基于时频特征融合的伪造语音检测方法

袁程胜[1,2] 张雪原[2,3] 周志立[4] 李欣亭[5] 付章杰[1,2]

(1. 南京信息工程大学 计算机学院、网络空间安全学院，南京 210044)
(2. 南京信息工程大学 数字取证教育部工程研究中心，南京，210044)
(3. 南京信息工程大学 软件学院，南京，210044)
(4. 广州大学 人工智能研究院，广州，510006)
(5. 国防科技大学 外国语学院，南京，210039)

摘 要：针对伪造语音检测精度不高和泛化性弱的难题,提出一种基于时频特征融合的伪造语音检测算法。首先,为了挖掘语音片段能量分布不均、基频波动异常,以及提取语义连贯性的细微差别,提出一种多分支特征融合网络,分别从音高、音强以及能量分布来挖掘真假语音的差异痕迹,以更好地表征真假语音的频率变化、振幅变化和峰值差异,提高伪造语音检测的准确率。其次,经典的坐标注意力机制未能对语音时频域的细粒度差异进行有效挖掘,为此提出一种时频坐标注意力机制,分别从时域和频域两个方向对能量分布和基频波动异常进行联合编码,以更好地表征频谱图中的共性高频能量异常,提升模型的泛化性。最后,又设计了一种自适应联合损失优化函数,通过平衡不同分支网络的权重,进一步提升模型对伪造语音中高频能量异常及语义不连贯性的学习能力。在 ASVspoof 2019 的逻辑访问数据集上进行了性能评估,实验结果表明,与现有的工作相比,所提方法在等错误率（Equal Error Rate, EER）和最小归一化串联检测代价函数（Minimum Normalized Tandem Detection Cost Function, min t-DCF）两个指标均取得较好性能,分别降低了 0.34%和 0.014。此外,在应对极难检测的未知攻击 A17 时,同样展现出较高的泛化性,其中 EER 和 min t-DCF 分别下降了 3.9522%、0.1364。当应对未知类型的欺骗攻击时,同样表现出较好的泛化性。

关键词：伪造语音检测；特征融合；时频特征融合；频谱图；语音波形
中图分类号：TN912 **文献标识码：**A

Forged Speech Detection Algorithm Based on Time-Frequency Feature Fusion

YUAN ChengSheng[1,2], ZHANG XueYuan[2,3], ZHOU ZhiLi[4], LI XinTing[5], FU ZhangJie[1,2]

(1. School of Computer Science, School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China)
(2. Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China)
(3. School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China)
(4. Institute of Artificial Intelligence, Guangzhou University, Guangzhou 510006, China)
(5. School of Foreign Languages, National University of Defense Technology, Nanjing 210039, China)

Abstract: To solve the problem of low accuracy and weak generalization of forged speech detection, a new algorithm based on time-frequency feature fusion is proposed. Firstly, in order to excavate the uneven energy distribution of speech fragments or the abnormal fundamental frequency fluctuation, and extract the subtle difference of semantic coherence, a multi-branch feature fusion network is proposed to excavate the difference traces of true and false speech from the pitch, pitch intensity and energy distribution respectively, so as to better represent the frequency change, amplitude change and peak difference of true and false speech, and improve the accuracy of forged speech detection. Secondly, the classical coordinate attention mechanism fails to effectively mine the fine-grained differences in the time-frequency domain of speech. Therefore, a time-frequency coordinate attention mechanism is proposed to jointly encode the energy distribution

基金项目：国家自然科学基金项目(62102189; 62122032)，国家社会科学基金(2022-SKJJ-C-082)，国防科大科研项目(JS21-4; ZK21-43)
收稿日期：2024-10-13；**修订日期**：2024-**-**

and fundamental frequency fluctuation anomalies from the time domain and the frequency domain respectively, so as to better characterize the common high frequency energy anomalies in the spectral graph and improve the generalization of the model. Finally, an adaptive joint loss optimization function is designed to balance the importance of different branch networks to further improve the model's ability to learn high frequency energy anomalies and semantic incoherence in forged speech. Performance was evaluated on the logical access (LA) dataset of ASVspoof 2019, and experimental results showed that compared with the current methods, the proposed method achieved good performance in both EER(Equal Error Rate) and mint-DCF(Minimum Normalized Tandem Detection Cost Function) indicators, which decreased by 0.34% and 0.014, respectively. In addition, when dealing with unknown attack A17, which is extremely difficult to detect, it also showed good generalization, where EER and mint-DCF decreased by 3.9522% and 0.1364, respectively. When dealing with unknown types of spoofing attacks, it also shows better generalization.

Key words: forged speech detection; feature fusion; time frequency feature fusion; spectrogram; speech waveform

1 引言

移动互联和智能穿戴设备的普及给人们的日常生活带来了诸多便捷。但是近些年, 因敏感内容保护不当而导致的数据泄露、隐私窃取问题愈发凸显, 对个人信息的安全维护构成了严峻挑战。因此, 验证用户身份的真实性和合法性成为保障用户隐私安全的首要防线。相较于指纹、人脸和行为等各类生物特征识别技术,语音识别技术凭借其非接触式交互、低认知负荷及场景普适性的特性, 深受大众青睐[1][2]。尤其是新冠疫情期间，其“零接触”特性更使其成为公共健康安全与数字身份核验平衡的优选技术。然而,最新的研究成果揭示, 语音识别技术正遭受严峻的伪造语音欺诈攻击挑战。特别是, 借助深度伪造技术, 如今已能够生成在频谱特性、情感韵律乃至呼吸节奏方面都与真人无异的语音样本。更严峻的是, 攻击者正将目标从通用语音库转向特定个体声纹克隆, 仅需少量目标语音样本,就能构建出高精度的声纹模型, 这使得“听觉欺骗”从技术可能演变为现实威胁。因此, 鉴别待认证的语音是否源自自然人, 对于提升语音识别系统的安全性和可靠性至关重要。鉴于伪造语音欺骗攻击的严峻态势, 伪造语音检测技术应运而生。然而, 如何切实有效地提升伪造语音检测任务的准确率与泛化能力, 依旧是当前研究人员亟待解决的关键难题。

目前,伪造语音技术主要分为传统式和深度学习式两大类。前者主要依赖于波形拼接和参数统计两种方法。但是,该类方法伪造的语音时常面临韵律信息丢失（如重音、音调、节奏和停顿等）、噪声干扰以及信号失真等问题, 使其在自然度和流畅度上表现不佳, 且易被人耳辨识。而后者则是通过设计定制化的伪造模型, 自动生成高度逼真的合成语音。为了抵御伪造语音的欺骗攻击, 提升语音识别系统的安全性和可靠性, 一系列伪造语音检测技术相继被提出。通过对伪造语音检测技术梳理归纳后发现, 现有方法大多仅对单一频谱或波形特征进行分析, 忽略了伪造语音片段间的语音连贯性, 以及能量分布不均、基频波动异常情况, 致使现有方法在检测精度方面仍然不高。此外, 伪造语音在生成过程中为了提高与原声的相似度, 会将多个声道的语音叠加, 从而导致声谱分布连续性较差, 并且在高频区域容易出现异常的波峰。

本文受多模态特征融合思想的启发,提出一种基于时频特征融合的伪造语音检测方法, 分别从音高、音强以及能量分布信息三个角度挖掘伪造语音中的差异信息, 以更好地表征真假语音中的频率变化、振幅变化和峰值差异等表征真伪语音本质差异的信息。本文的主要贡献有,

(1) 为了挖掘语音片段能量分布不均、基频波动异常, 以及提取语义连贯性的细微差别, 提出一种多分支特征融合网络(Multi-Branch Feature Fusion Network, MFFNet),分别从音高、音强以及能量分布信息角度挖掘语音中的高频能量异常以及语义不连贯信息,以更好地表征真假语音的频率变化、振幅变化和峰值差异。

(2) 经典的坐标注意力机制对语音的时频域处理能力较弱, 未能提取频谱图中的细粒度伪造痕迹, 为此提出一种时频坐标注意力机制(Time-frequency Coordinate Attention Mechanism, TCAM), 通过对时域和频域上的能量分布和基频波动异常进行联合编码,以更好表征真假语音的细微差异。

(3) 为了评估不同分支特征学习网络的重要性,剔除多分支中的冗余特征,还设计了一种自适应联合损失优化函数(Adaptive Loss Optimization Function, ALOF),通过融合交叉熵损失和 MultimarginLoss, 以平衡不同分支网络的权重, 通过联合训练来挖掘不同分支下的高频能量共性特征。

(4) 在 ASVspoof 2019 逻辑访问(Logical Access, LA)数据集上进行了性能测试, 实验结果显示, 所提方法相较于现有工作, 在 EER 和 min t-DCF 两项指标上分别为 0.35%和 0.011,表现出较好的检测精度。此外, 在

应对未知攻击 A17 时同样表现出极高的泛化性。

2 相关工作

通过对现有的工作梳理后发现,伪造语音检测技术主要分为四类: 传统检测方法、基于深度特征学习的检测方法、基于端到端的检测方法和基于特征融合的检测方法。本节将对上述方法进行详细介绍。

2.1 传统检测方法

传统检测方法首先从原始语音信号中提取频谱特征,然后深入分析这些频谱特征,挖掘出频率的细微变化、峰值差异等关键信息,最后利用分类器对上述特征进行建模与测试。其中最具代表性的频谱特征主要包含三类,即 Mel 频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)、常数 Q 倒谱系数(Constant Q Cepstral Coefficients, CQCC)和线性频率倒谱系数(Linear Frequency Cepstral Coefficients, LFCC)。基于此,Patel 等人[3]将其应用于语音识别任务中,提出了一种基于帧级特征的伪造语音检测方法,通过提取 Mel 频率倒谱系数、耳蜗滤波器倒谱系数(Cochlear Filter Cepstral Coefficients, CFCC)以及耳蜗滤波器倒谱系数与瞬时频率变化的结合特征(Cochlear Filter Cepstral Coefficients and Change in Instantaneous Frequency, CFCC-IF)来捕捉跨帧的特征变化, 降低噪声的干扰,实验结果表明, 该方法在检测准确率上取得了不错的性能。在 2016 年举办的 BTAS 反欺骗语音挑战比赛中,KORSHUNOV 等人[4]提出了一种基于长期平均频谱和标准偏差信息的伪造语音检测方法,利用线性判别分析(Linear Discriminant Analysis, LDA)模型充分学习通道退化和语音自然性等信息,最终在鲁棒性测试中表现出不错的性能。Sahidullah 等人[5]使用最大似然高斯混合模型(Gaussian Mixture Model with Maximum Likelihood, GMM-ML)和局部二值模式支持向量机(Local Binary Pattern Support Vector Machine, LBP-SVM)分类器, 分别对 LFCC 特征进行训练和测试, 使模型更关注语音的高频区域。Todisco 等[6]将 CQCC 特征引入到语音伪造检测中,使用高斯混合模型捕获语音中的高低频伪造痕迹,取得当时的最佳性能(State of the arts, SOTA)。

虽然传统的检测方法在应对伪造语音检测任务取得了不错的性能, 但是该类方法极度依赖手工特征工程, 以及核函数的选择和设计, 普适性不高。

2.2 基于深度学习的检测方法

该类方法通过对手工特征进一步学习,以挖掘更深层次的伪造痕迹。鉴于深度学习在图像分类任务中的出色表现,部分研究人员也开始将其应用于伪造语音检测中[7]。Lavrentyeva 等人[8]提出了一种基于轻量化卷积神经网络(Lightweight Convolutional Neural Network, LCNN)的伪造语音检测方法。该方法首先提取语音的频谱特征，然后通过 LCNN 网络对其中的伪造痕迹进行学习。尽管该方法在解决模型参数规模问题上表现出色，但随着模型层数的急剧增加，却极易陷入梯度消失的困境。为此, Lai 等人[9]将残差神经网络模型引入伪造语音检测任务中, 提出一种基于残差连接的伪造语音检测方法,实验结果表明,所提方法能够解决梯度消失问题。但是,该方法无法直接应用于序列性的语音检测任务。为此, GS 等人[10]提出一种基于双向长短期记忆网络(Bi-directional Long Short-Term Memory Network, Bi-LSTM)的伪造语音检测方法,通过对时间序列和远程依赖进行建模,以捕获频谱特征中的序列性伪造痕迹。Gomez-Alanis 等人[11]将轻量化的门控递归神经网络作为特征提取器,概率线性判别分析作为分类器, 进一步提高伪造语音检测的性能。

当前,尽管基于深度特征学习的检测方法在提升模型检测性能方面展现出了一定的优势, 但在频谱特征提取的过程中仍会不可避免地丢失部分关键信息。此外, 分类器的选择过于依赖人工操作, 难以实现特征提取与分类器之间的双向最优化,从而限制了检测性能的进一步提升。

2.3 基于端到端的检测方法

该类方法是直接构建一个能够鉴别语音真伪的检测模型,无需执行一系列的复杂设计。为了提取更深层次的语义特征, Jung 等人[12]设计了一个 RawNet 模型,提高了真假语音检测任务的准确率。Tak 等人[13]提出一种基于图注意力网络的检测方法, 能够更好地提取不同频率子带和时间段之间的关联性。此外,通过将图神经网络和其它深度学习模型融合,一定程度上弥补了单一网络模型检测精度不高的困扰。但是,该方法过

于依赖语音中的相位信息。为此, Wang 等人[14]利用 RawNet2 网络进行特征提取, 并设计了一个名为 Simple Attention Module (SimAM)的注意力模块来对特征图进行深入分析。同时, 他们还引入了均方误差作为模型的元训练策略,有效提升了检测模型的泛化能力。在 Wang 等[15]人的另一项工作中, 通过将正交卷积与时间卷积网络相结合, 并输入 RawNet 网络,构建了 TO-RawNet 网络, 能够进一步提高伪造语音的检测精度。

2.4 基于特征融合的检测方法

基于特征融合的伪造检测方法通过融合各种相同或不同的特征,以增强特征的表征能力,进一步提高任务的检测性能。Xue 等人[16]提出了一种基于生理-物理特征融合的伪造语音检测方法, 通过卷积神经网络提取音频中与人脸相关的生理特征(如性别、嘴型等)以及物理特征来识别语音欺骗攻击。Wang 等人[17]分别使用隐藏单元 Bert(Hidden-Unit BERT, HuBERT)和 Conformer 模型提取语音的音素特征, 并将其与 Wav2Vec2 模型提取的特征进行融合, 通过深入分析原始语音信号的语调、音高等韵律信息来检测伪造语音。Liu 等人[18]提出了一种创新方法, 将原始语音转换为立体声, 并从左右声道两个方面获取深层特征表示,最终将这两种深层特征进行融合,以获取更为全面且细致的伪造痕迹信息。

尽管基于特征融合的方法在提升伪造语音的检测性能上取得了一定成效, 但是仍然面临特征冗余和细粒度伪造痕迹难以挖掘等挑战。

3 基于时频特征融合的伪造语音检测模型

针对当前伪造语音检测任务所面临的诸多挑战,如难以精准捕获语音连贯性、难以检测高频区域出现的异常波峰、以及语音波形特征对相位和幅度信息过度依赖等不足, 提出了一种新型伪造语音检测方法。本节首先将阐述多分支特征融合网络的设计过程。然后对现有的坐标注意力机制的局限性进行说明, 并提出一种时频坐标注意力机制,能够增强差异特征的提取和表征能力。最后, 为了评估不同分支特征的重要性并剔除冗余特征, 本文还设计了一个自适应联合损失优化函数,进一步提高伪造语音检测任务的准确性。

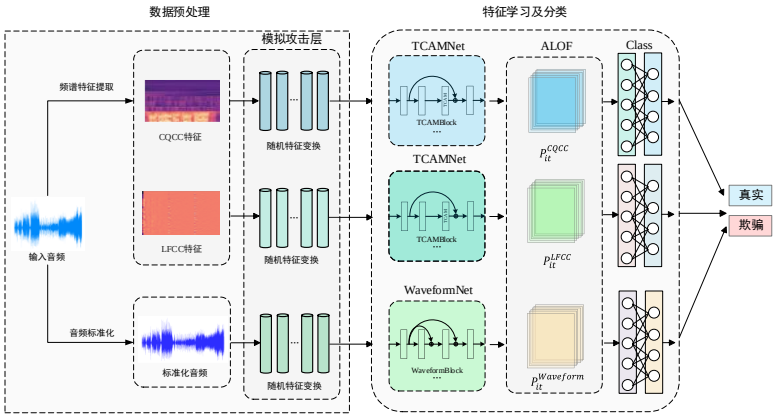


图 1 多分支特征融合网络(MFFNet)的总体架构

Fig.1 Overall architecture of Multi-branch Feature Fusion network (MFFNet)

3.1 多分支特征融合网络

为解决当前特征学习方法在频谱信息丢失和挖掘不同伪造语音技术的共性特征方面难的问题,提出一种基于时频特征融合的伪造语音检测算法,具体实现流程如图 1。与直接在物理层面进行特征融合不同，所提方法采用一种基于分类结果层面的均值融合策略,旨在避免因直接融合不同维度信息而导致的特征冗余,以及规避可能造成的特征维度灾难。通过将时域和频域特征在分类层面进行融合,能够使模型专注于各自特征空间的模式学习,充分利用不同特征维度的独特性。在分类结果融合阶段,整合了时域和频域特征所捕获的关键分类信息,有效提升了检测性能和泛化能力。

本文方法旨在从语音波形、LFCC 和 CQCC 三个维度充分挖掘伪造语音的细粒度特征。首先, 借助频谱变换技术,从原始的语音信号中提取出 CQCC 和 LFCC 两类频谱特征。随后,将这些特征输入到模拟攻击层（Simulation Attack Layer, SAL）之中, 通过模拟随机噪声和信号增益等操作，在时频域内实施双重扰动策

略,使得模型能够在信号干扰和传输等复杂环境下,能够精准地学习到伪造语音所独有的特征,频谱变换 $S(\cdot)$ 和时域波形变换 $W(\cdot)$ 实现公式如下,

$$S_{shift}(t,f)=X(|(t-\Delta t)\bmod T|,|(f-\Delta f)\bmod F|) \quad (1)$$

首先,为了模拟语音信号在时间和频率上的特征偏移干扰,如公式(1)所示, 对二维频谱信号 $X(t,f)$ 在时频两个维度上分别执行周期性偏移(偏移量分别为 Δt 和 Δf),并采用模运算 mod(T 和 F 分别表示时频维度的最大范围)来限制偏移范围。为了进一步模拟语音信号在时间或频率范围内的攻击场景,分别引入了时域 $M_t(t_{mask})$ 和频域 $M_f(f_{mask})$ 掩码函数,

$$S_{mask}(t,f)=X(t,f)\odot M_t(t_{mask})\ M_f(f_{mask}) \quad (2)$$

对频谱信号 $X(t,f)$ 部分区域(t_{mask} 和 f_{mask} 范围内)执行时频域的信号掩蔽操作。其中, $M(\cdot)$ 值为 0 表示该区域内信号被掩蔽,值为 1 表示该区域信号未被掩蔽，计算公式如下,

$$M_t(t_{mask})=\begin{cases}0, & t_{mask}\in[t_{start},t_{start}\quad L_t]\\ 1, & otherwise\end{cases} \quad (3)$$

$$M_f(f_{mask})=\begin{cases}0, & f_{mask}\in[f_{start},f_{start}\quad L_f]\\ 1, & otherwise\end{cases} \quad (4)$$

其中, t_{start} 和 f_{start} 分别为指定掩蔽区域的起始位置, L_t 和 L_f 用来确定掩蔽区域的范围。通过随机掩蔽时频域中部分信息,增强模型对语音局部信息丢失的适应性,进而提升对低质量信号的感知能力。为了进一步增强模型在面临各种时频域攻击噪声时的鲁棒性,还在时频域引入随机噪声,具体如下，

$$S_{noise}(t,f)=X(t,f)+\sigma\cdot M,\quad M\sim N(0,1) \quad (5)$$

对频谱信号 $X(t,f)$ 添加服从标准正态分布 $N(0,1)$ 的高斯噪声矩阵 M , 通过参数 σ 控制噪声强度。为了增强模型对时域信号变化时的鲁棒性,对频谱信号 $X(t,f)$ 执行线性变换来模拟信号增益,求解公式为

$$S_{adjust}(t,f)=X(t,f)\cdot\lambda,\quad \lambda\sim U(1\quad\varepsilon,1+\varepsilon) \quad (6)$$

其中, $\lambda\sim U(1-\varepsilon,1+\varepsilon)$ 为服从连续型均匀分布的随机增益因子。为了进一步增强模型对语音信号时间轴偏移干扰的适应能力,还对语音波形信号 $X(t)$ 执行时间轴上的平移操作,即

$$W_{shift}(t)=X(|(t-\Delta t)\bmod T|) \quad (7)$$

此外,为增强模型应对不同波形信号强度下的检测能力,对语音波形信号 $X(t)$ 执行信号增益操作,随机增益因子 $\lambda\sim U(1-\varepsilon,1+\varepsilon)$ 服从连续型均匀分布,计算公式如下,

$$W_{adjust}(t)=X(t)\cdot\lambda,\quad \lambda\sim U(1\quad\varepsilon,1+\varepsilon) \quad (8)$$

接下来,模型需进一步提升对语音波形噪声的适应性。为此,向语音波形信号 $X(t)$ 添加随机噪声 I ,并通过参数 γ 调节噪声强度,

$$W_{noise}(t)=X(t)+\gamma\cdot I,\quad I\sim N(0,1) \quad (9)$$

最终,为了模拟语音复杂干扰场景,通过双重变换实现组合攻击,具体如下,

$$\begin{aligned}&X_{enhanced}=\begin{cases}f_k(f_j(X)), & if\ X\in\{CQCC,LFCC\}\\ g_k(g_j(X)), & if\ X\in Waveform\end{cases}\\ &f_k,f_j\in\{S_{shift},S_{mask},S_{noise},S_{adjust}\},g_k,g_j\in\{W_{shift},W_{adjust},W_{noise}\}\end{aligned} \quad (10)$$

其中, $f(\cdot)$ 和 $g(\cdot)$ 分别表示单次特征变换($S(\cdot)$ 或 $W(\cdot)$ 变换),信号 X 经过组合特征变换后得到的 $X_{enchanced}$ 可更真实地模拟复杂语音干扰场景。如果信号 X 为频谱特征，则对频谱信号应用两次频谱变换 f_k , f_j ; 若信号 X 为语音波形特征,则应用两次波形变换 g_k , g_j 。

为了进一步挖掘不同伪造技术中的不变共性特征,将增强后的特征分别送入 TCAMNet 和 WaveformNet 两个专用网络进行处理。在 CQCC 特征提取分支中, 利用 TCAMNet 骨干网络（如图 2 所示）深入分析长期窗口变换的 CQCC 特征,以学习频谱图中各位置的能量分布模式。此外,该分支还能有效捕捉低频部分的高分辨率信息,从而更精确地刻画真伪语音在能量分布上的差异。WaveformNet 网络（内部结构如图 3）则从语音波形中提取深度特征。利用其内置的池化模块,能够对这些特征进行特征降维,以更精准地挖掘波形特征中的伪造痕迹。

在 LFCC 特征提取分支中,利用线性滤波器精准提取语音信号中的谐波结构、共振峰以及高频区域能量等关键信息。该分支与 CQCC 特征提取分支采用相同的网络结构和学习策略,以确保特征提取的一致性和有效性。选择频谱特征作为输入的优势在于有助于降低模型对相位信息的过度依赖,提高模型的检测精度和稳定性。最后, 设计了 ALOF (Adaptive Loss Optimization Function) 模块,能够自适应地调整模型的联合损失,减少模型对语音中噪声以及特征中冗余信息的学习,进一步提升伪造语音检测的性能。

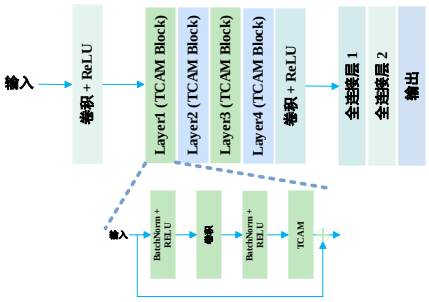


图 2 TCAMNet 学习架构

Fig.2 TCAMNet learning architecture

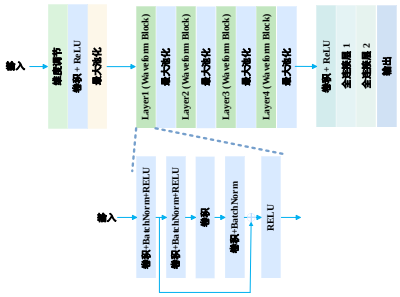
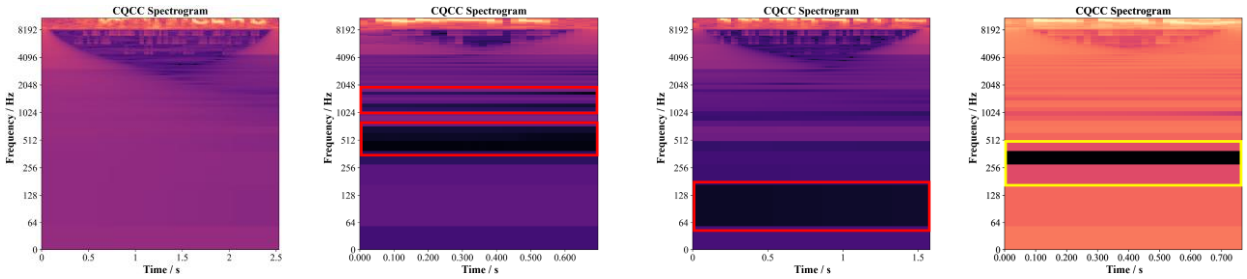


图 3 WaveformNet 学习架构

Fig.3 WaveformNet learning architecture

3.2 时频坐标注意力机制

坐标注意力机制在计算机视觉领域得到了广泛应用,核心思想是通过在通道注意力的基础上引入位置信息,以更好地捕捉特征的局部判别性信息[19]。为了有效提取语音中的伪造痕迹和语义不连贯性特征,本文首次将坐标注意力机制引入到伪造语音检测任务中。文本转语音（Text-To-Speech, TTS）作为当前主流的语音伪造技术,充分考虑了人耳对部分频带敏感度较低的特点,在生成语音时降低了部分频率的质量。这种处理方式致使合成语音在不同频带之间的能量分布不均匀,呈现出不正常的能量连贯性,如图 4 所示。伪造语音在部分频带之间普遍存在能量异常,具体体现在频谱图的不同频带部分出现能量损失或增益不均衡的现象。通过观察图 4(b)中的伪造语音频谱图,可以明显发现部分频带的能量分布相较于真实语音存在明显的缺失,呈现出能量异常的断层。这些频带能量异常现象是鉴别伪造语音的重要特征,为检测模型提供了有力的判别依据。



(a) 真实语音 (b) 伪造语音 I (c) 伪造语音 II (d) 伪造语音 III

图 4 不同语音伪造痕迹在 CQCC 频谱图中的位置

Fig.4 Positions of different speech forgery traces in CQCC spectrum

坐标注意力机制此前主要应用于图像分类任务,但在处理音频的时频域方面,其能力相对受限,难以挖掘真假语音间细微的伪造特征,从而影响了语音伪造检测任务的性能。鉴于此,本文对其进行了优化改进,提出了一种专门用于伪造语音检测的时频坐标注意力机制。

在具体实施中,首先通过一组卷积操作对输入的频谱图进行时频信息提取。然而,经典的坐标注意力机制采用的是平均池化,很难捕捉到某些能量异常区域,反而可能导致伪造痕迹信息被模糊处理,进而对伪造检测的性能产生不利影响。为了挖掘全局特征中的局部变化,改用了针对于时频域的最大自适应池化操作,以捕获全局特征中的局部变化,如频谱中峰值变化的尖锐区域和能量不连续的跳变区域。通过该操作能够在保证全局信息不受损失的条件下,准确提取伪造语音中微小能量变化和频谱失真信息,避免微小的时频特征被忽略。随后,池化后的时频域特征会经过特征连接和分离等操作,进一步增强对不同频带及时间段特征能量信息的提取。具体而言,通过特征连接操作,将来自时频域尺度的特征信息融合在一起,构建一个多尺度的特征表示。该融合能够让模型同时捕捉到频谱中不同频率段的全局信息和局部细节,提升模型的时频特征表达能力。特征分离操作则通过将频域和时域的特征执行分离操作,有助于在时频域内分别强化伪造痕迹的表征,特别是语音信号的高频异常和时序上的语义不连贯性。在注意力中间层,将最大通道数固定为 64,以确保在特征提取过程中保持足够的特征丰富度。此外,考虑到注意力网络中深层的通道数量会逐渐衰减,通过设计适当的通道衰减策略来避免模型的特征学习能力过早下降。通过在每一层中加入一定衰减比例的约束通道数,使得时频信息的压缩和提炼保持在合理的范围内,进而最大程度地提高检测精度。为了进一步提高语音检测模型的稳定性和泛化能力,本文在该注意力机制激活函数的选择上还进行了优化。本文采用 ReLU4 激活函数,它是一种调整后的 ReLU 函数,能够限制激活值的上界为 4。相比于标准的 ReLU, ReLU4 在伪造语音检测模型收敛过程中表现出更强的稳定性,特别是在高频特征和能量异常区域的检测中,可大幅减少特征值过大导致的检测误差。

其内部结构如图 5 所示,输入 X 经过 TCAM 后,首先通过自适应最大池化层对频谱图的时间方向和频率方向进行池化操作,得到两个池化后的频域和时域张量 x_f和 x_t。然后将这两个张量拼接为一个新的张量 y,并对其执行双层卷积、批量归一化和激活等操作,得到处理后的 y。接着,将 y 进行特征分离并再次生成张量 x_f 和 x_t,对其再次执行时频域卷积操作。最后,利用激活函数分别计算两个注意力模块的权重,并将这两个注意力权重与输入特征 X 相乘,作为该模块的最终输出。

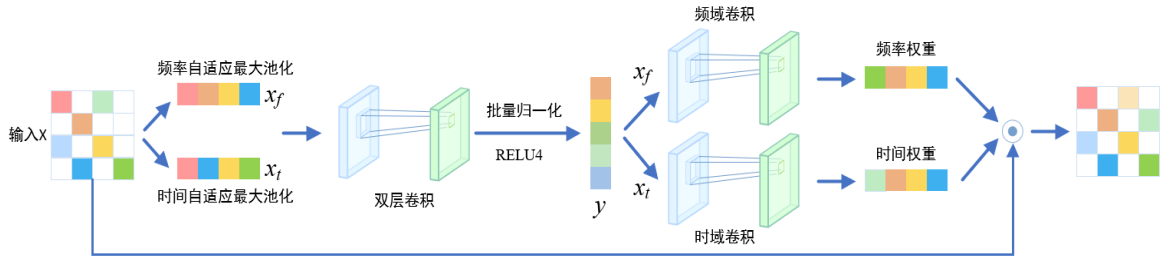


图 5 时频坐标注意力机制

Fig.5 Time-frequency Coordinate Attention Mechanism

3.3 自适应联合损失优化函数

为了评估不同分支特征的重要性,以便充分学习细粒度伪造痕迹,设计了一种自适应联合损失优化函数,其包括 MultimarginLoss 函数[20]和交叉熵损失函数[21]。在训练初期,采用交叉熵损失函数,它能够有效处理类别的分布，引导模型学习真假语音的差异性纹理，实现模型训练的快速收敛。一旦模型趋向收敛状态,则切换到 MultimarginLoss 函数,并通过 margin 对模型的分类边界进行微调。MultimarginLoss 函数能够通过设置类别间的距离，帮助模型细化决策边界，提高对复杂伪造类型样本的分类。目前主流的伪造语音数据的数据样本往往分布不够均衡，因此采用 MultimarginLoss 进行微调也能够使模型对不同类型伪造语音特征进行更为全面的学习，缓解正负样本失衡带来的分类偏差问题。其中,交叉熵损失函数的计算公式如下,

$$Loss = -\frac{1}{S}\sum_{i=0}^{S-1}\sum_{t=0}^{1} y_{it}\ln(P_{it}) \quad (11)$$

在公式(11)中，S 表示样本数量, y_{it} 表示真实标签的类别, P_{it} 表示模型的预测类别, t 表示样本标签数。通过运用交叉熵损失函数来比较预测类别与真实标签之间的差异,可以使模型初步达到收敛的状态。其中 MultimarginLoss 函数的定义如下,

$$Loss = \frac{1}{S}\sum_{i=0}^{S-1}\sum_{t=0}^{1}\max(0, margin - P_{it} + y_{it})^{p} \quad (12)$$

在公式(12)中, P_{it} 代表模型的预测值, y_{it} 对应真实的标签类别, p 是计算损失时的指数值,用来约束对模型错误分类的惩罚力度,取值只能为 1 或 2, i 表示数据编号。p 值越大, 模型对错误分类所施加的惩罚力度也就越强。margin 是指在分类任务中, 正确分类得分与其他类别得分之间的最小间隔。当正确类别的得分超过一个阈值, 并且超过所有类别的得分之和时, 损失函数的取值为 0；反之, 则为正值。当 margin 较大时, 会增加标签值和其他类别间的间隔,能在一定程度上提高模型的预测能力。而当 margin 值过大时,也会使得模型产生过拟合,而若 margin 值较小,会缩小真实标签和预测错误值间的间隔,使得模型的决策边界更趋于平滑,泛化性能更强。但是若 margin 过小,则会导致样本学习不充分。为了评估不同分支特征的重要性,又提出了一种自适应联合损失优化函数 ALOF,通过调节不同分支网络在任务中的权重, 提升模型对不同特征伪造痕迹的学习能力。其公式如下,

$$Loss = \begin{cases} Loss_{front}, loss_{dev} \geq Threshold \\ Loss_{rear}, else \end{cases} \quad (13)$$

$$Loss_{front} = \max\begin{pmatrix} -\frac{1}{S}\sum_{i=0}^{S-1}\sum_{t=0}^{1} y_{it}\ln\left(P_{it}^{CQCC}\right), \\ -\frac{1}{S}\sum_{i=0}^{S-1}\sum_{t=0}^{1} y_{it}\ln\left(P_{it}^{LFCC}\right), \\ -\frac{1}{S}\sum_{i=0}^{S-1}\sum_{t=0}^{1} y_{it}\ln\left(P_{it}^{waveform}\right) \end{pmatrix} \quad (14)$$

$$Loss_{rear} = \max\begin{pmatrix} \frac{1}{S}\sum_{i=0}^{S-1}\sum_{t=0}^{1}\max(0, margin - P_{it}^{CQCC} + y_{it})^{p}, \\ \frac{1}{S}\sum_{i=0}^{S-1}\sum_{t=0}^{1}\max(0, margin - P_{it}^{LFCC} + y_{it})^{p}, \\ \frac{1}{S}\sum_{i=0}^{S-1}\sum_{t=0}^{1}\max(0, margin - P_{it}^{Waveform} + y_{it})^{p} \end{pmatrix} \quad (15)$$

有关 ALOF 的推理过程如公式(13)、(14)和(15)所示。其中在公式(13)中, $Loss$ 表示真伪语音分类的损失值, $Loss_{front}$ 和 $Loss_{rear}$ 分别为前期和后期训练阶段的真伪分类损失, $Loss_{dev}$ 则表示开发集的损失。$Threshold$ 是控制损失函数的阈值,将其设为 0.05。当 $Loss_{dev}$ 低于阈值时进入后期训练阶段, t 代表分类的类别数。在公式(14)、(15)中, y_{it} 表示真实标签的类别, S 表示样本数量, P_{it}^{CQCC} ， P_{it}^{LFCC} 和 $P_{it}^{Waveform}$ 分别表示不同分支下的预测类别, p 表示对不同类型分支进行调节的惩罚项。通过自适应地调节联合损失, 能够解决模型的过拟合问题，提高模型的检测精度。

4 实验和分析

4.1 数据集

为了构建一个共享的性能测试平台,推动学术界和产业界的研究人员在伪造语音检测领域的交流与合作,英国爱丁堡大学联合多个研究机构成功举办 ASVspoof 反欺骗挑战赛,至今已连续举办多届,吸引了包括中山大学在内的众多研究人员的参与。在每届挑战赛启动前,主办方都会发布一批最新的数据集,其涵盖语音合成、语音转换等多种攻击类型的数据。同时，为了便于不同算法之间的性能对比和优秀算法的筛选,主办方还会给出相应的评价指标。本文采用的数据集为主流的 ASVspoof 2019 逻辑访问(LA)数据集,其详细信息参见表 1。ASVspoof 2019 LA 数据集在无通道或噪声干扰的环境下,从语音克隆工具箱(Voice Cloning Toolkit, VCTK)语料库中收集了 46 名男性和 61 名女性的真实语音样本,并通过多种语音合成和语音转换技术生成伪造语音。该数据集分为训练集(Train)、开发集(Dev)和测试集(Eval)三部分,其中训练集(Train)和开发集(Dev)主要包含六种已知的伪造语音攻击,而测试集(Eval)则涵盖了 2 种已知攻击和 11 种未知攻击的伪造语音,以全面检验算法的性能[22]。ASVspoof2021 数据集包括两种类型：逻辑访问（Logical Access, LA）数据集和深度伪造（Deep Fake, DF）数据集。LA 数据集中的伪造语音是由 13 种不同的语音

合成和语音转换攻击算法生成,在生成过程中引入了编码和传输干扰,进一步增加了伪造的复杂性。相比之下，DF 数据集的伪造语音由 110 种不同的欺骗性伪造算法生成,并通过有损编码器来处理真实语音和伪造语音,以模拟语音失真场景。本文方法在 ASVspoof 的训练集中进行训练,在 2019 LA、2021 LA 以及 2021 DF 的测试集中进行测试[23]。

表 1 ASVspoof **数据集**

Table 1 ASVspoof dataset

-		真实语音	欺骗语音	
		数量	数量	欺骗攻击
ASVspoof2019 LA	训练集	2580	22800	A01 – A06 (6 种)
	开发集	2548	22296	A01 – A06 (6 种)
	测试集	7355	63882	A07 – A19 (13 种)
ASVspoof2021 LA	测试集	14816	133360	A07-A19 (13 种)
ASVspoof2021 DF	测试集	14869	519059	110 种

4.2 评价指标

为了确保对不同算法的性能评估的公正性和一致性,本文采用 7 个公开的评价标准用于性能的评估,包括等错误率、最小归一化串联检测代价函数、召回率（Recall）、F1 得分（F1-score）、检测误差权衡曲线（Detection Error Tradeoff, DET）和接受者操作特征曲线（Receiver Operating Characteristic, ROC）。此外,为了便于观察所提算法的分类效果,本文还采用 t 分布-随机邻近嵌入（t-Distributed Stochastic Neighbor Embedding, t-SNE）指标,对模型的输出结果进行可视化展示。EER（等错误率）是指在错误接受率与错误拒绝率相等时的特定数值,作为一个综合性指标,能够全面反映模型的检测性能。其计算公式为,

$$EER = P_{fa_cm} = P_{fr_cm} \tag{16}$$

其中,P_{fa_cm} 表示错误接受率,P_{fr_cm} 为错误拒绝率。最小归一化串联检测代价函数 min t-DCF 作为一种伪造检测的评价指标,能够更好地评估模型的性能。值越小,则表明方法性能愈佳。其计算公式如下,

$$\text{min t-DCF} = C_{fr} \cdot P_{fr_cm} + C_{fa} \cdot P_{fa_cm} \tag{17}$$

其中,C_{fr} 为错误拒绝率的风险系数,本文取值为 2.40595,C_{fa} 表示错误接受率的风险系数,取值为 1。min t-DCF 的值越小,说明算法的性能越好。精确率是指所有模型预测为正样本的样本数量中,模型预测为正样本且分类正确的样本数量的概率。值越大,表示算法的性能越好。其计算公式(18)为,

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

TP 代表实际为正样本且被模型正确预测为正样本的数目,*FP* 则指实际为负样本却被模型错误地预测为正样本的数目,而 *FN* 则是实际为正样本却被模型错误地预测为负样本的数目。召回率,作为评估指标之一,表示在所有真实正样本中,被模型正确预测为正样本的比例,其计算公式(19)如下,

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

虽然准确率和召回率是用于评估不同算法性能的重要指标,但其仍然不能全面反映算法的表现。例如,某算法可能在准确率方面表现出色,但在召回率上却表现不佳。为了解决这一问题,引入了 F1 得分作为一个综合评价指标,以平衡不同指标的影响,从而更全面地评估算法性能。F1 得分越高,代表着算法的性能越好。其计算公式(20)如下所示,

$$F1_{Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + recall} \tag{20}$$

此外,t-SNE 作为一种经典的降维可视化技术,能够充分考虑数据点之间的相对距离关系,从而将高维空间的模型分类结果映射至低维空间。这一过程不仅保留了数据的局部结构特征,还使得研究者能够直观地观察和理解模型的分类效果,为后续的模型优化提供有力的支持[24]。DET 和 ROC 曲线作为评估二元分类模型性能的指标,能够直观反映模型分类性能。其中,DET 曲线的横轴表示假正例率(False Positive Rate, FPR),即负样本被错误分类为正样本的概率[25]；纵轴表示假负例率(False Negative Rate, FNR),即正样本被错误分类为负样本的概率。DET 曲线越接近图像的左下角,其性能越好。曲线下的面积与模型性能成反比,面积越大,则方法性能越低。ROC 曲线的横轴表示假正例率[26],纵轴表示真正例率(True Positive Rate,TPR),即真实样本被正确分类为真实样本的概率。若曲线越靠近左上角,则表明方法性能越好。

4.3 实验环境与参数设置

本文所采用的网络模型是基于 Pytorch 1.7.1 深度学习框架,运行在 Ubuntu 20.04.5 LTS 操作系统上。硬件配置包括 12th Gen Intel(R) Core(TM) i5-12400 处理器以及 2 块 NVIDIA Tesla P40 24GB 显卡。在骨干网络的训练过程中,采用 Adam 优化器进行参数的优化,学习率设置为 0.0001,并将反向传播参数 lambda 设为 0.05。对于骨干网络中的 TCAMNet,频谱图的特征长度设置为 750,编码维度设置为 256，WaveformNet 的输入音频设置为 6 秒。

4.4 实验结果与分析

本文首先在 ASVspoof 数据集上进行了攻击成功率测试和泛化性测试,详细实验结果如表 2。

表 2 不同方法在 ASVspoof 数据集上的实验结果

Table 2 Experimental results of different methods on ASVspoof dataset

数据集	模型	EER (%)	min t-DCF
ASVspoof 2019 LA	ProsoSpeaker[27]	5.39	-
	CFE-ResNet[28]	5.23	-
	ARawNet2[29]	4.61	-
	Stat-SE-Res2Net50[30]	2.86	0.068
	Student Net-KA[31]	2.39	0.067
	OC-softmax[32]	2.19	0.059
	Spec+Multi-branch[33]	2.17	0.118
	Spec+LFCC+DenseNet[34]	1.98	0.047
	CQT+MCG-Res2Net50[35]	1.78	0.052
	Waveform+Raw PC-DARTS[36]	1.77	0.051
	MFCC+GTCC+SMOTE-LSTM[37]	1.60	-
	L-VQT+ DenseNet[38]	1.54	0.045
	LFCC+OCT[39]	1.06	0.034
	Fusion-PA-SE-ResNet[40]	0.96	0.030
	SpoTNet[41]	0.95	0.045
	Proposed Method	**0.35**	**0.011**
ASVspoof 2021 LA	CQCC-GMM[6]	15.62	0.4794
	LFCC-GMM[42]	19.30	0.5758
	LFCC-LCNN[43]	9.26	**0.3445**
	RawNet2[44]	9.50	0.4257
	Proposed Method	**8.94**	0.4063
ASVspoof 2021 DF	CQCC-GMM[6]	25.56	-
	LFCC-GMM[42]	25.25	-
	LFCC-LCNN[43]	23.48	-
	RawNet2[44]	22.38	-
	Proposed Method	**19.81**	-

通过与若干现有的方法对比可知,所提算法在 EER 和 min t-DCF 两个指标上均取得最佳性能。其中,与之前在原始波形上表现的方法 Waveform+Raw PC-DARTS[36]相比,所提方法在 EER 和 min t-DCF 两个指标上分别降低了 1.42%和 0.040。在检测精度方面,与当前最佳的频谱图特征方法 SpoTNet[41]对比后,所提方法在 EER 和 min t-DCF 两个指标上分别降低 0.60%和 0.034,均取得了 SOTA 性能。通过与现有工作相比,所提方法在特征丰富度表示上更为全面,能够削弱伪造语音检测任务对幅度和相位信息的严重依赖,伪造检测的性能也得到了显著的提升。此外,与 Yue 等人[40]提出的将频谱特征和频谱矩阵相融合方法相比,本文所提的方法性能更好,其中,EER 降低了 0.61%、min t-DCF 降低了 0.019。

此外，所提方法在 ASVspoof 2021 LA 和 DF 数据集上的表现同样出色, 展现了较强的泛化能力。在 ASVspoof 2021 LA 数据集上 EER 上实现了 8.94%, 显著优于现有的其他方法。与传统的 GMM 方法相比，所提方法的性能显著提升。而与 LFCC-LCNN[43]方法相比, 所提方法在 EER 上降低 0.32%, 在 min t-DCF 上略高, 说明本方法在特征信息挖掘方面更关注于伪造检测, 在一定程度上忽视了对说话人信息的区分。与此同时, 在 ASVspoof 2021 DF 数据集上, 所提方法的 EER 为 19.81%, 相比 CQCC-GMM[6]、LFCC-GMM[42]和 LFCC-LCNN[43]方法均表现出显著的性能提升。与 RawNet2[44]方法相比, 本文方法也取得了 2.57%的性能提升。结果表明, 所提方法不仅在 ASVspoof 2019 LA 上表现优异, 而且在其他数据集上同样表现出较好的跨数据集泛化性。

表 3 面对不同攻击类型时的 EER(%)和 min t-DCF

Table 3 EER (%) and min t-DCF for different attack types

模型	OCSoftmax[32]		Inc-TDSSDNET[45]		Res-TDSSDNET[45]		Proposed Method	
攻击类型	EER(%)	min t-DCF	EER(%)	min t-DCF	EER(%)	min t-DCF	EER(%)	min t-DCF
A07	0.1222	0.0034	0.6927	0.0195	1.4430	0.0395	**0.0985**	**0.0020**
A08	**0.1800**	**0.0042**	16.5625	0.3004	0.7504	0.0206	0.2037	0.0054
A09	0.1222	0.0094	0.4312	0.0341	0.0407	0.0025	**0.0000**	**0.0000**
A10	1.1408	0.0244	0.8319	0.0242	1.7860	0.0493	**0.2207**	**0.0064**
A11	0.1222	0.0025	0.6689	0.0191	0.0645	0.0015	**0.0170**	**0.0005**
A12	0.4652	0.0107	0.7096	0.0199	0.1800	0.0043	**0.0985**	**0.0018**
A13	0.2210	0.0061	0.1630	0.0041	0.0645	0.0015	**0.0170**	**0.0005**
A14	0.6927	0.0187	0.3260	0.0088	0.1392	0.0032	**0.0238**	**0.0005**
A15	1.1023	0.0376	0.7979	0.0219	2.0949	0.0541	**0.1392**	**0.0039**
A16	0.3260	0.0074	1.1646	0.0278	1.2461	0.0276	**0.0985**	**0.0024**
A17	9.2184	0.6158	12.6171	0.7860	6.0064	0.3254	**1.1986**	**0.1201**
A18	0.8964	0.0378	1.7282	0.1119	1.1171	0.0642	**0.0985**	**0.0092**
A19	0.8964	0.0317	2.0780	0.0793	1.4430	0.0511	**0.1222**	**0.0045**
Total	2.1788	0.0558	4.0431	0.0976	1.6421	0.0481	**0.3536**	**0.0105**

为了验证本文所提方法的泛化能力,还对 ASVspoof 2019 LA 中未知类型的欺骗攻击进行了性能测试,实验结果见表 3。通过观察结果可知,在面对目前最具挑战性的 A17 攻击时,所提方法在 EER 指标的评估中仍表现出色。A17 攻击采用语音波形滤波技术来处理伪造语音,通过对原始语音进行时域处理,实现伪造语音中音色的变化和噪声的降低,使伪造痕迹难以被提取。本文通过对语音波形的时域信息进行精细处理,成功提取了时域共性特征,显著提升了对经波形滤波操作生成的伪造语音的检测能力。与 Hua 等人提出的 Res-TDSSDNET[45]方法对比,本文方法在 A17 类型攻击检测的 EER 指标上实现了 4.8078%的降低,展现了更高的检测精度和优越性。然而,当面对 A08 这类攻击时,本文提出的算法的 EER 稍逊于现有方法。对数据集进行内容分析后发现,A08 攻击是一种基于神经网络的语音伪造技术,因此在语音伪造过程中,会在语音波形上留下一些不自然的痕迹。而本文所提算法与原始语音在时域上的波形变化密切相关,这种依赖性会严重影响原始任务的检测性能[22]。此外,本文还进行了 min t-DCF 的测试,实验结果如表 3 所示。观察结

果发现,本文提出的方法在应对大部分未知攻击时仍表现出较强的泛化性能。相较于 Res-TDSSDNET[45], 本文所提的方法在 A17 类型攻击检测的 min t-DCF 降低了 0.2053,然而在 A08 类型攻击下,min t-DCF 的效果稍显不佳。

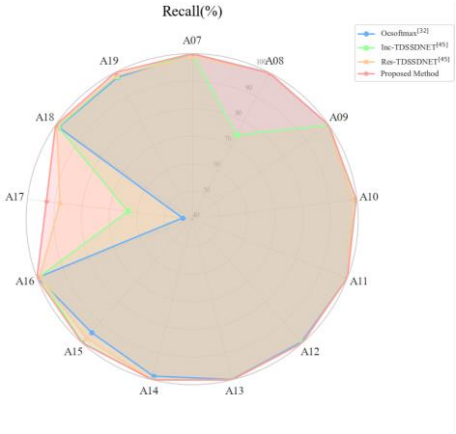


图 6 对 A07-A19 的召回率(%)

Fig.6 Recall rate of A07-A19 (%)

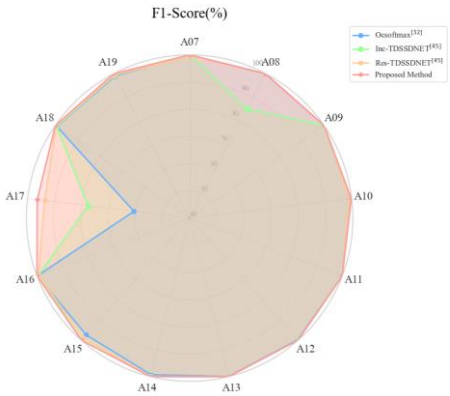
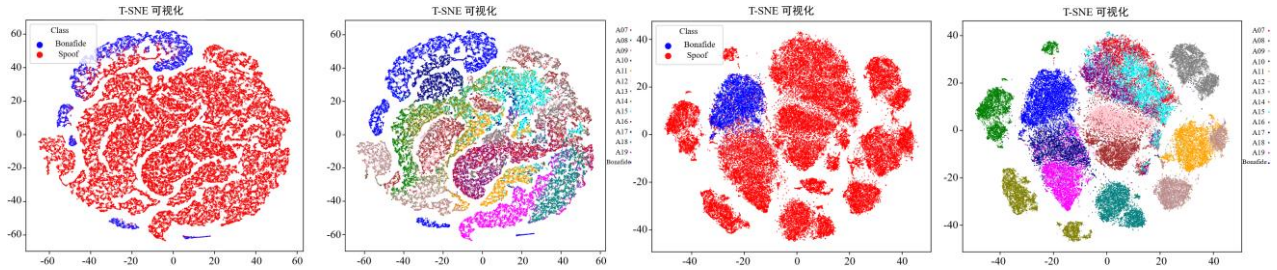


图 7 对 A07-A19 的 F1 得分(%)

Fig.7 F1-score(%) of A07-A19

鉴于 ASVspoof 2019 LA 数据集中真实语音与伪造语音的样本不平衡问题, 本文选用召回率和 F1 得分作为评价指标,以评估算法在应对多种未知类型攻击时的泛化能力。在测试中,将欺骗语音标记为正样本, 真实语音则被视为负样本,通过观察图 6 和图 7 可知,本文方法在召回率和 F1 得分上的表现均优于基准方法, 进一步表明本文算法具有较好的泛化性。



(a) (b) (c) (d)

图 8 在 ASVspoof 2019 LA Eval 数据集上的 T-SNE 可视化 其中(a) OCSoftmax[32], (b) OCSoftmax[32]在不同攻击下的可视化, (c) 所提方法, (d) 所提方法在不同攻击下的可视化。

Fig.8 T-SNE visualization on the ASVspoof 2019 LA Eval dataset in which (a) OCSoftmax[32], (b) OCSoftmax[32] visualization under different attacks, (c) the proposed method, and (d) the proposed method visualization under different attacks.

为了便于观察模型的分类效果,在图 8(a)和图 8(c)中,分别用蓝色和红色标识真实语音（Bonafide）和欺骗语音（Spoof）, 可发现本文所提方法对真实语音和欺骗语音展现出更高效的聚类能力。进一步分析图 8(b)和 8(d),蓝色代表真实语音, 而其他不同颜色则分别对应不同类型的攻击。通过对比图 8(b)和 8(d), 可以清晰地看到,相较于方法 OCSoftmax[32], 本文方法在聚类不同类型攻击时表现出了更高的准确性,从而验证了其在语音伪造检测领域的优越性。此外,在 DET 指标测评中,本文所提的方法表现出更佳的性能,如图 9。左下角的覆盖面积低于其他三种方法, 但假负例率略高于 Res-TDSSDNET[45], 表明存在部分欺骗语音被识别为真实语音的情况。在 ROC 指标测试中, 如图 10 所示,本文所提的方法在指标 AUC 达到了 99.99%, 超越其余三种方法, 进一步证明了本文方法的有效性。

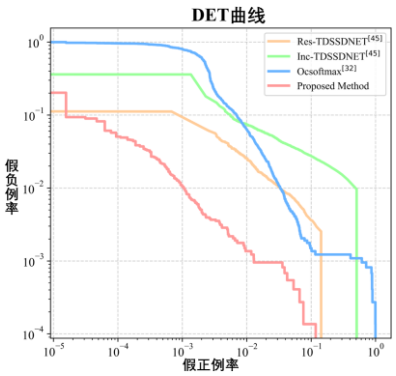


图 9 DET 曲线

Fig.9 DET curve

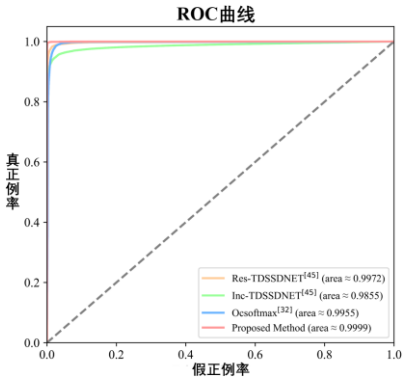


图 10 ROC 曲线

Fig.10 ROC curve

为了验证不同参数对算法性能的影响,还进行了消融实验,实验结果如表 4 所示。

表 4 消融实验

Table 4 Ablation experiments

模型	EER (%)	t-DCF
Signal-Spectrum-branch (CQCC)	3.41	0.103
Signal-Spectrum-branch (LFCC)	3.70	0.097
Signal-Waveform-branch w/o Dual-Spectrum-branch	2.14	0.062
Dual-Spectrum -branch w/o Waveform-branch, w/o TCAM	1.82	0.062
Dual-Spectrum -branch w/o Waveform-branch, w/ TCAM	1.37	0.040
Multi-Feature Fusion w/o TCAM, w/o SAL	0.70	0.022
Multi-Feature Fusion w/o TCAM, w/ SAL	0.58	0.018
Multi-Feature Fusion w/ TCAM, w/o SAL	0.50	0.015
Proposed Method (Multi-Feature Fusion w/ TCAM, w/ SAL)	**0.35**	**0.011**

相较于单频谱图,双频谱图在 EER 方面表现出 1.82%的较低数值, 同时在 min t-DCF 指标上取得了更佳的效果。因此,采用双频谱图作为验证 TCAM 模块是有效的。当进行双频谱图的对比时, 引入了 TCAM 模块后, EER 的值下降了 0.45%。对于引入 TCAM 模块的多特征融合方法, 能够从时频角度挖掘语音中的伪造痕迹, EER 达到了 0.50%。同时, 将 SAL 模块引入多特征融合方法后, 提高了数据多样性, 加强了模型对未知伪造语音能量分布的学习, 使得 EER 达到 0.58%。将 TCAM 和 SAL 引入多特征融合方法中后, EER 为 0.35%, min t-DCF 为 0.011,再次证明本文所提算法的高效性。

5 总结

语音识别作为一种身份认证技术,目前已被广泛应用于各类身份认证场景。然而,随着深度伪造技术的发展,伪造语音被用来尝试欺骗各类语音识别系统,引发了一系列严重的隐私泄露、非授权访问等安全纠纷,因此对待认证语音的真伪进行鉴别至关重要。现有的伪造语音检测方法,往往基于单一特征,难以捕捉伪造语音中的语义连贯性,且过度依赖时域的相位信息,致使检测精度和泛化能力受限。因此,本文提出了一种多分支特征融合网络,分别从频率、共振峰以及时域等角度挖掘伪造语音中的能量异常痕迹及语义不连贯性细微差别,以更好地表征真伪语音的音高变化、峰值差异和波形变化。然后,进行时频伪造定位,在时域上聚焦局部伪造特性,且在频域上捕捉伪造语音中的偏移或异常频率,并对时域和频域上的空间信息进行联合编码,以便更好地关注伪造区域。实验结果表明,本文所提方法在伪造语音检测和未知攻击识别方面均取得了显著的优势。

在接下来的工作中,我们还将探索更高效的特征融合和伪造语音片段的定位方法,以进一步解决伪造语

音检测任务中面临的可解释难的问题。

参 考 文 献

[1] 李云峰,闫祖龙,高天,等.基于多任务学习的语音情感识别[J].数据采集与处理,2024,39(02):424-432.
LI Yunfeng, YAN Zulong, GAO Tian, et al. Speech Emotion Recognition with Multi-task Learning [J]. Journal of Data Acquisition and Processing, 2024, 39 (02): 424-432.

[2] 康瑶,康坊,杨飞然.利用互子带滤波器和稀疏特性的多通道线性预测语音去混响方法[J].数据采集与处理,2024,39(05):1135-1146.
KANG Yao, KANG Fang, YANG Feiran. Multi-channel Linear Prediction for Speech Dereverberation Using Cross-Band Filters and Sparse Priors [J]. Journal of Data Acquisition and Processing, 2024, 39 (05): 1135-1146.

[3] PATEL T B and PATIL H A. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech [C]// Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: the International Speech Communication Association, 2015: 2062-2066.

[4] KORSHUNOV P, MARCEL S, MUCKENHIRN H, et al. Overview of BTAS 2016 speaker anti-spoofing competition [C]// Proceedings of 2016 IEEE 8th international conference on biometrics theory, applications and systems. Niagara Falls, NY, USA: IEEE, 2016: 1-6.

[5] SAHIDULLAH M, KINNUNEN T, and HANILÇI C. A comparison of features for synthetic speech detection [C]// Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: the International Speech Communication Association, 2015: 2087-2091.

[6] TODISCO M, DELGADO H, and EVANS N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification [J]. Computer Speech & Language, 2017, 45: 516-535.

[7] QIAN Yanmin, CHEN Nanxin and YU Kai. Deep features for automatic spoofing detection [J]. Speech Communication, 2016, 85: 43-52.

[8] LAVRENTYEVA G, NOVOSELOV S, MALYKH E, et al. Audio replay attack detection with deep learning frameworks [C]// Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: the International Speech Communication Association, 2017: 82-86.

[9] LAI Cheng-I, CHEN Nanxin, VIllalba J, et al. ASSERT: Anti-spoofing with squeeze-excitation and residual networks[EB/OL]. https://arxiv.org/pdf/1904.01120, 2019-04-01.

[10] GS A S, GANESHKUMAR V, THIRUMAVALAVAN V C, et al. Synthetic Speech Classification using Bidirectional LSTM Networks [C]// Proceedings of 2022 IEEE 3rd Global Conference for Advancement in Technology. Bangalore, India: IEEE, 2022: 1-6.

[11] GOMEZ-ALANIS A, PEINADO A M, GONZALEZ J A et al. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection [C]// Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: the International Speech Communication Association, 2019: 1068-1072.

[12] JUNG J W, KIM S B, SHIM H J, et al. Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms [C]// Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: the International Speech Communication Association, 2020: 1496-1500.

[13] TAK H, JUNG J, PATINO J, et al. Graph attention networks for anti-spoofing [EB/OL]. https://arxiv.org/pdf/2104.03654, 2021-04-08.

[14] WANG Zhenyu, HANSEN J H L. Audio anti-spoofing using a simple attention module and joint optimization based on additive angular margin loss and meta-learning [EB/OL]. https://arxiv.org/pdf/2211.09898, 2022-11-17.

[15] WANG Chenglong, YI Jiangyan, TAO Jianhua, et al. TO-Rawnet: improving Rawnet with TCN and orthogonal regularization for fake audio detection [EB/OL]. https://arxiv.org/pdf/2305.13701, 2023-05-23.

[16] XUE Junxiao, ZHOU Hao. Physiological-physical feature fusion for automatic voice spoofing detection [J]. Frontiers of Computer Science, 2023, 17(2): 172318.

[17] WANG Chenglong, YI Jiangyan, TAO Jianhua, et al. Detection of cross-dataset fake audio based on prosodic and pronunciation features [EB/OL]. https://arxiv.org/pdf/2305.13700, 2023-05-23.

[18] LIU Rui, ZHANG Jinhua, GAO Guanglai, et al. Betray Oneself: a novel audio deepfake detection model via mono-to-stereo conversion [EB/OL]. https://arxiv.org/pdf/2305.16353v1, 2023-05-25.

[19] HOU Qibin, ZHOU Daquan, and FENG Jiashi. Coordinate attention for efficient mobile network design [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville, TN, USA: IEEE, 2021: 13713-13722.

[20] PyTorch Contributors. Multi Margin Loss [EB/OL]. https://pytorch.org/docs/stable/generated/torch.nn.MultiMarginLoss.html, 2023-12-01.

[21] PyTorch Contributors. Cross Entropy Loss [EB/OL]. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html, 2023-12-01.

[22] WANG Xin, 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.

[23] LIU Xuechen, WANG Xin, 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.

[24] CAI T T and MA Rong. Theoretical foundations of t-sne for visualizing high-dimensional clustered data [J]. The Journal of Machine Learning Research, 2022, 23(301):1-54.

[25] ADLER A and SCHUCKERS M E. Calculation of a composite DET curve [C]// Proceedings of Audio-and Video-Based Biometric Person Authentication: 5th International Conference. Berlin, Germany: Springer, 2005: 860-868.

[26] FAWCETT T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[27] ATTORRESI L, SALVI D, BORRELLI C, et al. Combining automatic speaker verification and prosody analysis for synthetic speech detection [C]// Proceedings of International Conference on Pattern Recognition. Montréal Québec, Canada: Springer, 2022: 247–263.

[28] ZHANG Jinghong, YI Xiaowei and ZHAO Xianfeng. A Compressed Synthetic Speech Detection Method with Compression Feature Embedding [C]// Proceedings of the 24th Annual Conference of the International Speech Communication Association. Dublin, Ireland: The International Speech Communication Association, 2023: 5376–5380.

[29] LI Jing, LONG Yanhua, LI Yijie, et al. Advanced RawNet2 with Attention-based Channel Masking for Synthetic Speech Detection [C]// Proceedings of the 24th Annual Conference of the International Speech Communication Association. Dublin, Ireland: the International Speech Communication Association, 2023: 2788–2792.

[30] LI Xu, LI Na, WENG Chao, et al. Replay and synthetic speech detection with res2net architecture [C]// Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Ontario, Canada: IEEE, 2021: 6354-6358.

[31] REN Yeqing, PENG Haipeng, Li Lixiang, et al. Generalized Voice Spoofing Detection via Integral Knowledge Amalgamation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2461-2475.

[32] ZHANG You, JIANG Fei and DUAN Zhiyao. One-class learning towards synthetic voice spoofing detection [J]. IEEE Signal Processing Letters, 2021, 28: 937-941.

[33] WANG Ruoyu, DU Jun, and WANG Chang. Multi-branch Network with Circle Loss Using Voice Conversion and Channel Robust Data Augmentation for Synthetic Speech Detection [C]// Proceedings of Chinese Conference on Biometric Recognition. Beijing, China: Springer, 2022: 613-620.

[34] WANG Zheng, CUI Sanshuai, KANG Xiangui, et al. Densely connected convolutional network for audio spoofing detection [C]// Proceedings of 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Auckland, New Zealand: IEEE, 2020: 1352-1360.

[35] LI Xu, WU Xixin, LU Hui, et al. Channel-wise gated res2net: Towards robust detection of synthetic speech attacks [C]// Proceedings of the 22th Annual Conference of the International Speech Communication Association. Brno, Czech Republic: the International Speech Communication Association, 2021: 4314-4318.

[36] GE W, PATINO J, TODISCO M, et al. Raw differentiable architecture search for speech deepfake and spoofing detection[EB/OL]. https://arxiv.org/pdf/2107.12212, 2021-10-06.

[37] CHAKRAVARTY N and DUA M. Data augmentation and hybrid feature amalgamation to detect audio deep fake attacks [J]. Physica Scripta, 2023, 98(9): 096001.

[38] LI Jialong, WANG Hongxia, HE Peisong, et al. Long-term variable Q transform: A novel time-frequency transform algorithm for synthetic speech detection [J]. Digital Signal Processing, 2022, 120: 103256.

[39] LI Changtao, YANG Feiran and YANG Jun. The Role of Long-Term Dependency in Synthetic Speech Detection [J]. IEEE Signal Processing Letters, 2022, 29: 1142-1146.

[40] YUE Feng, CHEN Jiale, SU Zhaopin, et al. Audio Spoofing Detection Using Constant-Q Spectral Sketches and Parallel-Attention SE-ResNet [C]// Proceedings of Computer Security–ESORICS 2022: 27th European Symposium on Research in Computer Security. Copenhagen, Denmark: Springer, 2022: 756-762.

[41] KHAN A and MALIK K M. SpoTNet: A spoofing-aware Transformer Network for Effective Synthetic Speech Detection [C]// Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation. Thessaloniki, Greece: Association for Computing Machinery, 2023: 10-18.

[42] TAK H, PATINO J, NAUTSCH A, et al. Spoofing attack detection using the non-linear fusion of sub-band classifiers [EB/OL]. https://arxiv.org/pdf/2005.10393, 2020-05-20.

[43] WANG Xin, YAMAGISHI J. A comparative study on recent neural spoofing countermeasures for synthetic speech detection [EB/OL]. https://arxiv.org/pdf/2103.11326, 2021-06-13.

[44] TAK H, PATINO J, TODISCO M, et al. End-to-end anti-spoofing with rawnet2 [C]// Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Ontario, Canada: IEEE, 2021: 6369-6373.

[45] HUA Guang, TEOH A B J, and ZHANG Haijian. Towards end-to-end synthetic speech detection[J]. IEEE Signal Processing Letters, 2021, 28: 1265-1269.

作者简介：

袁程胜(1989-) 男, 博士, 副教授, 研究方向: 信息隐藏、多媒体内容安全等。
Email: yuancs@nuist.edu.cn

张雪原(2000-) 男, 硕士研究生, 研究方向: 人工智能安全。

周志立(1982-) 男, 博士, 教授, 研究方向: 信息安全、区块链安全等。

李欣亭(1991-) **通信作者**, 女, 博士, 副教授, 研究方向: 数据挖掘、情报分析等。
Email: lixt@tju.edu.cn

付章杰(1983-) 男,博士, 教授, 研究方向: 信息安全、区块链安全等。