$$

其中, ω 和 ϕ 分别从标准正态分布和均匀分布中采样。

将变量 A 和 B 的独立性检验统计量 I_{AB} 定义为部分互协方差矩阵的 Frobenius 范数,

$$I_{AB}=\left\|\hat{\Sigma}_{AB}\right\|_{\mathrm{F}}^{2} \tag{11}$$

当 $I_{AB}\to 0$ 时, 变量 A 和 B 趋于独立。由于 I_{AB} 是非负的, 所以 I_{AB} 越小, 变量 A 和 B 越独立。

在建立了独立性度量标准以后, 受文献[46]启发, 可以通过对样本加权来消除隐藏状态特征之间的依赖性, 并利用 RFF 度量总体独立性。令样本权重为 $\omega\in\mathbb{R}_{+}^{n}$, 且满足 $\sum_{i=1}^{n}\omega_i=n$。加权后的互协方差矩阵更新为

$$\hat{\Sigma}_{AB;w}=\frac{1}{n-1}\sum_{i=1}^{n}\left(w_iu(A_i)-\frac{1}{n}\sum_{j=1}^{n}w_ju(A_j)\right)^{\mathrm{T}}\cdot\left(w_iv(B_i)-\frac{1}{n}\sum_{j=1}^{n}w_jv(B_j)\right) \tag{12}$$

具体而言, 对于上述提到的特征对 $Z_{:,i}$ 和 $Z_{:,j}$, 通过优化权重 ω 来实现去相关, 目标函数为

$$\omega^{*}=\arg\min_{\omega\in\Delta_n}\sum_{1\le i<j\le m}\left\|\hat{\Sigma}_{Z_{:,i},\,Z_{:,j};\omega}\right\|_{\mathrm{F}}^{2} \tag{13}$$

其中，$\Delta_n=\left\{\omega\in\mathbb{R}_{+}^{n}\mid\sum_{i=1}^{n}\omega_i=n\right\}$，通过使用最优样本权重 ω^{*} 对训练样本加权，可以有效减少所有隐藏特征之间的依赖性。

该算法通过以下步骤迭代优化样本权重 ω、特征提取器 f 和分类器 g，即

$$f^{(t+1)},\,g^{(t+1)}=\arg\min_{f,g}\sum_{i=1}^{n}\omega_i^{(t)}\cdot\mathcal{L}(g(f(X_i)),\,y_i),$$

$$\omega^{(t+1)}=\arg\min_{\omega\in\Delta_n}\sum_{1\le i<j\le m}\left\|\hat{\Sigma}_{Z_{:,i},\,Z_{:,j};\omega}\right\|_{F}^{2} \tag{14}$$

其中，$Z^{(t+1)}=f^{(t+1)}(X)$，$\mathcal{L}(,)$ 表示加权交叉熵损失函数，t 为迭代时间步，初始样本权重设置为 $\omega^{(0)}=(1,\,1,\cdots,\,1)^{\mathrm{T}}$。

3.4.2 迭代学习全局样本权重

式(14)要求在训练过程中为每个样本学习一个权重。然而，对于基于深度神经网络且训练数据量庞大的 DSD 模型而言，全局学习所有训练样本的权重将带来极大的计算和存储开销。

因此，将每个训练阶段中遇到的特征与权重保存并合并，以作为后续训练中学习全局样本权重的依据。在每个训练批次中，特征与样本权重的组合方式为

$$\begin{aligned}Z_O&=\mathrm{Concat}(Z_{G1},\,Z_{G2},\,\ldots,\,Z_{Gk},\,Z_L)\\ w_O&=\mathrm{Concat}(w_{G1},\,w_{G2},\,\ldots,\,w_{Gk},\,w_L)\end{aligned} \tag{15}$$

其中，Z_O 和 W_O 表示当前批次中用于优化的特征和权重，$Z_{G1},\,\ldots,\,Z_{Gk}$ 和 $w_{G1},\,\ldots,\,w_{Gk}$ 分别表示从之前所有批次中累积得到的全局特征和全局权重，Z_L 和 w_L 表示当前批次的局部特征和局部权重。这些信息会用于优化下一批次中的样本权重。然而，在训练数据量较大的场景下，这种拼接操作会导致显著的内存消耗，并增加计算复杂度。

为了解决这个问题，在每轮训练结束时，将历史的全局信息与当前的局部信息进行融合，更新后的全局表示为

$$\begin{aligned}Z'_{Gi}&=\alpha_i Z_{Gi}+(1-\alpha_i)Z_L\\ w'_{Gi}&=\alpha_i w_{Gi}+(1-\alpha_i)w_L\end{aligned} \tag{16}$$

然后在下一批次中，使用更新后的 $(Z'_{Gi},\,w'_{Gi})$ 作为新的全局信息继续训练。其中，$\alpha\in[0,1]$ 用来控制信息的保留程度，较大的 α 值代表更强的长期记忆(更依赖过去的全局信息)，较小的 α 值则代表更强的短期记忆(更侧重当前的局部特征)。

4 实验结果与分析

本小节首先介绍实验使用的数据集及评价指标，以及实验的一些超参数等设置，最后展示并分析本文方法在不同数据集上进行的评估实验、与其他工作的对比实验和消融实验的结果。

4.1 实验设置

4.1.1 数据集

本文使用 ASVspoof 2019 LA[48]数据集的训练集对模型进行训练，为了测试模型的鲁棒性和泛化性，使用 ASVspoof 2019 LA(19LA)、ASVspoof 2021 LA(21LA)[15]和 ASVspoof 2021 deepfake(21DF)[15]三种具有代表性的 DSD 任务数据集的测试集对不同模型进行了评估。本文实验所使用的数据集信息总览如表 1 所示。

表 1 实验使用的数据集

Table 1 Overview of the dataset used in the experiment

数据集	说话人数量 男性/女性	伪造方法种数	条件	采样率	真实语音数量	伪造语音数量
19LA 训练集	8/12	6 种	干净	16 kHz	2580	22800
19LA 测试集	30/37	13 种	干净	16 kHz	7355	63822
21LA 测试集	21/27	13 种	干净&编码	8 或 16 kHz	18452	163114
21DF 测试集	21/27	100+种	干净&编码	8 或 16 kHz	22617	589212

19LA 数据集基于一个在半消声房间中以 96kHz 采样率录制的多说话人英语数据库 Voice Cloning Toolkit(VCTK)[47]构建，它将语音文件以 16 比特的位深降采样到 16 kHz，并使用了 17 种完全不同类型的 TTS 和 VC 方法构建，其中训练集使用了 4 种 TTS 和 2 种 VC 方法，测试集使用了 7 种 TTS 和 6 种 VC 方法。

21LA 数据集基于 ASVspoof 2015[49]和 19LA 数据集构建，为扩大挑战，对真实和伪造语音进行了多种电话编码和传输方式，包括基于 IP 的语音传输(Voice over Internet Protoco, VoIP)和公用电话交换网(Public Switched Telephone Network, PSTN)，使用的

伪造方法与 19LA 的测试集相同(共 13 种)。

21DF 数据集基于 19LA 的测试集以及其他(未公开)来源的数据建立, 包含真实语音与超过 100 种不同伪造方法生成的伪造语音, 这些语音被多种有损编解码器处理, 包括 mp3、m4a、ogg 等格式, 不同的编码器类型和参数设置会引入不同程度的失真。

本实验使用的噪声数据集中, 房间脉冲响应共 325 条音频, 由 Real World Computing Partnership 声音场景数据库[39]、REVERB 挑战数据库[40]和亚琛脉冲响应数据库[41]组成, 各向同性噪声共 92 条音频, 同样来自以上三个数据库。点声源噪声采自 Music Speech and Noise Corpus(MUSAN)[42]中的 Freesound 部分, 共 843 条音频。

4.1.2 评价指标

本文实验使用两个指标对模型进行评估: 等错误率(Equal Error Rate, EER)和最小串联检测损失函数(Minimum Tandem Detection Cost Function, min t-DCF)[50]。这两个指标分别从系统整体和组件单独性能两个层面, 对伪造检测系统进行了全面的分析。其中, EER 是评估伪造检测系统自身性能的经典指标, 它表示伪造语音被误判为真实语音与真实语音被误判为伪造语音的概率相等时的错误率。EER 反映了伪造检测系统在不依赖 ASV 系统的前提下对伪造语音的识别能力, 是衡量系统分类边界质量的重要指标。越低的 EER 表明模型在平衡两类错误的前提下具有更好的判别能力。

min t-DCF 则是在 ASVspoof 挑战中提出的重要评估指标, 旨在评估伪造检测系统与 ASV 系统在串联使用时的综合性能。该指标基于特定的应用场景成本权重设置, 反映了在现实部署条件下, CM 与 ASV 系统协同工作的安全性与实用性。较低的 min t-DCF 值表示系统在区分真实语音与伪造语音的同时, 尽可能减少对合法用户的拒绝, 有助于提升整体用户体验与系统鲁棒性。

4.1.3 实验参数

实验使用 librosa 库[51]读取音频, 目标采样率设置为 16kHz, 并通过裁剪或重复将音频统一为 4 秒(64000 个采样点)。在数据增强阶段, 设置信噪比为 10dB, 把噪声音频添加到语音音频中。

实验使用一张 NVIDIA L40 和一张 NVIDIA A40。优化器使用 Adam[52], 学习率设置为 10^{-5}, 权重衰减系数 weight_decay 设置为 0.0001。损失函数使用交叉熵, 其中真实语音样本的权重设置为 0.8983, 伪造语音样本的权重设置为 0.1017。训练批次大小设置为 32。

特征解相关模块的实验参数按照开源实现配置[16]执行。局部权重优化器使用 AdamW, 局部学习率 lrbl=0.9, 每个 minibatch 对权重执行固定 epochb=20 次局部迭代。随机傅里叶特征数 num_f=20, sum=True(使用 cos+sin 的和作为特征映射)。权重正则幂次 decay_pow=2, 损失缩放常数 lambdap=70.0, 首轮缩放系数 first_step_cons=0.9, 历史统计与当前特征的融合采用指数平滑, 融合因子 α =presave_ratio=0.9(历史贡献 90%, 新数据贡献 10%)。该模块在每次调用后返回归一化的 softmax 权重用于样本加权。

对比实验中, 本文复现了 AASIST[35]和 AASIST+SWL[17]方法在 19LA、21LA、21DF 三个数据集上的实验结果, 由于其他方法未找到开源代码等, 直接使用作者原论文中的实验结果。

4.2 19LA 数据集实验结果

为评估本文方法对干净环境下的伪造语音的检测能力, 本文首先在 19LA 数据集上进行了测试。使用 19LA 数据集的训练集进行训练, 并在测试集上选择表现最优的模型进行测试, 实现了 0.44%的 EER 和 0.0124 的 min t-DCF, 在两项指标上均显著优于现有多种主流伪造检测方法。

表 2 展示了本文方法与当前代表性方法的详细对比。与 Jung 等人[35]提出的基于图注意力机制的 AASIST 模型相比, EER 从 0.83%降至 0.44%, 降幅约 47%, min t-DCF 亦从 0.0275 显著下降至 0.0124, 性能提升明显。此外, 与 Huang 等人[53]提出的基于频域判别特征建模的方法(EER 为 0.52%)相比, 降幅约 15%, 比 Wang 等人[54]提出的 Mixture of Experts 方法(EER 为 0.74%)降低了 40.5%。

表 2 19LA 数据集评估实验结果

Table 2 Results of 19LA dataset evaluation experiment

方法	EER(%)↓	min t-DCF↓
DM-Res2Net[55]	1.21	0.0360
AASIST+SWL[17]	1.02	0.0288
AASIST[35]	0.83	0.0275
Mixture of Experts[54]	0.74	-
SE-Rawformer[56]	0.59	0.0184
DFSincNet[53]	0.52	0.0176
本文方法	**0.44**	**0.0124**

值得注意的是, 19LA 测试集中的部分攻击类型在训练集中未出现, 因此测试结果不仅反映了模型在已知攻击条件下的学习能力, 也说明本文方法所

提出的结构设计与优化策略能够有效捕捉语音伪造中的关键特征, 具备良好的判别能力。

4.3 21LA 数据集实验结果

为进一步验证本文方法在复杂环境下的鲁棒性, 特别是在受噪声和信道失真影响的实际语音场景中的检测能力, 本文直接使用 19LA 数据集上训练得到的模型在 21LA 数据集上进行了测试。该数据集相比 19LA 数据集, 增加了更具现实性的扰动因素, 包括编码压缩、信道传输和重新采样等, 使得伪造语音与真实语音之间的差异更加微弱, 从而对检测系统提出了更高的鲁棒性要求。

实验结果如表 3 所示, 本文方法实现了 1.87%的 EER 和 0.2564 的 min t-DCF, 与 Wang 等人[17]基于稳定学习框架对 AASIST 改进后的模型(EER 为 3.66%)相比, 降低了 49%。与 Wang 等人[54]提出的 Mixture of Experts 方法(EER 为 0.74%)相比, 降低了 36.8%。

表 3 21LA 数据集评估实验结果

Table 3 Results of 21LA dataset evaluation experiment

方法	EER(%)↓	min t-DCF↓
AASIST[35]	9.25	0.4599
SE-Rawformer[56]	4.53	0.3088
AASIST+SWL[17]	3.66	0.3083
DFSincNet[53]	3.38	0.2732
Mixture of Experts[54]	2.96	-
Ensembling Model[57]	2.32	**0.2339**
本文方法	**1.87**	0.2564

Rosello 等人[57]提出的动态权重分配集成模型方法在 21LA 数据集上实现了 2.32%的 EER, 其已被 Interspeech 2024[58]收录, 代表当前该任务的领先水平之一。本文方法相比其降低了 19%, 充分展示了本文方法在跨环境条件下的鲁棒性。

在 21LA 数据集上的实验结果表明, 本文方法不仅能在应对未知伪造手段时保持良好的泛化能力, 同时还拥有在嘈杂、非理想传输条件下稳健识别深度伪造语音的能力, 具有较好的鲁棒性。

4.4 21DF 数据集实验结果

为进一步评估本文方法在面对更多复杂环境情况下语音的检测能力, 本文在 21DF 数据集上进行了测试。该数据集专为模拟真实世界中深度伪造语音的复杂性而设计, 包含超过 100 种由不同 TTS 和 VC 系统生成的伪造语音, 并经过多种有损压缩编解码器处理, 显著加大了检测任务的难度。相较于 19LA 和 21LA 数据集, 该数据集引入了更为多样化的伪造来源与通道失真, 特别适用于评估模型在开放环境下的鲁棒性与泛化性。

在该数据集上, 本文提出的检测方法依然展现出性能优势, 如表 4 所示。本文方法实现了 3.57%的 EER, 显著优于多项现有代表性方法。例如, Jung 等人[35]提出的 AASIST 模型在该数据集上的 EER 为 21.08%, 而 Wang 等人[17]提出的方法 EER 为 21.77%, 均存在明显性能差距。与在该数据集上取得良好表现的 Rosello 等人[57]、Wang 等人[59]和 Martín-Doñas 等人[60]的方法相比, 本文方法的 EER 仍展现出超过 1%的绝对优势。

表 4 21DF 数据集评估实验结果

Table 4 Results of 21DF dataset evaluation experiment

方法	EER(%)↓
AASIST+SWL[17]	21.77
AASIST[35]	21.08
Ensembling Model[57]	5.60
Wav2Vec2.0+Classifier[60]	4.98
XLS-R+LGF[59]	4.75
本文方法	**3.57**

实验结果表明, 尽管 21DF 数据集具有丰富的伪造手段和失真噪声, 本文方法依然能够稳定有效地识别出真实与伪造语音。这一性能提升不仅得益于自监督预训练 WavLM 模型的强大表征能力, 还有赖于特征解耦与数据增强模块的贡献, 有效提升了系统对复杂环境和未知伪造方法的检测性能。

4.5 综合分析实验

为系统评估本文方法在跨数据集检测不同伪造环境与失真条件下的语音的泛化能力, 我们与当前几种具有代表性的先进方法在 19LA、21LA 和 21DF 数据集上的表现进行了综合对比。

表 5 展示了本文方法与其他方法在三个数据集上的 EER(%)表现, 以及其总和作为综合评估指标。实验结果表明, 尽管 Guo 等人[61]和 Wang 等人[54]的方法在 21DF 数据集上取得了较低的 EER(分别为 2.56%和 2.54%), 但其在 21LA 数据集上的性能相对欠佳, EER 分别为 5.08%和 2.96%, 导致整体表现受限。而 Rosello 等人[57]和 Martín-Doñas 等人[60]虽在 21LA 数据集上具备较强检测能力, 但在 21DF 数据集上的误差率分别达到 5.60%和 4.98%, 表明其对广泛深度伪造手段的适应能力仍存在不足。

本文方法在 19LA 数据集上 EER 为 0.44%, 在 21LA 数据集上 EER 为 1.87%, 在 21DF 数据集上 EER

为 3.57%。综合 EER 合计仅为 5.88%, 相比当前最先进的 Wang 等人[54]的方法降低了 5.77%。这一结果不仅体现了本文方法对未知伪造技术和不同攻击模式的泛化能力, 也验证了其在干扰条件下的鲁棒性。

表 5　综合分析实验结果①

Table 5　Comprehensive analysis of expermental results

方法	EER(%)↓				
	19LA	21LA	21DF	21LA+21DF	19LA+21LA+21DF
XLS-R+LGF[59]	1.28	6.53	4.75	11.28	12.56
Wav2Vec2.0+Classifier[60]	-	3.54	4.98	8.52	8.52+
Ensembling Model[57]	-	2.32	5.60	7.92	7.92+
WavLM+Multi-Fusion Attentive[61]	**0.42**	5.08	2.56	7.64	8.06
Mixture of Experts[54]	0.74	2.96	**2.54**	5.50	6.24
本文方法	0.44	**1.87**	3.57	**5.44**	**5.88**

4.6　消融实验

4.6.1　特征提取器分析实验结果

为探究使用不同的自监督预训练模型作为特征提取器的效果, 本文对比了 Wav2Vec2-XLS-R-300M (XLS-R-300M)、HuBERT-Large-Ls960-Ft、WavLM-Large 三种主流语音大模型在 19LA、21LA 和 21DF 数据集上的表现。实验中, 所有模型均使用基于 AASIST 模型的后端分类器与权重学习模块, 仅更换特征提取器, 且没有使用任何数据增强策略, 使用 EER 与 min t-DCF 作为评估指标。

实验结果如表 6 所示, 三种特征提取器在各数据集上的性能存在显著差异。在 19LA 数据集上, WavLM 实现了 0.53%的 EER 和 0.0137 的 min t-DCF, 显著优于 Wav2Vec2.0 和 HuBERT。在更具挑战性的 21LA 数据集上, WavLM 的 EER 为 1.76%, min t-DCF 为 0.2520, 也低于其他两个模型, 具有更好的鲁棒性。

在检测真实应用场景中具有多样化深度伪造攻击的 21DF 数据集时, WavLM 同样取得了 5.95%的最低的 EER, 相比之下, Wav2Vec2.0 的 EER 为 17.33%, 检测性能明显下降, 而 HuBERT 的表现略逊于 WavLM, EER 为 6.11%。WavLM 模型在三个数据集上均取得了最优结果, 在 DSD 任务上展现出更强的特征表示能力。这一结果表明 WavLM 模型在提取复杂伪造场景下的语音特征时具有更好的泛化能力。因此, 本文选取 WavLM 模型作为特征提取器。

表 6　特征提取器分析实验结果

Table 6　Results of feature extractor analysis experiment

特征提取器	19LA		21LA		21DF
	EER(%)↓	min t-DCF↓	EER(%)↓	min t-DCF↓	EER(%)↓
XLS-R-300M[8]	1.14	0.0383	4.09	0.3148	17.33
HuBERT[10]	0.73	0.0185	2.34	0.2697	6.11
WavLM[9]	**0.53**	**0.0137**	**1.76**	**0.2520**	**5.95**

4.6.2　数据增强分析实验结果

为系统评估不同数据增强策略在 DSD 任务中的实际效用, 本文分别从 RawBoost[14]、Room Impulse Response and Noise Database(RIRs)[38]和 MUSAN[42]三个具有代表性的数据增强方法或数据集中选择噪声进行对比。本文测试了 RawBoost 方法中的 7 种数据增强策略, 分别是线性与非线性卷积噪声(RB1)、脉冲信号相关的加性噪声(RB2)、平稳信号无关的加性噪声(RB3)及它们的组合。RIRs 和 MUSAN 数据集中选取了房间脉冲响应与各向同性噪声(RirIso)、点声源噪声(Point)和来源于 Sound Bible 网站②的真实噪声录音(Real)以及这些噪声的组合。最后, 本文还对比了分别对语音进行 MP3、AAC 和混合使用 MP3 和 AAC 编码的效果。其中 MP3 和 AAC 的比特率都

① XLS-R+LGF[59]未使用数据增强方法; Wav2Vec2.0+Classifier[60]使用低通有限脉冲响应滤波器进行数据增强; Ensembling Model[57]未使用数据增强方法; WavLM+Multi-Fusion Attentive[61]未使用数据增强方法; Mixture of Experts[54]使用 RawBoost 的算法 3 进行数据增强。

② https://soundbible.com/

设置为 96kbps。实验结果如表 7 所示。

表 7 数据增强分析实验结果

Table 7 Results of data augmentation analysis experiment

方法	EER(%)↓			
	19LA	21LA	21DF	总和
RB1	0.80	2.35	5.43	8.58
RB2	**0.34**	18.88	4.60	23.82
RB3	0.54	15.78	6.72	23.04
RB1+RB2+RB3	0.68	**1.08**	6.41	8.17
RB1+RB2	0.61	9.01	5.10	14.72
RB1+RB3	0.57	2.65	5.95	9.17
RB2+RB3	0.64	24.40	9.88	34.92
Point	0.38	4.21	8.18	12.77
RirIso	0.87	5.35	4.76	10.98
Real	0.53	6.84	5.67	13.04
Point+RirIso	0.44	1.87	**3.57**	**5.88**
Point+Real	0.61	5.09	5.11	10.81
RirIso+Real	0.45	21.63	7.40	29.48
Point+Rir+Real	0.67	5.51	5.82	12.00
MP3	0.58	3.41	4.69	8.68
AAC	0.38	3.72	4.39	8.49
MP3+AAC	0.42	12.63	11.17	24.22

从实验结果来看, 在不使用外部数据进行增强的 RawBoost 方法中, 单独使用线性与非线性卷积噪声(RB1)在三个子集上表现相对稳定, 总 EER 为 8.58%, 显著优于仅使用 RB2 或 RB3 的策略, 说明基于卷积失真的建模更具代表性, 能够有效模拟现实中的信道效应与播放设备引起的失真。值得注意的是, RB2 和 RB3 虽然在 19LA 数据集上表现良好(EER 分别为 0.34%和 0.54%), 但在 21LA 上 EER 分别飙升至 18.88%和 15.78%, 这表明当噪声类型缺乏结构信息或与语音信号完全无关时, 模型可能被过度扰动, 反而难以学习稳健的判别特征。相较而言, 组合使用 RB1、RB2 和 RB3 的组合增强策略(RB1+RB2+RB3)能够有效中和不同噪声带来的扰动, 总 EER 降至 8.17%, 且在 21LA 数据集上取得了最优结果(EER 为 1.08%), 优于各自单独使用, 验证了多模式噪声的协同优势。

在使用外部噪声数据增强方面, 单独使用 Point、RirIso 或 Real 等策略的结果较为分散, 其中 Point 噪声在 21DF 数据集上的 EER 为 8.18%, 高于 RirIso 的 4.76%, 说明点声源模拟的方向性噪声对模型构成更大干扰。而组合使用 Point 和 RirIso 的方法取得了最优的综合表现(三个数据集的 EER 总和为 5.88%), 不仅在 21DF 数据集上取得了 3.57%最低的 EER, 在 21LA 数据集上也取得了 1.87%次优的 EER, 表明结构性噪声(如空间混响)与方向性噪声的联合增强更贴近实际应用场景, 有利于模型学习具有判别性的稳健特征。

在压缩编码增强策略方面, 单独使用 MP3 或 AAC 编码进行增强与不使用数据增强的表现相差不大。然而, 当随机混合使用 MP3 和 AAC(MP3+AAC)进行编码时, 性能反而大幅下降, 其原因可能为不同的压缩编码方法对音频引入的失真差异较大, 而频繁切换编码方法会过多干扰导致模型的优化方向, 从而导致模型难以学习到稳定的特征表示。

实验结果表明, 基于现实空间建模的房间脉冲响应与点声源、各向同性噪声的联合增强策略(Point+RirIso)能显著提升模型在复杂语音环境、跨伪造手段下的检测鲁棒性与泛化性, 而 RawBoost 与压缩编码增强策略在调整策略组合与扰动强度后亦具有提升潜力。这些发现为后续增强策略设计提供了经验依据, 同时验证了本文方法在复杂环境下依赖数据增强策略以提升模型鲁棒性的设计合理性。

4.6.3 消融实验结果

为进一步验证本文提出方法中各个关键模块在提升模型鲁棒性与泛化性方面的实际贡献, 本文在以上三个具有不同伪造攻击特征和失真条件的数据集上设计并开展了一系列消融实验。通过逐步引入或去除各功能组件, 本文系统评估了 SWL 机制、自监督语音模型 WavLM 作为特征提取器, 以及数据增强机制(Data Augmentation, DA)对检测性能的影响。实验结果如表 8 所示。

表 8 消融实验结果

Table 8 Results of ablation experiment

模型结构	19LA		21LA		21DF
	EER(%)↓	min t-DCF↓	EER(%)↓	min t-DCF↓	EER(%)↓
AASIST[35]	0.83	0.0275	9.25	0.4599	21.08
AASIST+SWL[17]	1.017	0.0288	3.66	0.3083	21.77
WavLMAST	0.45	**0.0116**	1.78	0.2578	7.25
WavLMAST+SWL	0.53	0.0137	**1.76**	**0.2520**	5.95
WavLMAST+SWL+DA	**0.44**	0.0124	1.87	0.2564	**3.57**

首先, 以原始的 AASIST 模型为基线, 该模型在 19LA 数据集上取得了 0.83%的 EER 和 0.0275 的 min t-DCF, 但在 21LA 和 21DF 数据集上表现明显下降, 分别达到 9.25%和 21.08%的 EER, 表明其在跨分布和带失真场景下的泛化能力有限。引入 SWL 机制后(AASIST+SWL), 在 21LA 测试集上 EER 显著下降至 3.66%, min t-DCF 降至 0.3083, 验证了 SWL 在提升模型稳定性方面的有效性; 但在 21DF 数据集上, EER 反而略有上升, 说明模型的特征表征能力依然不足。

当使用自监督预训练模型 WavLM 替代原始基于 RawNet2 的编码器后(WavLMAST), 在所有数据集上检测性能均有显著提升, 尤其在 21DF 数据集上, EER 从 21.08%降至 7.25%, 显示出更强的跨域鲁棒性。进一步叠加 SWL 机制后(WavLMAST+SWL), EER 继续下降至 5.95%, 表明 SWL 与 WavLM 的协同作用有效增强了模型对复杂伪造手段的判别能力。

最终, 加入数据增强策略后(WavLMAST+SWL+DA), 模型在 19LA 上取得了 0.44%的最低 EER 与 0.0124 的最小 t-DCF, 并在 21LA 和 21DF 上分别达到 1.87%和 3.57%的 EER, 进一步巩固了模型在真实多样场景中的适应能力。这表明数据增强不仅有助于提高模型对噪声与失真的容忍度, 也进一步提升了整体检测系统的泛化性能。

为进一步验证所提出的特征解相关模块在提升模型判别能力方面的有效性, 本文在 19LA、21LA 及 21DF 三个数据集上进行了特征可视化实验, 结果如图 3 所示。可以观察到, 在未引入 SWL 模块的 WavLMAST 模型中, 真实语音与伪造语音的特征在空间中存在一定的重叠, 尤其在 21LA 数据集上重叠部分较多, 在 21DF 数据集上分布边界较为模糊。而在加入 SWL 模块后, 真实与伪造样本在特征空间中的区分度明显增强, 类间间隔更加清晰, 伪造样本的聚类结构也更为紧凑。这一结果直观证明了解相关策略能够有效抑制虚假相关特征, 引导模型聚焦于任务相关的判别性特征, 从而提升检测的泛化性与鲁棒性。

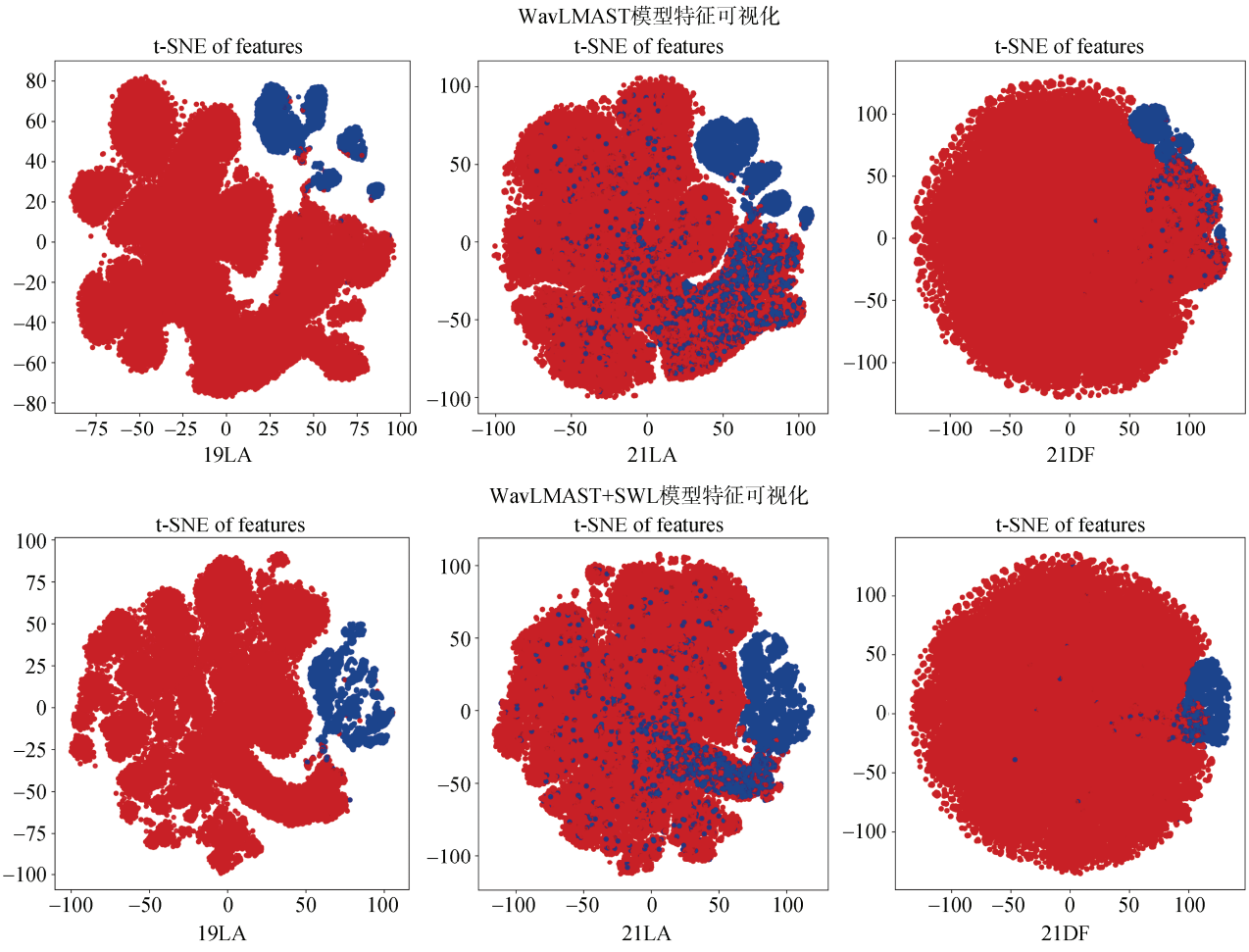

图 3　模型特征可视化图

Figure 3　Visualization of model features

此外, 本文从模型的参数量、计算量、推理延迟以及吞吐量四个维度对模型的效率进行评估。参数量用于衡量模型规模, 其大小直接反映存储和部署的开销。计算量(FLOPs)以浮点运算次数为单位, 表示完成一次前向推理所需的理论运算复杂度, 基于乘加累积操作数(Multiply–Accumulate, MACs)换算得到。推理时间测量的为单个样本的推理时长, 是在 GPU 环境下, 以批量大小 32 的语音样本为输入进行多次测试后取平均值及标准差, 用于衡量单次前向传播的耗时性能。吞吐量则表示模型在稳定运行条件下每秒可处理的语音样本数。

表 9 给出了不同模型在同一实验环境下的效率对比结果。从结果来看, AASIST 及其改进版本 AASIST+SWL 参数量极小, 仅约 0.30M, 其计算量约为 578.62 GFLOPs, 推理延迟维持在 3.71ms 左右, 吞吐量达到 269 samples/s, 体现了其轻量化模型的优势。相比之下, WavLMAST 及其加入 SWL 后的版本在特征建模能力上更强, 但代价是参数量高达 315M, 计算量超过 3.3 TFLOPs, 推理延迟约为 8.7ms, 吞吐量在 114 samples/s 左右。总体而言, AASIST 系列在效率上占优, 更适合资源受限场景, 而 WavLMAST 系列在保证较高推理效率的同时提供了更强的表示能力, 适用于高性能检测任务。

5 结论

本文针对深度伪造语音检测任务中普遍存在的特征表示能力不足、模型学习到虚假相关特征的问题, 设计了一种具有良好鲁棒性与泛化性的检测方法。首先, 本文通过对多种自监督预训练语音模型的特征表示能力进行探索, 提出了一个结合 WavLM 模型与图注意力网络的 WavLMAST 模型, 提高了对深层伪造特征的感知能力。

表 9 模型效率对比

Table 9 Comparison of model efficiency

模型结构	参数量(M)	计算量(FLOPs/G)	推理时间(ms)	吞吐量(samples/s)
AASIST[35]	0.30	578.62	3.71±0.01	269.58
AASIST+SWL[17]	0.30	578.62	3.71±0.01	269.44
WavLMAST	315.54	3328.44	8.73±0.08	114.58
WavLMAST+SWL	315.54	3328.44	8.72±0.22	114.70

其次, 本文引入特征解相关模块, 通过调整训练样本的重要性权重引导模型聚焦与深度伪造语音检测任务稳定、因果相关的特征, 从而在不引入额外目标域数据的条件下, 显著提高了模型在面对训练数据域外的伪造方法与语音环境场景下的泛化性。

此外, 为提升模型在复杂语音环境下的检测性能, 本文设计了多维度数据增强模块, 引入了包括点声源、各向同性和房间脉冲响应等多种噪声来模拟不同真实语音环境下的噪声与混响。该模块有效拓展了训练数据的分布范围, 增强了模型对复杂语音环境条件的鲁棒性。

局限性与未来工作: 尽管本文提出的方法在特征表征能力、泛化性与鲁棒性方面均取得了较好效果, 但仍然存在一定的局限性。首先, WavLMAST 模型参数量较大, 导致在资源受限设备上难以高效部署, 限制了其在实时检测场景中的应用。其次, 特征解相关模块在优化过程中对超参数较为敏感, 不同任务和数据集下需要人工调节, 增加了方法迁移的复杂度。此外, 当前的数据增强策略虽然能够提升噪声环境下的鲁棒性, 但在面对跨语种和未知合成方法时仍存在性能下降。未来工作中将重点探索更高效的轻量化建模方法, 进一步提升特征解相关的自适应性, 并结合跨语种自监督语音模型与生成对抗式数据增强技术, 提升检测系统在真实多样化应用场景下的实用性与普适性。

致 谢 感谢审稿专家提供的宝贵修改意见及改进建议。

参考文献

[1] Alali A, Theodorakopoulos G. Partial Fake Speech Attacks in the Real World Using Deepfake Audio[J]. *Journal of Cybersecurity and Privacy*, 2025, 5(1): 6.

[2] Yuan W G, Zhang X Y, Yuchi X B. Impact Analysis and Recommendations for the Network Security Field by Generative Artificial Intelligence Technology[J]. *Information and Communications Technology and Policy*, 2025(1): 2-9.
(苑卫国, 张新跃, 尉迟学彪. 生成式人工智能技术对网络安全领域的影响分析与启示建议[J]. *信息通信技术与政策*, 2025(1): 2-9.)

[3] Wu Z Z, Yamagishi J, Kinnunen T, et al. ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2017, 11(4):

588-604.

[4] Kang W H, Alam J, Fathan A. Investigation on Activation Functions for Robust End-to-End Spoofing Attack Detection System[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 83-88.

[5] 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(4): 372-1-372-11.

[6] 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.

[7] Cáceres J, Font R, Grau T, et al. The Biometric Vox System for the ASVspoof 2021 Challenge[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 68-74.

[8] 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.

[9] 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.

[10] 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.

[11] Xie Y, Zhang Z C, Yang Y C. Siamese Network with Wav2vec Feature for Spoofing Speech Detection[C]. *Interspeech 2021*, 2021: 4269-4273.

[12] Wang C L, Yi J Y, Tao J H, et al. Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features[C]. *INTERSPEECH 2023*, 2023: 3844-3848.

[13] Schneider S, Baevski A, Collobert R, et al. Wav2vec: Unsupervised Pre-Training for Speech Recognition[EB/OL]. 2019: arXiv: 1904.05862. https://arxiv.org/abs/1904.05862.

[14] 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.

[15] Yamagishi J, Wang X, Todisco M, et al. ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection[EB/OL]. 2021: arXiv: 2109.00537. https://arxiv.org/abs/2109.00537.

[16] Zhang X X, Cui P, Xu R Z, et al. Deep Stable Learning for Out-of-Distribution Generalization[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 5368-5378.

[17] Wang Z Y, Fu R B, Wen Z Q, et al. Generalized Fake Audio Detection *via* Deep Stable Learning[C]. *Interspeech 2024*, 2024: 4773-4777.

[18] Wani T M, Gulzar R, Amerini I. ABC-CapsNet: Attention Based Cascaded Capsule Network for Audio Deepfake Detection[C]. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024: 2464-2472.

[19] Reynolds D. Gaussian Mixture Models[M]. Encyclopedia of Biometrics. Boston, MA: Springer US, 2009: 659-663.

[20] Suthaharan S. Support Vector Machine[M]. Machine Learning Models and Algorithms for Big Data Classification. Boston, MA: Springer US, 2016: 207-235.

[21] Popescu M C, Balas V E, Perescu-Popescu L, et al. Multilayer Perceptron and Neural Networks[J]. *WSEAS Transactions on Circuits and Systems*, 2009, 8(7): 579-588.

[22] Villalba J, Miguel A, Ortega A, et al. Spoofing Detection with DNN and One-Class SVM for the ASVspoof 2015 Challenge[C]. *Interspeech 2015*, 2015: 2067-2071.

[23] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7132-7141.

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Advances in neural information processing systems*, 2017, 30.

[25] 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.

[26] 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.

[27] Wu X, He R, Sun Z N, et al. A Light CNN for Deep Face Representation with Noisy Labels[EB/OL]. 2015: arXiv: 1511.02683. https://arxiv.org/abs/1511.02683.

[28] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks[M]. Backpropagation. Psychology Press, 2013: 35-61.

[29] 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.

[30] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-Augmented Transformer for Speech Recognition[EB/OL]. 2020: arXiv: 2005.08100. https://arxiv.org/abs/2005.08100.

[31] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.

[32] Lin M, Chen Q, Yan S C. Network in Network[EB/OL]. 2013: arXiv: 1312.4400. https://arxiv.org/abs/1312.4400.

[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] Babu A, Wang C H, Tjandra A, et al. XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale[EB/OL]. 2021: arXiv: 2111.09296. https://arxiv.org/abs/2111.09296.

[35] 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.

[36] 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/OL]. 2021: arXiv: 2107.12710. https://arxiv.org/abs/2107.12710.

[37] Yi J Y, Wang C L, Tao J H, et al. Audio Deepfake Detection: A Survey[EB/OL]. 2023: arXiv: 2308.14970. https://arxiv.org/abs/2308.14970.

[38] Ko T, Peddinti V, Povey D, et al. A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition[C]. *2017 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2017: 5220-5224.

[39] Nakamura S, Hiyane K, Asano F, et al. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition[J]. *Proceedings of ICLRE*, 2000.

[40] Kinoshita K, Delcroix M, Yoshioka T, et al. The Reverb Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech[C]. *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2014: 1-4.

[41] Jeub M, Schafer M, Vary P. A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms[C]. *2009 16th International Conference on Digital Signal Processing*, 2009: 1-5.

[42] Snyder D, Chen G G, Povey D. MUSAN: A Music, Speech, and Noise Corpus[EB/OL]. 2015: arXiv: 1510.08484. https://arxiv.org/abs/1510.08484.

[43] Rahimi A, Recht B. Random Features for Large-Scale Kernel Machines[C]. *The 21st International Conference on Neural Information Processing Systems*, 2007: 1177-1184.

[44] Bahng H, Chun S, Yun S, et al. Learning De-Biased Representations with Biased Representations[C]. *The 37th International Conference on Machine Learning*, 2020: 528-539.

[45] Strobl E V, Zhang K, Visweswaran S. Approximate Kernel-Based Conditional Independence Tests for Fast Non-Parametric Causal Discovery[J]. *Journal of Causal Inference*, 2019, 7: 20180017.

[46] Kuang K, Xiong R X, Cui P, et al. Stable Prediction with Model Misspecification and Agnostic Distribution Shift[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(4): 4485-4492.

[47] Veaux C, Yamagishi J, MacDonald K. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit[J]. *University of Edinburgh. The Centre for Speech Technology Research*, 2017, 6: 15.

[48] Wang X, Yamagishi J, Todisco M, et al. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech[J]. *Computer Speech & Language*, 2020, 64: 101114.

[49] Wu Z Z, Kinnunen T, Evans N, et al. ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge[C]. *Interspeech 2015*, 2015: 2037-2041.

[50] Kinnunen T, Lee K A, Delgado H, et al. T-DCF: A Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification[EB/OL]. 2018: arXiv: 1804.09618. https://arxiv.org/abs/1804.09618.

[51] McFee B, Raffel C, Liang D W, et al. Librosa: Audio and Music Signal Analysis in Python[J]. *Proceedings of the 14th Python in Science Conference*, 2015: 18-24.

[52] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[EB/OL]. 2014: arXiv: 1412.6980. https://arxiv.org/abs/1412.6980.

[53] Huang B Y, Cui S S, Huang J W, et al. Discriminative Frequency Information Learning for End-to-End Speech Anti-Spoofing[J]. *IEEE Signal Processing Letters*, 2023, 30: 185-189.

[54] Wang Z Y, Fu R B, Wen Z Q, et al. Mixture of Experts Fusion for Fake Audio Detection Using Frozen Wav2vec 2.0[C]. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2025: 1-5.

[55] Yang M J, Zheng K F, Wang X J, et al. Comparative Analysis of ASV Spoofing Countermeasures: Evaluating Res2Net-Based Approaches[J]. *IEEE Signal Processing Letters*, 2023, 30: 1272-1276.

[56] Liu X H, Liu M, Wang L B, et al. Leveraging Positional-Related Local-Global Dependency for Synthetic Speech Detection[C]. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2023: 1-5.

[57] Rosello E, Gomez A M, López-Espejo I, et al. Anti-Spoofing Ensembling Model: Dynamic Weight Allocation in Ensemble Models for Improved Voice Biometrics Security[C]. *Interspeech 2024*, 2024: 497-501.

[58] Interspeech 2024. ISCA Archive. https://www.isca-archive.org/interspeech_2024/index.html. Sept. 2024.

[59] Wang X, Yamagishi J. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures[C]. *The Speaker and Language Recognition Workshop*, 2022: 100-106.

[60] Martín-Doñas J M, Álvarez A. The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge[C]. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2022: 9241-9245.

[61] 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.

王帅斌 于 2023 年在青海大学计算机科学与技术专业获得学士学位。现在中国科学院信息工程研究所网络空间安全专业攻读硕士学位。研究领域为多媒体信息安全。研究兴趣包括深度伪造语音检测。Email: wangshuaibin@iie.ac.cn

易小伟 于 2014 年在中国科学院软件研究所获得博士学位, 现任中国科学院信息工程研究所副研究员、CCF 会员。研究领域为网络空间安全、多媒体内容智能分析。研究兴趣包括信息隐写与隐写分析、多媒体内容取证、人工智能安全。Email: yixiaowei@iie.ac.cn

刘长军 于 2003 年获得国防科技大学硕士学位, 现任中国科学院信息工程研究所高级工程师。研究领域为网络体系结构与安全防护、系统安全理论与技术。Email: liuchangjun@iie.ac.cn

苏小苏 于 2023 年在北京工商大学模式识别与智能控制专业获得硕士学位。现在中国科学院信息工程研究所网络空间安全专业攻读博士学位, 研究领域为语音理解与生成、语音大模型。Email: suxiaosu@iie.ac.cn

曹纭 于 2012 年在中国科学院软件研究所获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为多媒体内容安全。研究兴趣包括隐写与隐写分析、数字内容取证等。Email: caoyun@iie.ac.cn

吕美杨 于 2023 年在北京工业大学信息安全专业获得学士学位。现在中国科学院信息工程研究所网络空间安全专业攻读硕士学位。研究领域为多媒体信息安全。研究兴趣包括信息隐藏、隐写与隐写分析。Email: lvmeiyang@iie.ac.cn