2025 年 第 25 卷 第 22 期
2025, 25(22): 09427-09

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2409674
引用格式:冯嘉琪, 王华朋, 刘天赐. 融合通道-时间注意力和深度可分离卷积的欺骗语音检测[J]. 科学技术与工程, 2025, 25(22): 9427-9435.
Feng Jiaqi, Wang Huapeng, Liu Tianci. Spoof speech detection with channel-temporal attention and depthwise separable convolutions[J]. Science Technology and Engineering, 2025, 25(22): 9427-9435.

融合通道-时间注意力和深度可分离卷积的欺骗语音检测

冯嘉琪, 王华朋*, 刘天赐
(中国刑事警察学院公安信息技术与情报学院, 沈阳 110854)

摘 要 自动说话人验证系统在应对日益逼真的深度伪造语音时,面临显著的欺骗攻击威胁。现有基于卷积神经网络的反欺骗模型在捕捉全局特征与应对未知类型语音伪造的泛化性能方面存在不足。为提升反欺骗检测效果,提出了一种融合通道-时间注意力机制与深度可分离卷积的网络模型 CT-DSCNet。该模型在 RawNet2 基础上引入通道-时间注意力模块,增强对重要语音特征的关注,减少无关区域的干扰;同时采用深度可分离卷积残差块,优化计算效率与模型实时性。实验在 ASVspoof2019、ASVspoof2021 和 FMFCC-A 数据集上进行,结果显示 CT-DSCNet 在 ASVspoof2019 LA 测试集上的等错误率(equal error rate,EER)达到 1.53%,较基线模型降低 70.58%。在泛化能力方面相较其他模型也表现出色,在 FMFCC-A 评估集上的 EER,较改进前模型相比提高了 25.35%。实验验证了该方法在提升伪造语音检测性能和跨数据集适应性方面的有效性。

关键词 深度伪造语音; 注意力机制; 深度可分离卷积; 语音反欺骗
中图法分类号 TP391; **文献标志码** A

Spoof Speech Detection with Channel-temporal Attention and Depthwise Separable Convolutions

FENG Jia-qi, WANG Hua-peng*, LIU Tian-ci
(College of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang 110854, China)

[**Abstract**] The growing sophistication of deepfake speech poses significant security threats to ASV(automatic speaker verification) systems. Current anti-spoofing models based on CNNs(convolutional neural networks) are constrained by inadequate global feature extraction and limited generalization capability against unseen spoofing attacks. To address these challenges, a novel network architecture integrating CT-DSCNet(channel-temporal attention mechanisms with depthwise separable convolutions) was proposed. Building upon the RawNet2 framework, the developed model incorporates dual-domain attention modules to enhance discriminative feature representation while suppressing irrelevant acoustic artifacts. Furthermore, depthwise separable convolutional residual blocks were strategically implemented to optimize computational efficiency and real-time processing capabilities. Comprehensive evaluations were conducted across three benchmark datasets: ASVspoof2019 LA, ASVspoof2021 DF, and FMFCC-A. Experimental results demonstrate state-of-the-art performance with EER(equal error rate) of 1.53% on ASVspoof2019 LA, representing a 70.58% relative improvement over baseline systems. Notably, the proposed architecture exhibits superior cross-dataset generalization, achieving a 25.35% lower EER on the FMFCC-A evaluation set compared with conventional approaches. These findings validate the effectiveness of the hybrid attention-convolution design in advancing spoofing detection robustness and domain adaptability.

[**Keywords**] deepfake speech; attention mechanism; depthwise separable convolution; speech anti-spoofing

自动说话人验证系统(automatic speaker verification, ASV)[1]能够对输入语音数据中的声学特征和说话人特征进行分析,自动识别和验证说话人身份。随着深度学习与人工智能的不断发展,文本到语音合成(text to speech, TTS)和语音转换(voice conversion, VC)技术能够生成高度逼真的伪造语

收稿日期: 2024-12-29; **修订日期**: 2025-05-19
基金项目: 国家重点研发计划(2017YFC0821000);司法部司法鉴定重点实验室(司法鉴定科学研究院,KF202117);中国刑事警察学院研究生创新能力提升项目(2024YCZD05)
第一作者: 冯嘉琪(2001—),女,汉族,河南新乡人,硕士研究生。研究方向:深度学习、语音检验。E-mail:18240668287@163.com。
*通信作者: 王华朋(1979—),男,汉族,山东菏泽人,博士,教授。研究方向:说话人识别、深度学习、人工智能。E-mail:huapeng.wang@hotmail.com。

音,严重威胁 ASV 系统的安全性[2-3]。为应对这一挑战,近年来的研究尤其关注深度伪造语音(特别是合成语音)的检测,其中 ASVspoof 挑战系列[2-5]提供了通用评估规则、性能指标和数据集,对该领域的发展具有重要推动作用。

传统的反欺骗说话人验证系统由前端和后端两部分组成[6]。前端负责提取声学特征,如常数 Q 倒谱系数(constant-Q cepstral coefficients, CQCC)[7]、线性频率倒谱系数(linear frequency cepstral coefficients, LFCC)[8]、梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)[9]等。后端则直接使用高斯混合模型(Gaussian mixture model, GMM)或支持向量机(support vector machine, SVM)等分类器对语音真实性进行判断。然而,传统方法在面对复杂多样的伪造语音时表现出局限性。

深度学习的进一步发展使基于深度神经网络的反欺骗系统逐渐成为主流。当前的语音反欺骗模型大多以卷积神经网络(convolutional neural network, CNN)为主,常用的经典卷积模型有 LCNN(light convolutional neural network)[10]、AASIST(audio anti-spoofing using integrated spectro-temporal graph attention network)[11]等;还有 ResNet(residual network)[12]及其扩展的说话人特征提取网络如 ResNeXt、Res2Net[13]等,近年来也展现出巨大的潜力;图注意力网络(graph attention network, GAT)[14-15]则通过利用图注意力捕获跨越频率和时间位置的判别线索,这些模型在各种挑战赛中都有较好的表现。

尽管近年来基于深度学习的语音反欺骗检测方法展现了较好的效果,但依然存在诸多问题亟待解决。一方面,现有模型大多依赖于大量标注数据,面对数据集分布差异和未见类型的伪造语音时,往往表现出较差的泛化能力。另一方面,传统 CNN 模型因卷积和池化操作的局限性[16],更多关注局部的时频信息,难以充分捕获伪造语音中的全局依赖特征。此外,部分引入复杂机制的模型尽管提升了性能,但代价是显著增加了计算复杂度和模型参数量,影响了实际应用的实时性和效率。

在上述背景下,现以 RawNet2[17]系统为基线系统,融合通道-时间注意力机制,对输入的原始波形特征在通道和时维度上的注意力机制进行加权,增强模型对局部关键特征的关注能力,减少时频信息的损失。在 RawNet2 网络结构中引入通道-时间注意力机制,通过对原始波形特征在通道和时间维度上的依赖信息进行动态加权,增强模型对局部关键特征的关注能力,弥补传统网络难以捕获全局依赖的不足。在残差模块中引入深度可分离卷积,以高效提取局部模式和特征,显著降低模型参数量和计算复杂度,同时保持模型的高效性和准确性。设计跨数据集对比实验,验证本文方法在不同类型伪造语音上的检测性能,尤其是面对未知伪造语音时的泛化能力,并对中文数据集进行专项评估,展示模型在多语言场景下的适用性和优势。

1 基线模型与关键技术

1.1 Rawnet2 系统

RawNet2[17]是一种用于从原始音频波形中直接学习高级特征图(high-level feature map, HFM)的端到端神经网络架构,被用作 ASVspoof2019 和 ASVspoof2021 挑战赛的官方基线系统,主要由帧级特征提取器和分类器两部分组成,其结构如图 1 所示。首先,给定原始波形作为输入,SinConv 层使用一组参数化的 sinc 函数对波形进行卷积,生成 128 个梅尔频率带通滤波器,使网络能够关注对滤波器形状和带宽具有广泛影响的高层可调参数。接下来从 SinConv 层提取的局部声学特征会被送入 6 个残差块,以提取帧级说话人表示。此外,为了获得更具辨别性的说话人信息,在每个残差块的输出执行特征图的过滤缩放。具有 1 024 个隐藏节点的 GRU 层用于将帧级表示聚合为单一的话语级表示。最后,GRU 的输出经过两个全连接层和一个 softmax 激活函数的分类器,以预测输入的语音是真实的还是伪造的。

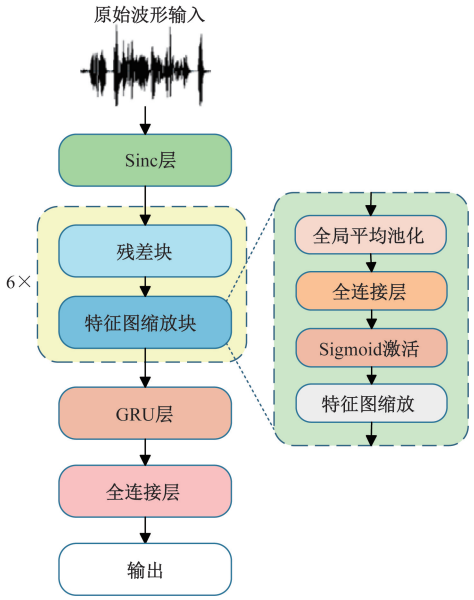


图 1 RawNet2 网络结构图

Fig. 1 RawNet2 network architecture

1.2　注意力机制

注意力机制(attention machanism)源自人对外部信息的处理能力,最早起源于图像视觉领域,通过对不同部分赋予不同的权重,这样可以使模型更关注输入序列中的关键信息,从而提高模型的效率和精度。

近年来,注意力机制被广泛应用于语音反欺骗检测任务,显著提升了检测性能。Zhang 等[18]引入 SE(squeeze and excitation)模块,旨在捕获信道维度上的全局关联和关键区域。Liu 等[19]的 Rawformer 模型则利用了与位置紧密相关的局部和全局依赖性,以识别合成语音。另一方面,Tak 等[14]提出名为 RawGAT-ST 的端到端时频图注意力网络,利用图注意力网络(GAT)的机制捕捉覆盖整个频率带或时间段的欺骗伪影特征。Ta 等[20]通过结合卷积操作和自注意力机制,有效地提取了语句中的局部细节和全局交互特性。此外,其他多种注意力机制方法也得到了广泛的应用,然而这些方法大多仅从单一维度来分析注意力机制[21]。

1.3　深度可分离卷积

深度可分离卷神经网络(depthwise separable convolutional neural network, DSCNN)[22-23]被学者们提出作为标准 CNN 的有效替代方案。标准 CNN 利用若干个多通道卷积核对输入的多通道图像进行处理,同时提取通道特征和空间特征。DSCNN 把传统卷积操作拆分为两个独立的步骤,深度卷积和逐点卷积。第一步是深度卷积,对每个通道独立地使用一个小卷积核(通常是 3×3 或 5×5)进行卷积,这一步只在空间维度上执行卷积操作,不会在通道之间进行混合,从而减少了计算成本。第二步是点卷积,是一个 1×1 的卷积,用来混合不同通道的信息,通过这个过程可以重新调整通道数,增加或减少特征维度。它能够将深度卷积输出的特征融合到新通道上,使网络获得跨通道的信息。当用深度可分离卷积来代替传统的 3×3×3 卷积核时,其框架如图 2 所示。

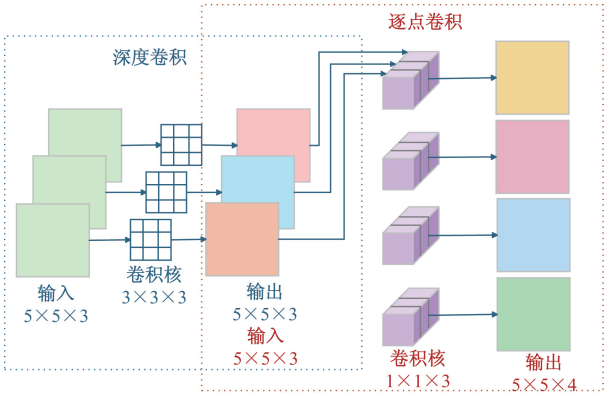


图 2　深度可分离卷积示意图

Fig. 2　Schematic of depthwise separable convolution

假设深度卷积的卷积核尺寸为 $D_K \times D_K \times 1$,其中 D_K 为卷积核的空间尺寸(如:3×3 的卷积核,则 $D_K = 3$)卷积核个数为 M,每个要做 $D_F \times D_F$ 次乘加运算,其中 D_F 表示特征图的空间尺寸。逐点卷积的卷积核尺寸为 $1 \times 1 \times M$,卷积核个数为 N,每个要做 $D_F \times D_F$ 次乘加运算,则深度可分离卷积与标准卷积的参数量比值为

$$\frac{D_K D_K M + MN}{D_K D_K MN} = \frac{1}{N} + \frac{1}{D_K^2} \tag{1}$$

计算量比值为

$$\frac{D_K D_K M D_F D_F + MN D_F D_F}{D_K D_K MN D_F D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{2}$$

一般来说,N 比较大,$1/N$ 可忽略不计,可以看出,使用深度可分离卷积后参数量和计算量可以下降到原来的 $1/D_K^2$ 左右,大大减少计算压力。DSCNN 有更少的参数、更快的速度、更加易于移植和更加精简的模型这些显著优势。

2　本文模型

2.1　总体网络结构

基线系统直接从原始波形中提取特征,为减轻其对时间和频率局部信息造成的信息丢失,本文研究在 RawNet2 网络中输出所有残差块后删除所有特征图缩放操作;在最后一个残差块和 GRU 层之间引入通道-时间注意力机制模块,学习语音特征在时频信息中的局部显著性特征,减少对不重要区域的关注,对不同维度的注意力特征进行动态权重计算并进行融合;同时将残差块中的所有卷积层替换为深度可分离卷积层,减轻计算负担,更好地捕捉到局部和全局的信息。本文提出的 CT-DSCNet 模型整体结构如图 3 所示。详细网络参数如表 1 所示。

表 1　CT-DSCNet 网络结构

Table 1　CT-DSCNet network structure

模块名称	内容	输出参数
SinConv	一维卷积 + BN + SeLU	(8,1 28, 64 472)
深度可分离卷积残差块(5)	BN + 深度可分离卷积 + SELU + BN +	(8, 512, 16 118)
通道-时间注意力	通道注意力块 + 时间注意力块 + 动态权重	(8, 512, 16 118)
GRU	GRU(1 024)	(8, 16 118, 1 024)
全连接层	1 024	(8, 16 118, 512)
全局池化	—	(8,512)
输出	—	(8, 2)

注:—表示单一层级。

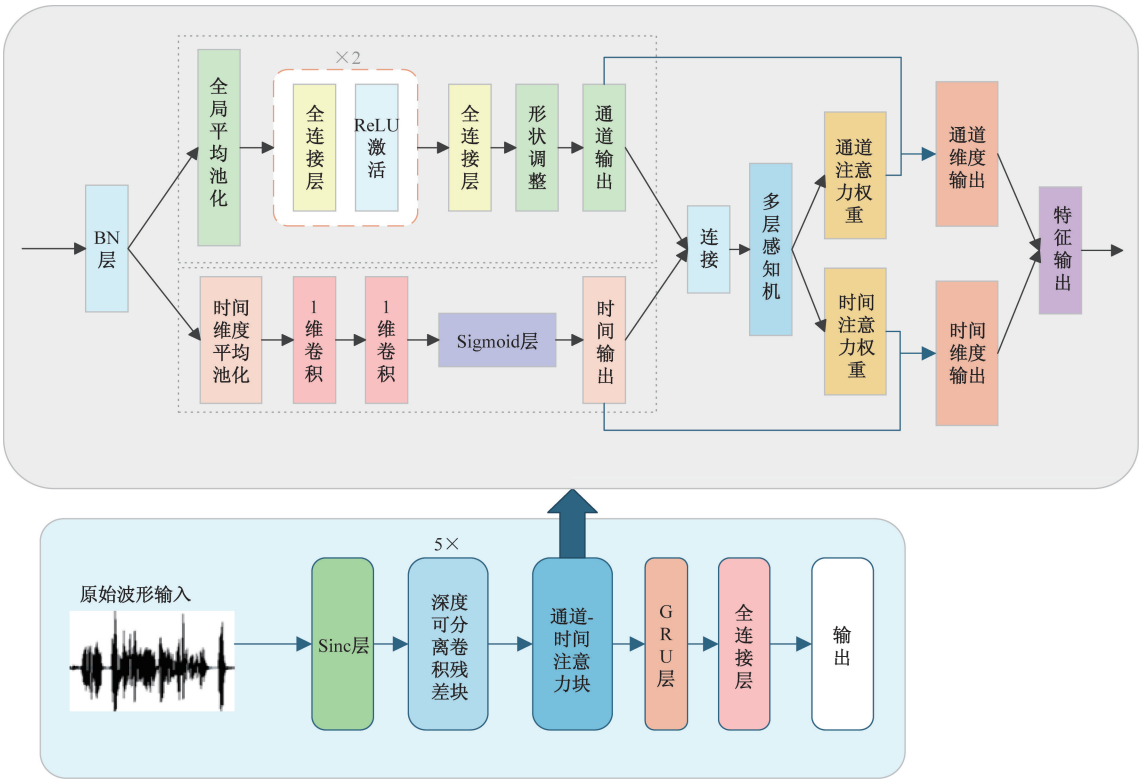


图 3 总体结构图

Fig. 3 General structure chart

2.2 通道-时间注意力机制模块

相比传统的单维度注意力机制,本文研究在通道和时间维度上联合建模,显著增强了模型对伪造语音中局部关键特征(如不自然的重音、断续的频谱异常)和全局过渡特性的关注能力,所提出的注意力模块整体结构如图 3 灰色部分所示。全局特征信息中的频谱子带能够准确地区分真实或欺骗的音频,使用全局平均池化层可以帮助模型捕获全局信息。具体来说,本文提出的注意力机制模块集成了通道注意力和时间注意力两个机制,并行处理不同维度的特征,旨在全面捕捉音频信号中的重要信息。

通道注意力机制通过自适应调整通道特征的权重,强化模型对重要特征的关注。通道注意力权重 A_c 表达式为

$$A_c = \sigma\{W_3\mathrm{ReLU}[W_2\mathrm{ReLU}(W_1\boldsymbol{y})]\} \tag{3}$$

式(3)中:$\boldsymbol{y}$ 为经过全局平均池化后的特征向量;W_i 为全连接层的权重;σ 为 Sigmoid 激活函数;ReLU 为 ReLU 函数的计算。

该机制主要由 3 个全连接层组成,通过全局平均池化,将输入特征 x 变为一维向量 $\boldsymbol{y}$,然后依次经过全连接层。最终通道注意力权重 A_c 通过 Sigmoid 激活得到,形成一个形状为 $(B,C,1,1)$ 的权重矩阵,其中 B 代表特征批量值,C 为通道数。

时间注意力通过两层一维卷积操作捕捉时间序列中的关键动态信息。首先将输入张量展平为适用于一维卷积的格式,输入特征 直接通过两层卷积层,得到一个时间特征表示,之后运用特征插值保持特征长度一致。

动态调整权重使用多层感知机(multilayer perceptron,MLP)对通道和时间注意力的输出特征进行加权融合,使用 Softmax 函数以确保权重的归一化。首先,Softmax 函数可以增强模型对多模态输入的自适应性,确保通道注意力与时间注意力特征在融合过程中不会出现不平衡;其次,通过实验对比静态加权求和与 MLP 加权,结果表明 MLP 结构能够捕捉更加复杂的特征关系,进一步提升模型性能。

将生成的动态权重应用到两个注意力模块的输出上,融合得到最后的输出,公式为

$$F = W_C C + W_T T \tag{4}$$

式(4)中:F 为融合后的特征表示;W_C 和 W_T 分别为生成的通道注意力权重与时间注意力权重;C 和 T 分别为通道注意力模块的输出和时间注意力模块的输出。

通过 MLP,模型可以根据输入特征动态调整通道和时间注意力的重要性,并通过权重融合机制更好地聚合这些特征,提高模型对不同特征的敏感度。通过联合建模与动态加权优化,注意力模块有效提升了伪造语音特征的区分性和检测精度。

2.3　深度可分离卷积残差块

深度可分离卷积结合残差模块,既能高效捕捉伪造语音中局部的特征模式,又能显著降低参数量和计算复杂度,使得模型更适用于资源受限的场景。本文研究将深度可分离卷积与残差连接、归一化和激活函数结合,构建了更关注局部特征的残差模块,将其应用于语音欺骗检测网络。

深度可分离卷积模块通过分解卷积操作作为深度卷积和逐点卷积,有效捕捉局部特征模式,同时大幅降低参数和计算复杂度,提升模型效率。对于时间序列数据,利用深度可分离卷积能够捕捉局部的模式和特征,捕捉时间的上局部依赖性,构建更为全面的特征表示。其中一个深度可分离卷积残差块结构如图4所示。

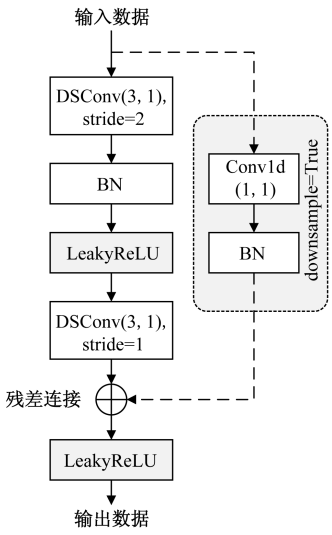
x 的输入形状为(B, C, L),对输入的张量逐个通道进行卷积,输出一个相同数量通道的特征映射,过程表示为

$$y_{\mathrm{d}}[c,l] = \sum_{k} x[c,l+k]w_{\mathrm{d}}[c,k] \tag{5}$$

式(5)中:x 为输入特征;w_{d} 为深度卷积核;c 为通道索引;l 为序列索引;k 为卷积核的索引,$k=3$;y_{d} 为深度卷积的输出。

随后 1×1 的卷积核用于跨通道整合特征,将输入通道映射的输出通道维度,将 y_{p} 表示为逐点卷积的输出,w_{p} 表示为 1×1 的卷积核,c_{in} 和 c_{out} 分别代表输入的通道索引和输出的通道索引,其中 c_{in} 设置为128、192、256、384、512,其公式表示为

$$y_{\mathrm{p}}[c_{\mathrm{out}},l] = \sum_{c_{\mathrm{in}}} y_{\mathrm{d}}[c_{\mathrm{in}},l+k]w_{\mathrm{p}}[c_{\mathrm{out}},c_{\mathrm{in}}] \tag{6}$$



DSConv为深度可分级卷积层;BN为批量归一化;Conv1d为一维卷积;downsample=True为执行下采样;LeakyReLU为LeakyReLU函数

图4　深度可分离卷积残差块

Fig. 4　Depthwise separable cconvolutional residual block

生成的新的特征组合通过BN层来稳定训练过程,加快收敛速度,使用LeakyReLU激活函数引入非线性特征,提高模型的表达能力。之后利用残差连接对输出进行更详细的特征变换与交互,保持输入特征并确保信息流畅传递,支持更深的网络结构。

堆叠的5个残差块逐层提取和学习更加复杂的特征,平衡模型复杂度与表达能力,通道数的扩展逐步增加特征图的深度,增强了网络容量。分阶段在第2个和第4个残差块中使用下采样,以减少特征图的尺寸。下采样操作降低了空间维度,使网络在更高的层次上能够观察到更大的感受野,也使得后续层的计算开销减小,同时逐渐聚焦于全局特征。本文研究通过设计特定的卷积分组策略和内核尺寸调整,使得深度可分离卷积对时域和频域特征的捕获更加精准,相较于标准卷积,提升了模型在伪造语音检测任务中的实时性和鲁棒性。

3　实验与结果分析

3.1　数据集介绍

本文研究在ASVspoof2019逻辑访问(logical access, LA)的数据集上进行训练[24]。ASVspoof挑战赛是Interspeech每隔两年举办的专门针对语音检测欺骗的赛事,每一届挑战赛都有专门的数据集以供研究者使用。ASVspoof2019挑战赛的LA数据集是基于VCTK[25]数据库进行划分的,包含多种TTS攻击、VC攻击以及TTS和VC的混合攻击,详情如表2所示。

ASVspoof2021[5]深度伪造(deepfake, DF)赛道,评估数据是使用通常用于媒体存储的不同有损编解码器处理的真实和欺骗语音的集合。对与LA相比,DF任务不涉及ASV系统的使用,其测试集有超过100种未公开的攻击算法。

FMFCC-A[26]数据集是中国内学者为促进语音深度伪造检测在中文场景下的发展,而推出的中文语音深度伪造数据集,是迄今为止最大的用于合成语音检测的公开普通话数据集,其伪造语音是根据11个普通话TTS系统和2个普通话VC系统生成的。

选用ASVspoof2019LA的测试集、ASVspoof2021DF的测试集和FMFCC-A的测试集进行跨数据集检测,进一步证明所提出CT-DSCNet模型的泛化性。测试集情况如表3所示。

表2　ASVspoof2019LA数据集详情

Table 2　Details of ASVspoof2019 LA dataset

数据集	说话人数量	真实语音	合成语音	伪造种类
训练集	20	2 580	22 800	A01-A06
开发集	20	2 548	22 296	A01-A06
测试集	48	7 355	63 382	A07-A19

表 3 不同测试集详情

Table 3 Details of different evaluation dataset

数据集	真实语音	合成语音	伪造种类
ASVspoof2019LA 测试集	7 355	63 382	13(A07-A19)
ASVspoof2021DF 测试集	18 452	163 114	>100(未知攻击)
FMFCC-A 测试集	3 000	17 000	13(TTS 系统)

3.2 数据处理

使用 Pytorch 框架来构建和训练所提出的伪造语音检测系统。本文提出的 CT-DSCNet 模型直接使用原始波形作为输入,通过裁剪长语音或连接短语音,所有输入语音的时长统一为 4 s。实验中未使用数据增强处理,分别使用 ASVspoof2019LA 的训练集和开发集来训练提出的模型,并选择最佳模型进行评估。所有模型均通过 ADAM 优化器进行训练,设置固定学习率为 0.000 1,批量大小为 8,训练轮数为 100,权重衰减系数为 0.000 1。

3.3 评估指标

等错误率(equal error rate, EER)用于评价单一的语音反欺骗系统的性能。在欺骗检测中,等错误率指的是真实语音被错误拒绝的比例和虚假语音被错误接受的比例相等的情况,EER 越小,检测系统的性能越好。

串联检测代价函数(tandem detection cost function, t-DCF)能够综合考虑反欺骗系统和 ASV 系统的共同作用。它考虑了错误接受率(false acceptance rate, FAR)、错误拒绝率(false rejection rate, FRR)以及二者之间的代价权衡。通过设置不同的代价超参数评估反欺骗系统和 ASV 系统对结果的影响,mint_{DCF}越小,系统泛化性越好。ASVSpoof2019LA 数据集的 min t_{DCF}评估指标计算公式为

$$t_{\mathrm{DCF}} = \min_{\tau}\left[C_1 P_{\mathrm{miss}}(\tau) + C_2 P_{\mathrm{fa}}(\tau)\right] \tag{7}$$

式(7)中:$P_{\mathrm{miss}}(\tau)$ 和 $P_{\mathrm{fa}}(\tau)$ 为反欺骗系统检测阈值为 τ 时的 FRR 和 FAR;C_1 和 C_2 为代价函数参数。

对于 ASVspoof2021DF 和 FMFCC-A 数据集,没有标准的 ASV 系统,仅采用 EER 作为度量指标评估单一的反欺骗系统性能。

3.4 消融实验

本节设计了消融实验,探讨设计选择对模型性能的具体影响,并验证所提出的通道-时间注意力机制和引入的深度可分离卷积的有效性。实验使用了 ASVspoof2019LA 数据集进行训练和评估。

首先探讨两种注意力融合方法(静态加权融合和 MLP 动态加权)与不同数量残差块堆叠的最佳组合。在注意力机制中首先使用静态加权融合方式,使用固定权重($w_{\mathrm{c}} = 0.5, w_{\mathrm{t}} = 0.5$)对通道和时间特征进行加权,并逐步增加残差块的堆叠数 N 以观察性能变化。实验结果显示,随着 N 的增加,模型的 EER 和 mint_{DCF}均呈现先下降后上升的趋势。接下来使用 MLP 动态加权的方法动态调整注意力权重,并结合不同数量的残差块进行实验,评估其对比效果。如图 5 所示,在 MLP 动态加权的情况下,模型性能相比静态加权融合进一步提升。具体表现为:EER 和 mint_{DCF}在 $N=5$ 时达到最低,随着 N 继续增加至 7,计算开销显著增加,训练时间延长,优化过程变得更为不稳定,模型容易捕捉训练数据中的无意义特征,最终导致性能明显下降。综合实验结果,当采用 MLP 动态加权和 5 层残差块的组合时,模型性能最佳,此时不仅在 EER 和 mint_{DCF}上均表现出色,同时也在计算效率和模型稳定性之间达成了较好的平衡。

在确定注意力机制模块的加权方法和残差块数量后,表 4 展示了引入不同模块后模型的性能表现,其中 CA 表示使用通道注意力模块,TA 表示使用时间注意力模块,DSConv 表示使用深度可分离卷积残差模块。在 ASVspoof2019LA 数据集上进行训练与评估后,结果显示,单独使用某一模块时,模型性能提升有限,CA 和 TA 的 EER 为 3.67% 和 2.52%,对应的 min t_{DCF}为 0.098 8 和 0.078 4。模块的增加和融合显著提升了模型检测效果,验证了全局与局部特征联合建模的可行性和有效性。然而,当移除深度可分离卷积模块时,模型性能略有下降,EER 上升至 1.98%,min t_{DCF}增加至 0.057 6。尽管性能下降幅度不大,但深度可分离卷积通过引入局部时间依赖性作为补充,进一步提升了模型的整体检测精度。消融实验充分验证了模型改进设计的有效性和所提出的方法在提高检测精度方面的显著优势,并在欺骗语音检测任务中具有较高的可行性。

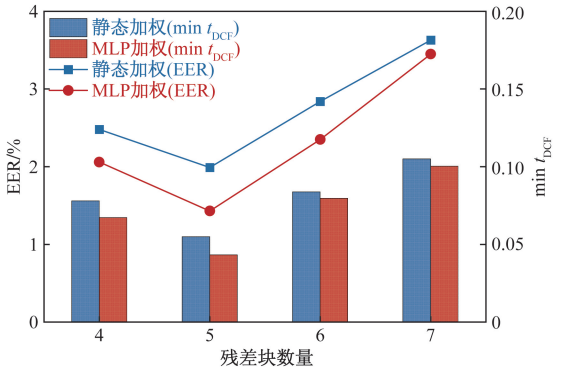


图 5 残差块数量和加权方法消融实验

Fig. 5 Residual block number and weighted method ablation experiment

表 4　ASVspoof2019LA 数据集消融实验结果

Table 4　Results of ablation experiments on the ASVspoof2019LA dataset

模型			EER/%	min t_{DCF}
CA	TA	DSConv		
—	—	—	4.86	0.119 5
√	—	—	3.67	0.098 8
—	√	—	2.52	0.078 4
√	√	—	1.98	0.057 6
√	√	√	**1.43**	**0.043 2**

注:加粗数值表示最优结果。

3.5　跨数据集消融实验

为了验证本文所提出的 CT-DSCNet 模型在中文数据集上的泛化能力,模型 3 个关键组件的消融研究的实验结果也 FMFCC-A 的评估集上进行,所得出的结果如表 5 所示。表 5 中,将去除 FMS 块的 RawNet2 表示为 RawNet2#。可以看出,由于 RawNet2 的泛化能力不高,基于 ASVspoof2019LA 训练集训练的模型,在 FMFCC-A 评估集上表现较差,仅达到 26.63% 的 EER。由于去除了原始 RawNet2 中的 FMS 模块,RawNet2#在 FMFCC-A 的评估集上仅达到 30.19% 的 EER。然而,当将 CA 添加到 RawNet2#时,系统达到了 25.28% 的 EER,表明通道注意力可以充分利用通道间的相互依赖性来捕获显著特征。只将 TA 添加到 RawNet2#时,系统的 EER 降低到 22.30%,这清楚地表明时域上的信息对于欺骗语音检测很重要,因为欺骗语音的伪影通常位于不同的子带上。同时将 CA 与 TA 共同插入 RawNet2#中时,性能得到进一步提升,EER 进一步下降,虽然与 RawNet2# + TA 相比只下降了 7.67%,但也证明了将通道特征与时序特征组合起来,也就是构建全局局部特征的依赖性,对语音反欺骗模型有重要意义。最后,将通道-时间注意力模块与深度可分离卷积残差块同时添加到 RawNet2#中时,CT-DSCNet 模型实现了 19.88% 的 EER,相对于原始的 RawNet2 模型提高了 25.35%,与 CT-DSCNet 模型在 ASVspoof2021DF 评估集上的表现(EER 为 18.01%),只差了 1.87%。可见深度可分离卷积残差块对提高模型的鲁棒性和泛化能力也有着非常重要的作用。

表 5　在 FMFCC-A 评估集上消融实验的 EER

Table 5　EER for ablation experiments on the FMFCC-A evaluation dataset

模型	EER/%
RawNet2[16]	26.63
RawNet2#	30.19
RawNet2# + CA	25.28
RawNet2# + TA	22.30
RawNet2# + CA + TA	20.59
RawNet2# + CA + TA + DSC(CT-DSCNet)	19.88

3.6　对比试验

表 6 展示了本文所提出的 CT-DSCNet 模型在 ASVspooof2019 和 2021 挑战赛不同评估集上的结果与官方基线结果之间的比较。CQCC-GMM[27]、LFCC-GMM[28]、LFCC-LCNN[29] 和 RawNet2[17] 是 ASVspoof 挑战赛中的官方基线系统。所有这些系统都是在 ASVspoof2019LA 的训练集上进行训练的,也就是说表 4 中的所有 EER 都是在同一训练集上训练的模型获得的。

通过对比官方系统的结果,可以明显看出,ASVspoof2019LA 评估集上的所有 EER 均远低于 ASVspoof2021 两个评估集上的 EER。这是因为训练集和评估集之间的欺骗算法非常接近,而 ASVspoof2021LA 和 DF 评估集中使用的合成欺骗算法与 ASVspoof2019LA 训练集相差甚远。此外,在 ASVspoof2019LA 上训练的模型不能很好地推广到其他评估条件,这表明它们非常容易受到跨数据集应用场景的影响。

仅从 ASVspoof2021LA 和 DF 评估集的结果来看,LFCC-LCNN[29] 和 RawNet2[17] 取得了相似的结果,但显着优于 CQCC-GMM[27] 和 LFCC-GMM[28] 系统。这表明 RawNet2[17] 是一个强大的基线系统,可以与提出的模型进行相对公平的比较。此外,从表 6可知,本文模型实现了最佳性能,并且其有效性可以很好地推广到交叉评估数据集上。具体来说,所提出的模型在 ASVspoof2019LA 任务上实现了 1.43% 的 EER,高出改进前模型 70.58%。在 ASVspoof2021LA 和 DF 任务上,模型的 EER 达到 3.86% 和 18.01%,分别比 RawNet2 基线相对高出 59.37% 和 19.53%。

表 6　CT-DSCNet 与其他模型 EER 的比较

Table 6　Comparison of the CT-DSCNet with other models EER

系统	ASVspoof2019LA	ASVspoof2021LA	ASVspoof2021DF
CQCC-GMM[27]	9.57	15.62	25.56
LFCC-GMM[28]	8.09	19.30	25.25
LFCC-LCNN[27]	—	9.26	28.48
RawNet2[16]	4.86	9.50	22.38
AASIST[11]	1.04	6.24	20.29
Rawformer[18]	1.05	4.98	—
CT-DSCNet	1.43	3.86	18.01

4　结论

构建了名为 CT-DSCNet 的欺骗语音检测模型,该模型通过在原始 RawNet2[17] 中引入通道-时间注意力机制模块和深度可分离卷积残差块,帮助模型

更好地提取全局-局部特征及其依赖性,来提高欺骗语音检测系统的性能和泛化能力。本文在 ASVspoof2019LA、ASVspoof2021LA、ASVspoof2021DF 和中文数据集 FMFCC-A 上进行了充分实验,结果表明,本文方法能够有效提高模型的泛化性能,提高模型对欺骗语音的检测能力。下一步工作将探讨涉及更多维度的多维全局注意力机制,以及不同的特征输入对模型检测性能的影响[30]。

参 考 文 献

[1] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: robust DNN embeddings for speaker recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018: 5329-5333.

[2] Wu Z, Kinnunen T, Evans N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge [C]//Interspeech 2015. Singapore: ISCA, 2015: 2037-2041.

[3] Nautsch A, Wang X, Evans N, et al. ASVspoof 2019: spoofing counter measures for the detection of synthesized, converted and replayed speech[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(2): 252-265.

[4] Kinnunen T, Sahidullah M D, Delgado H, et al. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection[C]//Interspeech 2017. Singapore: ISCA, 2017: 2-6.

[5] Yamagishi J, Wang X, Todisco M, et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection[C]//Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Singapore: ISCA, 2021: 47-54.

[6] 张雄伟, 李嘉康, 孙蒙, 等. 语音欺骗检测方法的研究现状及展望[J]. 数据采集与处理, 2020, 35(5): 807-823.
Zhang Xiongwei, Li Jiakang, Sun Meng, et al. Speech anti-spoofing: the state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2020, 35(5): 807-823.

[7] 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.

[8] Cui S, Huang B, Huang J, et al. Synthetic speech detection based on local autoregression and variance statistics[J]. IEEE Signal Processing Letters, 2022, 29: 1462-1466.

[9] Patel T B, Patil H A. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech[C]//Interspeech 2015. Singapore: ISCA, 2015: 2062-2066.

[10] Wu Z, Das R K, Yang J, et al. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks[J]. Interspeech, 2020, 2020: 1101-1105.

[11] 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 (ICASSP). Singapore: ICASSP, 2022: 6367-6371.

[12] Lei Z, Yan H, Liu C, et al. Two-path GMM-ResNet and GMM-SENet for ASV spoofing detection [C]//ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Singapore: ICASSP, 2022: 6377-6381.

[13] Hu C, Zhou R, Yuan Q. Synthetic speech spoofing detection based on online hard example mining[J]. IEEE Access, 2023, 11: 140443-140450.

[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]//Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Singapore: ISCA, 2021: 1-8.

[15] Tak H, Jung J W, Patino J, et al. Graph attention networks for anti-spoofing [C]//Interspeech: International Conference on Speech Communication and Technology. Brno: ISCA, 2021: 2356-2360.

[16] 杨海涛, 王华朋, 楚宪腾, 等. 基于卷积循环神经网络的语音逻辑攻击检测[J]. 科学技术与工程, 2022, 22(18): 7937-7944.
Yang Haitao, Wang Huapeng, Chu Xianteng, et al. Speech logic attack detection based on CNN-RNN-DNN network[J]. Science Technology and Engineering, 2022, 22(18): 7937-7944.

[17] 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(ICASSP). Singapore: ICASSP, 2021: 6369-6373.

[18] Zhang L, Li Y, Zhao H, et al. Backend ensemble for speaker verification and spoofing countermeasure[C]//Interspeech 2022. Singapore: ISCA, 2022: 4381-4385.

[19] Liu X, Liu M, Wang L, 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 (ICASSP). Rhodes Island: IEEE, 2023: 1-5.

[20] Ta B T, Nguyen T L, Dang D S, et al. A multi-task conformer for spoofing aware speaker verification[C]//IEEE Ninth International Conference on Communications and Electronics (ICCE). Nha Trang: ICCE, 2022: 306-310.

[21] 万玫汐, 王华朋, 闫道申, 等. 基于改进 ECAPA-TDNN 的法庭自动说话人识别[J]. 科学技术与工程, 2024, 24(27): 11763-11773.
Wan Meixi, Wang Huapeng, Yan Daoshen, et al. Forensic automatic speaker recognition based on enhanced ECAPA-TDNN[J]. Science Technology and Engineering, 2024, 24(27): 11763-11773.

[22] Howard A, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. arXiv: 2017, 1704.04861.

[23] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu: IEEE Computer Society, 2017: 1800-1807.

[24] Wang X, Yamagishi J, Todisco M, et al. ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech[J]. Computer Speech & Language, 2020, 64. DOI: 10.48550/arXiv.1911.01601.

[25] Zhang Z, Gu Y, Yi X, et al. FMFCC-A: a challenging mandarin dataset for synthetic speech detection[J]. arXiv: 2021, 2110.09441.

[26] 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.

[27] Sahidullah M, Kinnunen T, Hanilçi C. A comparison of features for synthetic speech detection[C]//Interspeech: International Conference on Speech Communication and Technology. Singapore: ISCA, 2015: 2087-2091.

[28] Wang X, Yamagishi J. A comparative study on recent neural spoofing countermeasures for synthetic speech detection[C]//Interspeech: International Conference on Speech Communication and Technology. Singapore: ISCA, 2021: 4259-4263.

[29] 李俊屿, 卜凡亮, 谭林, 等. 基于多模态共享网络的自监督语音-人脸跨模态关联学习方法[J]. 科学技术与工程, 2024, 24(7): 2804-2812.
Li Junyu, Bu Fanliang, Tan Lin, et al. Self-supervised voice-face cross-modal association learning method *via* multi-modal shared network[J]. Science Technology and Engineering, 2024, 24(7): 2804-2812.