研究与开发

基于上下文信息与注意力特征的欺骗语音检测

陈佳[1]，章坚武[1]，张浙亮[2]

（1. 杭州电子科技大学，浙江 杭州 310018；2. 浙江宇视科技有限公司，浙江 杭州 310051）

摘　要：随着语音合成和语音转换技术的快速发展，欺骗语音检测方法仍存在欺骗检测准确率低、通用性差等问题。因此，提出一种基于上下文信息与注意力特征的端到端的欺骗检测方法。该方法基于深度残差收缩网络（DRSN），利用双分支上下文信息协调融合模块（DCCM）聚集丰富的上下文信息，融合基于协调时频注意力机制（CTFA）的特征以获得具有上下文信息的跨维度交互特征，从而最大化捕获伪影的潜力。与最佳基线系统相比，在 ASVspoof 2019 LA 数据集中，所提方法在 EER 和 t-DCF 性能指标上分别降低 68%和 65%；在 ASVspoof 2021 LA 数据集中，所提方法的 EER 和 t-DCF 分别为 4.81 和 0.311 5，分别降低 48%和 10%。实验结果表明，所提方法能有效提高欺骗语音检测的准确率和泛化能力。

关键词：欺骗语音检测；上下文信息；注意力特征；端到端；伪影

中图分类号：TN912.3

文献标志码：A

doi: 10.11959/j.issn.1000–0801.2023006

Spoof speech detection based on context information and attention feature

CHEN Jia[1], ZHANG Jianwu[1], ZHANG Zheliang[2]

1. Hangzhou Dianzi University, Hangzhou 310018, China
2. Zhejiang Uniview Technologies Co., Ltd., Hangzhou 310051, China

Abstract: With the rapid development of speech synthesis and speech conversion technology, methods of spoof speech detection still have problems such as low spoof detection accuracy and poor generality. Therefore, an end-to-end spoof detection method based on context information and attention feature was proposed. Based on deep residual shrinkage network (DRSN), the proposed method used the dual-branch context information coordination fusion module (DCCM) to aggregate rich context information, and fused features based on coordinate time-frequency attention (CTFA) to obtain cross-dimensional interaction features with context information, thus maximizing the potential of capturing artifacts. Compared with the best baseline system, in the ASVspoof 2019 LA dataset, the proposed method had reduced the EER and t-DCF performance indicators by 68% and 65% respectively, in the ASVspoof 2021 LA dataset, the EER and t-DCF of the proposed method were 4.81 and 0.311 5 and dropped by 48% and 10% separately. The experimental results show that this method can effectively improve the accuracy and generalization ability of spoof speech detection.

Key words: spoof speech detection, context information, attention feature, end-to-end, artifacts

---

收稿日期：2022–11–28；**修回日期：**2023–01–05

通信作者：章坚武，jwzhang@hdu.edu.cn

基金项目：国家自然科学基金资助项目（No.U1866209，No.61772162）

Foundation Items: The National Natural Science Foundation of China (No.U1866209, No.61772162)

0　引言

自动说话人验证（automatic speaker verification，ASV）系统作为一种身份识别技术，旨在从语音信号中验证说话人的身份[1]，大力推动基于人类行为和生理特征监测及认证系统的发展[2]。ASV 系统验证过程不需要任何面对面的接触[3]，不会给用户带来不适和健康风险，但会导致该系统容易受到欺骗攻击。目前常用的反欺骗方法框架主要由前端特征提取和后端分类构成，将前端生成的手工声学特征输入后端分类器。徐剑等[4]直接从语谱图中提取完整局部二进制模式（completed local binary pattern，CLBP）纹理特征以提高欺骗语音检测的准确率。于佳祺等[5]将常量 Q 倒谱系数（constant Q cepstral coefficient，CQCC）声学特征与均匀局部二值模式（uniform local binary pattern，ULBP）纹理特征进行联合并输入随机森林分类模型以检测欺骗语音。手工声学特征在检测不可见的攻击时可能存在缺陷，因此已有工作提出了直接对原始音频波形进行操作的端到端（end-to-end，E2E）解决方案[6]，这种方案有效避免了手工声学特征带来的限制。Ge 等[7]探索了自动学习欺骗语音检测的方法，将架构搜索与 E2E 学习结合，提出了原始部分连接可差分结构搜索（raw partially-connected differentiable architecture search，Raw PC-DARTS）系统，该系统允许对网络架构和网络参数进行联合优化。为了有效捕获给定语音谱图中与欺骗攻击相关的伪影，Kang 等[8]建议在端到端欺骗对抗系统中采用注意力激活函数 AReLU[9]。尽管这些端到端系统的性能优于经典的欺骗检测系统，但研究结果表明其仍有很大的改进空间。

在 ASVspoof 2019[10]的逻辑访问（logical access，LA）场景中，合成语音欺骗攻击主要采取语音合成和语音转换的方式。用于指示欺骗攻击的人工制品称为欺骗伪影，人工制品的性能往往取决于攻击的性质和特定的攻击算法。在 ASVspoof 2021[11] LA 场景中，真实语音和欺骗语音通过各种电话网络进行未知编解码和传输。当语音数据在跨电话系统之间传输时，传输通道中可能会产生一些干扰性变化使数据中的欺骗伪影受到未知编解码和传输的影响，加大了欺骗检测的难度，从而提高了对欺骗检测系统的性能要求。在合成语音检测中，欺骗伪影用于区分真实语音与欺骗语音，主要存在于特定的时间和频谱间隔中，具有高区分性的时间特征和频率特征，但是目前没有一种较好的方法能够捕获存在于时域和频域间的伪装线索。无论在时域还是在频域，不同的注意力机制都会存在互补的、有区别的信息，且都适用于不同的欺骗攻击。Ling 等[12]利用频率注意力机制和通道注意力机制捕获频域和通道之间的关系，不仅将注意力集中到语音表示中信息量较大的频域中，还减少了通道冗余，但是该模型忽略了时域上的特征信息。Zhou 等[13]在欺骗语音检测中引入轻量级跨维度交互注意（lightweight cross-dimensional interaction attention，LCIA）模块以学习跨越不同频域和时域的欺骗线索，但该注意力机制没有充分融合上下文信息，导致容易忽略伪影的相关特征，高效地融合跨维度特征对于欺骗语音检测来说也十分重要。虽然现有方法的检测性能相比传统方法均有所提升，但随着各种高质量欺骗攻击的发展，现有的欺骗检测方法仍然缺乏对未知的欺骗攻击的有效性和通用性。针对以上问题，本文基于原始音频波形，提出一种上下文信息和注意力特征融合网络（context information and attention feature fusion network，CAFNet），该网络将上下文信息和基于注意力的跨维度交互特征进行融合以学习具有上下文信息的跨维度交互特征，同时克服由未知编解码和传输所带来的干扰，从而精确地识别并检测欺骗伪影。

本文的主要贡献包括以下 3 个方面。

- 设计了双分支上下文信息协调融合模块

（dual-branch context information coordination fusion module，DCCM），提取有价值的上下文信息以获得不同欺骗伪影之间的相关信息，融合基于注意力机制的跨维度交互特征以聚集区分性线索，集成具有上下文信息的跨维度交互特征来细化欺骗伪影的重要信息以获得全面的信息特征表示，有助于提高网络的抗干扰能力和高效地检测出欺骗伪影。

- 设计了协调时频注意力（coordinate time-frequency attention，CTFA）机制，捕获并融合时域和频域间的交互特征以及局部细粒度特征，最大限度地挖掘捕捉区分性线索的潜力，利用更多的细粒度特征信息以防止忽略细微伪影。
- 针对不同数据集之间存在数据组成、传输途径等差异，分析了所提网络的检测性能、通用性以及抗干扰能力。

1 相关工作

1.1 深度残差收缩网络

在卷积神经网络（convolutional neural network，CNN）中，深度残差网络（residual network，ResNet）[14]是其极具影响力的变体。对于早期的CNN模型，增加网络深度可能会使网络退化从而导致较高的训练误差，ResNet 使用恒等路径（identity shortcut）来解决这一问题以提高训练的正确率。Hua 等[15]基于原始语音波形，以 ResNet 的跳跃连接和 Inception[16]的并行卷积为网络架构，提出了一种端到端的轻量级欺骗检测模型，实现了较好的检测性能。但在处理噪声信号时，ResNet 的特征学习能力有待提升。深度残差收缩网络（deep residual shrinkage network，DRSN）[17]在ResNet的基础上学习基于注意力机制的阈值函数，并将学习到的最佳阈值提供给软阈值以自适应地从数据集中获得有用的特征并去除无关的噪声干扰。其中，阈值函数也称为收缩函数，通常用于信号去噪。周晔等[18]利用 DRSN 的去噪能力实现复杂声学环境下的欺骗语音检测，但其使用手工声学特征，容易丢失一些用于欺骗检测的有效信息。本文在 DRSN 的基础上，提出一种端到端的欺骗语音检测网络。

1.2 上下文信息和注意力特征

在实际应用场景中，欺骗对象不可能单独存在，其周围的对象一定会和该对象有或多或少的联系。当多个欺骗对象同时存在时，准确识别出欺骗对象是一项挑战，而增大感受野以获取有效的上下文信息有助于识别和检测欺骗对象。王金华等[19]提出一种基于卷积循环神经网络（convolutional recurrent neural network，CRNN）的语音情感识别算法，利用双向长短期记忆（bi-directional long short-term memory，BiLSTM）获得数据的序列上下文信息，有效提高算法的泛化性和区分性。Lei 等[20]设计分层上下文编码器来提取有效的上下文信息，显著提高合成语音的自然度和表达能力。注意力机制直观上可捕获全局和局部的依赖关系，防止网络过拟合，提高网络的泛化能力。挤压和激励网络（squeeze-and-excitation network，SENet）[21]在通道维度上增加注意力机制，但是没有考虑空间信息。卷积块注意力模块（convolutional block attention module，CBAM）[22]在 SENet 的基础上引入了空间注意力，同时对两个维度进行注意力分配，增强了注意力机制对模型性能的提升效果，但保留局部信息的效果较差。协调注意力（coordinate attention，CA）[23]将位置信息嵌入通道注意，有助于更准确地捕获方向和位置信息，但不能很好地整合全局和局部上下文信息。近年来，上下文信息、注意力机制在计算机听觉领域起着至关重要的作用，但是目前没有一种很好的方法将上下文信息和基于注意力的特征进行有效联合。

特征融合在现代网络架构中已被广泛使用，并且可以进一步提高 CNN 的性能。即便如此，大多数特征融合的工作为了实现多尺度特征的有效融合，需要构建复杂的路径，且不能很好地聚集上下文信息，以至于容易忽略欺骗对象的特征。注意力特征融合（attentional feature fusion，AFF）[24]可以融合不同层次或者分支的特征，来解决上下文聚合和初始集成的问题。该模块将接收到的特征与另一个 AFF 模块迭代集成，得到迭代注意力特征融合（iterative attentional feature fusion，iAFF）。iAFF 模块逐步优化初始集成，缓解特征的初始整合中基于注意力的特征融合的瓶颈，有效聚集上下文信息。本文提出了双分支上下文信息协调融合模块，将丰富的上下文信息和基于注意力的跨维度交互特征进行融合以准确识别区分性线索，具体介绍见第 2 节。

2　上下文信息和注意力特征融合网络

本文提出一种上下文信息和注意力特征融合网络，其结构如图 1 所示。本节首先介绍双分支上下文信息协调融合模块，其包含池化层分支和卷积层分支，然后介绍协调时频注意力机制，最后介绍该机制的两个组成模块。

2.1　双分支上下文信息协调融合模块

为了获取丰富的上下文信息和协调区分性线索的跨维度交互关系，本文设计了一种双分支上下文信息协调融合模块，以充分融合具有上下文信息的跨维度交互特征。该模块由卷积层分支和池化层分支组成，其结构如图 2 所示。

具体分析如下，首先将高级特征 $X \in \boldsymbol{R}^{C\times T\times F}$ 分别馈送到具有卷积层和池化层的两个并行分支中，其中 C、T、F 分别代表通道维度、时间维

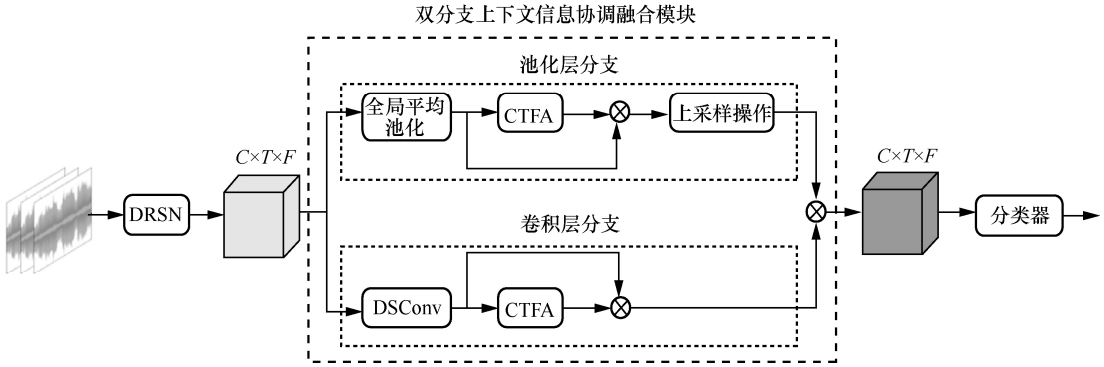

图 1　上下文信息和注意力特征融合网络结构

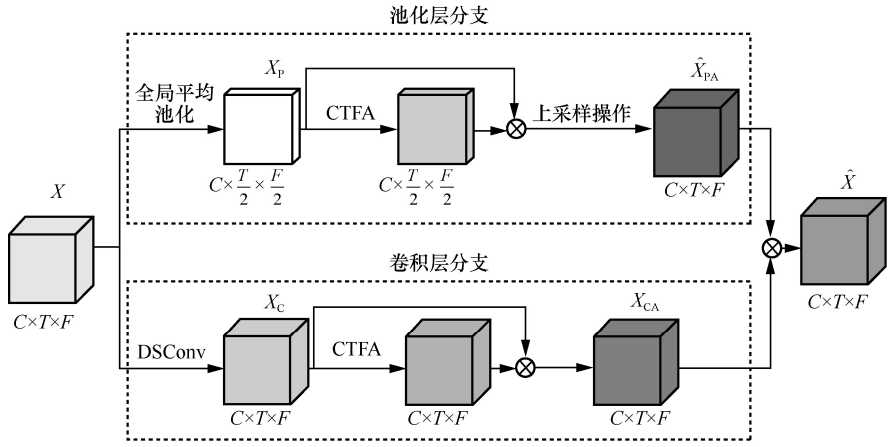

图 2　双分支上下文信息协调融合模块结构

度、频率维度。对于池化层分支，为了提取特征元素的上下文信息和防止过拟合，利用全局平均池化（global average pooling，GAPool）对输入特征 $X \in \boldsymbol{R}^{C\times T\times F}$ 进行降维，获得子特征 $X_{\mathrm{P}} \in \boldsymbol{R}^{C\times \frac{T}{2}\times \frac{F}{2}}$。其中，$\mathrm{GAPool}(X)$ 代表对高级特征 X 进行全局平均池化操作。

$$X_{\mathrm{P}} = \mathrm{GAPool}(X), X_{\mathrm{P}} \in \boldsymbol{R}^{C\times \frac{T}{2}\times \frac{F}{2}} \tag{1}$$

然后，使用协调时频注意力机制学习基于注意力的跨维度交互特征以获得跨维度伪装线索。接着，使用逐元素相乘对具有上下文信息的特征 X_{P} 与基于协调时频注意力的输出特征进行融合以最大化捕获区分性特征信息，得到含有上下文信息的跨维度交互特征 $X_{\mathrm{PA}} \in \boldsymbol{R}^{C\times \frac{T}{2}\times \frac{F}{2}}$。最后，使用上采样放大特征图以恢复其原始输入形状，得到特征 $\hat{X}_{\mathrm{PA}} \in \boldsymbol{R}^{C\times T\times F}$。

$$X_{\mathrm{PA}} = X_{\mathrm{P}} \otimes \mathrm{CTFA}(X_{\mathrm{P}}), X_{\mathrm{PA}} \in \boldsymbol{R}^{C\times \frac{T}{2}\times \frac{F}{2}} \tag{2}$$

$$\hat{X}_{\mathrm{PA}} = \mathrm{UpSample}(X_{\mathrm{PA}}), \hat{X}_{\mathrm{PA}} \in \boldsymbol{R}^{C\times T\times F} \tag{3}$$

其中，$\otimes$ 代表逐元素相乘，$\mathrm{CTFA}(X_{\mathrm{P}})$ 代表基于协调时频注意力的特征，UpSample 代表上采样操作。

对于卷积层分支，该分支利用卷积操作扩大特征元素的感受野以获取上下文信息。为了在减少训练参数量的同时提高模型的性能，使用深度可分离卷积（depthwise separable convolution，DSConv）代替普通 3×3 卷积，从而获得子特征 $X_{\mathrm{C}} \in \boldsymbol{R}^{C\times T\times F}$。其中，$\mathrm{DSConv}(X)$ 代表对高级特征 X 进行卷积操作。

$$X_{\mathrm{C}} = \mathrm{DSConv}(X), X_{\mathrm{C}} \in \boldsymbol{R}^{C\times T\times F} \tag{4}$$

与池化层分支相似，将具有上下文信息的子特征 X_{C} 与基于注意力的跨维度交互特征进行融合以获得具有上下文信息的跨维度交互特征 $X_{\mathrm{CA}} \in \boldsymbol{R}^{C\times T\times F}$。最后将两分支的跨维度交互特征逐元素相乘以获得更全面的特征表示，获得输出特征 $\hat{X} \in \boldsymbol{R}^{C\times T\times F}$。

$$X_{\mathrm{CA}} = X_{\mathrm{C}} \otimes \mathrm{CTFA}(X_{\mathrm{C}}), X_{\mathrm{CA}} \in \boldsymbol{R}^{C\times T\times F} \tag{5}$$

$$\hat{X} = X_{\mathrm{CA}} \otimes \hat{X}_{\mathrm{PA}}, \hat{X} \in \boldsymbol{R}^{C\times T\times F} \tag{6}$$

2.2 协调时频注意力机制

为了最大限度地从通道维度、时间维度、频率维度获取合理的特征表示以提供有效的欺骗线索，本文设计了协调时频注意力机制。该机制由时频融合模块（time-frequency fusion module，TFFM）和局部特征提取模块（local feature extraction module，LFEM）组成。协调时频注意力结构如图 3 所示。具体分析如下，首先将特征 $x \in \boldsymbol{R}^{C\times T\times F}$ 分别输入时频

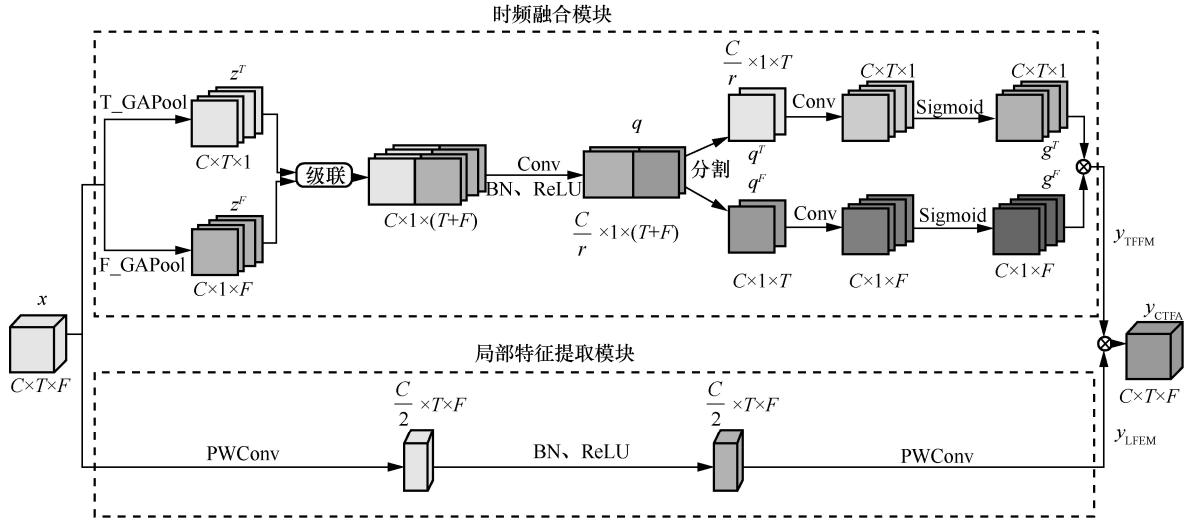

图 3 协调时频注意力结构

融合模块和局部特征提取模块，然后采用逐元素相乘来融合两分支的输出，得到基于注意力的跨维度交互特征 $y_{\mathrm{CTFA}} \in \boldsymbol{R}^{C\times T\times F}$。详细的处理步骤如式（7）所示。其中，TFFM($x$) 代表经过时频融合模块处理后的特征，LFEM($x$) 代表经过局部特征提取模块处理后的特征。

$$\begin{cases} y_{\mathrm{TFFM}} = \mathrm{TFFM}(x), y_{\mathrm{TFFM}} \in \boldsymbol{R}^{C\times T\times F} \\ y_{\mathrm{LFEM}} = \mathrm{LFEM}(x), y_{\mathrm{LFFM}} \in \boldsymbol{R}^{C\times T\times F} \\ y_{\mathrm{CTFA}} = y_{\mathrm{TFFM}} \otimes y_{\mathrm{LFEM}}, y_{\mathrm{CTFA}} \in \boldsymbol{R}^{C\times T\times F} \end{cases} \tag{7}$$

（1）时频融合模块

本文设计了时频融合模块，利用该模块的全局信息搜索能力挖掘时域和频域之间的交互关系以提高系统的检测性能。该模块既可以捕获通道信息，又可以最大限度地提取时频交互特征以捕获时间维度和频率维度之间潜在的区分性线索，使模型专注于特征信息最丰富的时间段和子带，增强感兴趣对象的特征表示。该模块将时间维度和频率维度上的信息嵌入通道维度，对输入的特征进行自适应特征细化。首先，该模块利用两个并行的一维全局平均池化进行特征编码以降低信息损失和帮助模型捕获全局上下文信息。即对于输入特征 $x \in \boldsymbol{R}^{C\times T\times F}$，使用池核 $(T,1)$ 和 $(1,F)$ 分别沿时间维度和频率维度对每个通道编码全局信息。

在时间维度上，池化后的输出特征为：

$$z^T = \mathrm{T\_GAP}(x), z^T \in \boldsymbol{R}^{C\times T\times 1} \tag{8}$$

在频率维度上，池化后的输出特征为：

$$z^F = \mathrm{F\_GAP}(x), z^F \in \boldsymbol{R}^{C\times 1\times F} \tag{9}$$

其中，T_GAP 代表时间维度上的全局平均池化，F_GAP 代表频率维度上的全局平均池化。将时间维度和频率维度上池化后的特征 z^T 和 z^F 进行级联，利用 1×1 卷积进行降维，捕获时间、频率维度的依赖关系，通过批量归一化 BN 和非线性激活函数 ReLU 对时间、频率维度的信息进行编码。

$$q \in \boldsymbol{R}^{\frac{C}{r}\times 1\times (T+F)} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}\left(\left[z^T, z^F\right]\right)\right)\right) \tag{10}$$

其中，$[\cdot,\cdot]$ 代表级联操作，Conv 代表 1×1 卷积，$q \in \boldsymbol{R}^{\frac{C}{r}\times 1\times (T+F)}$ 代表在时间、频率维度上的信息编码所获得的特征图。编码过程允许注意力机制更准确地捕获多维欺骗线索，从而帮助整个网络更有效地检测欺骗语音。在下一步中，将获得的特征图 q 分割，得到两个独立的张量 $q^T \in \boldsymbol{R}^{\frac{C}{r}\times T}$，$q^F \in \boldsymbol{R}^{\frac{C}{r}\times F}$。$r$ 是用于控制模块大小的缩减率。本文中，$r=2$。然后将两个张量分别馈送到 1×1 卷积运算中进行升维，以恢复与输入特征 $x \in \boldsymbol{R}^{C\times T\times F}$ 相同数量的通道。

$$g^T = \mathrm{Sigmoid}\left(\mathrm{Conv}(q^T)\right), g^T \in \boldsymbol{R}^{C\times T\times 1} \tag{11}$$

$$g^F = \mathrm{Sigmoid}\left(\mathrm{Conv}(q^F)\right), g^F \in \boldsymbol{R}^{C\times 1\times F} \tag{12}$$

其中，Sigmoid 代表归一化加权操作，$g^T \in \boldsymbol{R}^{C\times T\times 1}$、$g^F \in \boldsymbol{R}^{C\times 1\times F}$ 分别代表时间维度和频率维度的注意力权重。权重信息表示在特定时间和频率下特征信息的重要性。最后通过逐元素相乘将注意力权重 g^T 和 g^F 融合，实现并加强时间、频率维度上特征的互补性与相关性，最后输出的时频交互特征 $y_{\mathrm{TFFM}} \in \boldsymbol{R}^{C\times T\times F}$ 为：

$$y_{\mathrm{TFFM}} = \mathrm{TFFM}(x) = g^T \times g^F, y_{\mathrm{TFFM}} \in \boldsymbol{R}^{C\times T\times F} \tag{13}$$

（2）局部特征提取模块

由于欺骗语音系统中的关键特征是数据伪造后留下的欺骗伪影，这些伪影可能不包含语义信息，而包含一些细粒度特征信息，因此本文设计了局部特征提取模块来提取局部细粒度特征以帮助网络捕获更多细节信息。通过该模块在通道维度上获取并利用更多的细粒度特征信息以防止忽略细微伪影。

该模块利用逐点卷积（point wise convolution，PWConv）压缩和恢复通道维度上的特征，实现逐点通道交互，最后输出的细粒度特征 $y_{\mathrm{LFEM}} \in \boldsymbol{R}^{C\times T\times F}$ 的形状与输入特征 $x \in \boldsymbol{R}^{C\times T\times F}$ 一致，保留并突出显示有价值的细节线索以提升网络对细微伪影的捕捉能力。详细步骤如式（14）所示。

$$y_{\mathrm{LFEM}}=\mathrm{LFEM}(x)=\mathrm{PWConv}\left(\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{PWConv}(x)\right)\right)\right), y_{\mathrm{LFEM}}\in \boldsymbol{R}^{C\times T\times F} \tag{14}$$

3 实验及结果

3.1 数据集

ASVspoof 2019 LA 数据集基于 VCTK 语料库，使用最新的语音合成和语音转换算法生成欺骗语音信号。该数据集采用 107 名说话人（46 名男性、61 名女性）的语音作为语音样本，所有说话人的真实语音和欺骗语音被随机划分到互不相交的训练集、开发集和测试集。ASVspoof 2019 LA 数据集说话人和语音数量见表 1。

表 1 ASVspoof 2019 LA 数据集说话人和语音数量

ASVspoof 2019 LA	说话人/名		语音/条		
	男性	女性	真实	欺骗	总数
训练集	8	12	2 580	22 800	25 380
开发集	8	12	2 548	22 296	24 844
测试集	30	37	7 355	64 578	71 933

同时，本文选取 ASVspoof 2021 LA 数据集分析所提网络的抗干扰性能。与 ASVspoof 2019 LA 的测试集不同，ASVspoof 2021 LA 测试集由通过各种电话系统（包括 IP 电话（voice over Internet protocol，VoIP）和公共电话交换网（public switched telephone network，PSTN））传输的真实语音和欺骗语音组成，包含 181 566 条语音。ASVspoof 2021 LA 任务不会为单条语音提供编解码器元数据，该任务的重点是研究对未知编解码器和传输信道可变性干扰鲁棒的欺骗对策以区分由攻击生成的真实语音和欺骗语音。由于 ASVspoof 2021 LA 数据集中不包括单独的训练集和开发集，因此本文使用 ASVspoof 2019 LA 数据集的训练集和开发集作为 ASVspoof 2021 LA 数据集的训练集和开发集。

3.2 评价指标

本文使用官方评估指标：串联检测成本函数（tandem detection cost function，t-DCF）和等错误率（equal error rate，EER）检测不同网络的性能。两个指标值越小，网络性能越好。

$$\text{t-DCF}(\tau)=\frac{\mathrm{C}_1P_{\mathrm{miss}}^{\mathrm{cm}}(\tau)+\mathrm{C}_2P_{\mathrm{fa}}^{\mathrm{cm}}(\tau)}{\text{t-DCF}_{\mathrm{default}}} \tag{15}$$

$$\mathrm{EER}=P_{\mathrm{miss}}^{\mathrm{cm}}(\tau)=P_{\mathrm{fa}}^{\mathrm{cm}}(\tau) \tag{16}$$

其中，$P_{\mathrm{miss}}^{\mathrm{cm}}(\tau)$ 和 $P_{\mathrm{fa}}^{\mathrm{cm}}(\tau)$ 分别表示在阈值 τ 下欺骗语音检测系统的错误拒绝率和错误接受率；常数 C_1 和 C_2 表示两种错误成本，取决于 t-DCF 参数和 ASV 错误率；$\text{t-DCF}_{\mathrm{default}}$ 是接受或拒绝每个测试语音无信息的默认成本。

对于 ASVspoof 2021 LA 数据集，主要评估指标 t-DCF，对 t-DCF 指标稍微修改，具体如式（17）所示，其中，C_0 代表错误成本，其值取决于 t-DCF 参数和 ASV 错误率。

$$\text{t-DCF}\left(\tau_{\mathrm{cm}}\right)=\frac{\mathrm{C}_0+\mathrm{C}_1P_{\mathrm{miss}}^{\mathrm{cm}}\left(\tau_{\mathrm{cm}}\right)+\mathrm{C}_2P_{\mathrm{fa}}^{\mathrm{cm}}\left(\tau_{\mathrm{cm}}\right)}{\mathrm{C}_0+\min\left\{\mathrm{C}_1,\mathrm{C}_2\right\}} \tag{17}$$

3.3 不同注意力机制的性能对比

为了验证所提网络中 CTFA 对比其他注意力机制的优越性。本文在 ASVspoof 2019 LA 数据集上做了 4 组对比实验，不同注意力机制在 ASVspoof 2019 LA 数据集上的检测性能见表 2。具体而言，将 DCCM 中的 CTFA 用 SENet、CBAM、CA、LCIA 替换，其他条件均保持一致。

表 2 不同注意力机制在 ASVspoof 2019 LA 数据集上的检测性能

模型	t-DCF	EER
CAFNet-CTFA	0.044 7	1.48
CAFNet-LCIA	0.045 0	1.55
CAFNet-CBAM	0.051 8	1.78
CAFNet-CA	0.054 1	1.82
CAFNet-SENet	0.057 1	2.00

实验结果表明，与当前较流行的注意力机制相比，本文提出的 CTFA 检测性能更优越。为了分析以上 5 组模型，本文对模型进行多次训练并总结模型的检验结果，不同注意力机制的性能比

较如图 4 所示。从图 4 可以看出，CAFNet-CTFA 的 t-DCF 和 EER 最低，CAFNet-SENet 的 t-DCF 和 EER 最高。这表明，CTFA 能最大限度地提升特征学习能力，从而高效地提高网络的欺骗检测能力。与 LCIA 相比，CTFA 同时考虑时域和频域中潜在的欺骗线索以及局部细粒度信息，显著地提升了网络对目标伪影的捕获能力，证明了捕获局部细粒度特征的重要性。

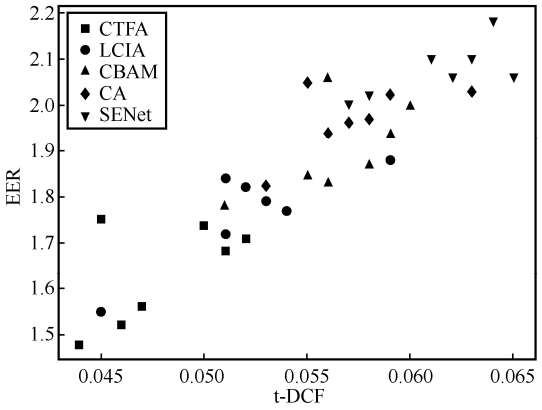

图 4　不同注意力机制的性能比较

3.4　消融实验

为了验证所提网络的有效性，本文在 ASVspoof 2019 LA 数据集上进行消融实验。具体而言，本文进行了 4 组消融实验：CAFNet 中未包含池化层分支（DRSN-convolution layer）、CAFNet 中未包含卷积层分支（DRSN-pool layer）、CAFNet 中未包含 CTFA（without CTFA）、CAFNet 中未包含 DCCM（DRSN）。CAFNet 在 ASVspoof 2019 LA 数据集上的消融实验结果见表 3。

表 3　CAFNet 在 ASVspoof 2019 LA 数据集上的消融实验结果

模型	t-DCF	EER
CAFNet	0.044 7	1.48
DRSN-convolution layer	0.056 9	2.18
DRSN-pool layer	0.062 0	2.23
without CTFA	0.071 0	2.84
DRSN	0.082 3	3.09

消融实验结果表明，卷积层分支和池化层分支都可以有效提高网络的检测性能。其中，卷积层分支的性能优于池化层分支，说明卷积更有助于扩大感受野以获取具有上下文信息的特征信息，从而更有效地检测区分性特征。此外，当添加 CTFA 模块时，EER 和 t-DCF 分别降低了 48% 和 37%。这表明将时间、频率维度间的交互特征和局部细粒度特征协调融合可帮助模型精准地捕获潜在的区分性线索。相比于 DRSN，DCCM 的引入可以更高效地提高模型的检测性能，充分证明将丰富的上下文信息与基于注意力的跨维度交互特征有效结合得到具有上下文信息的跨维度交互特征，能更加精准地捕获和鉴别欺骗伪影。CAFNet 和 ASV 系统的结合可以实现高效检测。

3.5　不同模型的对比实验

为了验证所提网络对比其他网络的性能优势，本文在 ASVspoof 2019 LA 数据集上将 CAFNet 与其他现有的竞争单系统进行比较，不同模型在 ASVspoof 2019 LA 数据集上的检测性能见表 4。

表 4　不同模型在 ASVspoof 2019 LA 数据集上的检测性能

模型	t-DCF	EER
CAFNet	0.044 7	1.48
Res-TSSDNet[15]	0.048 1	1.64
Raw CIANet-mul[13]	0.052 7	1.76
Capsule network[25]	0.053 8	1.76
Raw PC-DARTS Mel-F[7]	0.051 7	1.77
MCG-Res2Net50[26]	0.052 0	1.78
ResNet-FCA[12]	0.051 0	1.87
SE-ResNet-18-AReLU[8]	0.050 0	2.02
ResNet18-OC-softmax[27]	0.059 0	2.19
RawNet2[6]	0.129 4	4.66

从表 4 可以得出，在 ASVspoof 2019 LA 数据集上，CAFNet 实现了最佳检测性能，EER 和 t-DCF 分别降至 1.44 和 0.044 7。相比于 RawNet2，CAFNet 在 EER 和 t-DCF 性能指标上分别降低 68%和 65%。ResNet-FCA 将频率注意力和通道注意力进行融合，仅将注意力集中在语音表示中信

息较丰富的子带上，而忽略了重要的时域信息。Raw CIANet-mul 构建了一种新的时频注意力模块，可以捕获时间和频率间的跨维度交互线索，但忽略了上下文信息和局部细粒度信息对欺骗检测的重要性。CAFNet 在使用注意力机制聚焦有价值的特征的同时，有效地扩大感受野以获取上下文信息，从而实现高效的欺骗检测。本文提出的 DCCM 可以在 CTFA 提供的重要注意力线索的指导下有效集成包含丰富上下文信息的跨维度交互特征，提高网络检测欺骗线索的能力。实验结果表明，CAFNet 对于欺骗语音检测是有效的。

3.6 CAFNet 的通用性和抗干扰性能

为了验证所提网络的通用性和抗干扰性能，本文在 ASVspoof 2021 LA 数据集上将 CAFNet 与其他现有的竞争单系统进行比较，不同模型在 ASVspoof 2021 LA 数据集上的检测性能见表 5。

表 5 不同模型在 ASVspoof 2021 LA 数据集上的检测性能

模型	t-DCF	EER
CAFNet	0.311 5	4.81
ResNet-LogSpec[28]	0.293 0	5.18
LCNN[29]	0.319 7	5.27
RawNet2-RawBoost[30]	0.309 9	5.31
Lightweight TDNN-Focal[31]	0.364 5	7.51
SE-ResNet34-avg-OC-Softmax[32]	0.332 0	9.03
LFCC-LCNN[11]	0.344 5	9.26

在 ASVspoof 2021 LA 应用场景中，当所有语音数据在电话系统之间传输时，数据中的欺骗伪影可能会受到未知编解码和传输的干扰，从而使 ASVspoof 更接近实际的应用场景，大大增加了语音检测的复杂度，因此本文提出的网络需要很好地消除不同的干扰变化。从表 5 可以得出，在 ASVspoof 2021 LA 数据集上，CAFNet 实现了较好的泛化性能和抗干扰性能，EER 和 t-DCF 分别降至 4.81 和 0.311 5。相较于 ResNet-LogSpec 和 RawNet2-RawBoost，CAFNet 的 t-DCF 指标值略高，这两种系统均针对 ASVspoof 2021 LA 数据集的特点对数据进行不同方式的数据增强，显著提高了系统对电话场景中存在的未知干扰性变化的鲁棒性。相较于 LFCC-LCNN，CAFNet 在 EER 和 t-DCF 性能指标上分别降低 48%和 10%，进一步证明了基于端到端的 CAFNet 具有较强的通用性和抗干扰性能。

4 结束语

为了有效捕获并鉴别欺骗线索以及解决高质量欺骗攻击的通用性问题，本文提出一种端到端的上下文信息和注意力特征融合网络，设计了协调时频注意力机制以最大化捕获时域和频域中欺骗线索的潜力和有效利用局部细粒度特征，设计了双分支上下文信息协调融合模块以获得具有上下文信息的跨维度交互特征，从而提高网络的特征学习能力。消融实验表明，CAFNet 中使用的 DCCM、CTFA 是有效的。在不同数据集上的实验结果表明，CAFNet 在检测欺骗语音方面具有良好的实用性和普适性，并且比其他竞争单系统具有优势。在 ASVspoof 2021 LA 任务中，CAFNet 对电话场景中存在的未知干扰性变化的鲁棒性还有待提高。未来将在提升网络检测性能的同时，研究一种基于数据增强的轻量化欺骗检测网络，简化网络复杂度和参数量。

参考文献：

[1] KINNUNEN T, LI H. An overview of text-independent speaker recognition: from features to supervectors[J]. Speech communication, 2010, 52(1): 12-40.

[2] SINGH N, AGRAWAL A, KHAN R A. Voice biometric: a technology for voice based authentication[J]. Advanced Science, Engineering and Medicine, 2018, 10(7-8): 754-759.

[3] MITTAL A, DUA M. Automatic speaker verification systems and spoof detection techniques: review and analysis[J]. International Journal of Speech Technology, 2021(25): 1-30.

[4] 徐剑, 简志华, 于佳祺, 等. 采用完整局部二进制模式的伪装语音检测[J]. 电信科学, 2021, 37(5): 91-99.
XU J, JIAN Z H, YU J Q, et al. Completed local binary pattern based speech anti-spoofing[J]. Telecommunications Science, 2021, 37(5): 91-99.

[5] 于佳祺, 简志华, 徐嘉, 等. 基于联合特征与随机森林的伪

装语音检测[J]. 电信科学, 2022, 38(6): 91-99.
YU J Q, JIAN Z H, XU J, et al. Spoofing speech detection algorithm based on joint feature and random forest[J]. Telecommunications Science, 2022, 38(6): 91-99.

[6] 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 (ICASSP). Piscataway: IEEE Press, 2021: 6369-6373.

[7] GE W Y, PATINO J, TODISCO M, et al. Raw differentiable architecture search for speech deep fake and spoofing detection[EB]. 2021.

[8] KANG W H, ALAM J, FATHAN A. Attentive activation function for improving end-to-end spoofing countermeasure systems[EB]. 2022.

[9] CHEN D S, LI J, XU K. AReLU: attention-based rectified linear unit[EB]. 2020.

[10] WANG X, YAMAGISHI J, TODISCO M, et al. ASVspoof 2019: a large-scale public data base of synthesized, converted and replayed speech[J]. Computer Speech & Language, 2020, 64: 101-114.

[11] YAMAGISHI J, WANG X, TODISCO M, et al. ASVspoof 2021: accelerating progress in spoofed and deep fake speech detection[EB]. 2021.

[12] LING H F, HUANG L C, HUANG J R, et al. Attention-based convolutional neural network for ASV spoofing detection[C]// Proceedings of 2021 INTERSPEECH. [S.l.:s.n.], 2021: 4289-4293.

[13] ZHOU Y, ZHANG J W, ZHANG P G. Spoof speech detection based on raw cross-dimension interaction attention network[C]// Proceedings of 2022 Chinese Conference on Biometric Recognition. Cham: Springer, 2022: 621-629.

[14] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[15] HUA G, TEOH A B J, ZHANG H. Towards end-to-end synthetic speech detection[J]. IEEE Signal Processing Letters, 2021, 28: 1265-1269.

[16] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1-9.

[17] ZHAO M H, ZHONG S S, FU X Y, et al. Deep residual shrinkage networks for fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2019, 16(7): 4681-4690.

[18] 周晔, 章坚武, 程继承. 面向复杂声学环境的伪装语音检测[J]. 传感技术学报, 2022, 35(10): 1355-1362.
ZHOU Y, ZHANG J W, CHENG J C. Speech anti-spoofing for complex acoustic environments[J]. Chinese Journal of Sensors and Actuators, 2022, 35(10): 1355-1362.

[19] 王金华, 应娜, 朱辰都, 等. 基于语谱图提取深度空间注意特征的语音情感识别算法[J]. 电信科学, 2019, 35(7): 100-108.
WANG J H, YING N, ZHU C D, et al. Speech emotion recognition algorithm based on spectrogram feature extraction of deep space attention feature[J]. Telecommunications Science, 2019, 35(7): 100-108.

[20] LEI S, ZHOU Y X, CHEN L Y, et al. Towards expressive speaking style modelling with hierarchical context information for mandarin speech synthesis[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 7922-7926.

[21] HU J, SHEN L, ALBANIE S. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[22] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the 2018 European Conference on Computer Vision. [S.l.:s.n.], 2018: 3-19.

[23] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recongnition. Piscataway: IEEE Press, 2021: 13713-13722.

[24] DAI Y M, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2021: 3560-3569.

[25] LUO A W, LI E L, LIU Y L, et al. A capsule network based approach for detection of audio spoofing attacks[C]//Proceed ings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 6359-6363.

[26] LI X, WU X X, LU H, et al. Channel-wise gated Res2Net: towards robust detection of synthetic speech attacks[C]//Proceedings of 2021 INTERSPEECH. [S.l.:s.n.], 2021: 4314-4318.

[27] ZHANG Y, JIANG F, DUAN Z Y. One-class learning towards synthetic voice spoofing detection[J]. IEEE Signal Processing Letters, 2021, 28: 937-941.

[28] COHEN A, RIMON I, AFLALO E, et al. A study on data augmentation in voice anti-spoofing[J]. Speech Communication, 2022, 141: 56-67.

[29] DAS R K. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deep fake attacks: ASV spoof 2021[C]//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [S.l.:s.n.], 2021: 29-36.

[30] TAK H, KAMBLE M, PATINO J, et al. Raw boost: a raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6382-6386.

[31] CÁCERES J, FONT R, GRAU T. The biometric vox system for the ASVspoof 2021 challenge[C]//Proceedings 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [S.l.:s.n.], 2021: 68-74.
[32] PAL M, RAIKAR A, PANDA A, et al. Synthetic speech detection using meta-learning with prototypical loss[EB]. 2022.

[作者简介]

陈佳（2000– ），女，杭州电子科技大学通信工程学院硕士生，主要研究方向为语音检测与人工智能等。

章坚武（1961– ），男，博士，杭州电子科技大学通信工程学院教授、博士生导师，中国电子学会高级会员，浙江省通信学会常务理事，主要研究方向为移动通信、多媒体信号处理与人工智能、通信网络与信息安全。

张浙亮（1969– ），男，博士，浙江宇视科技有限公司副总裁，主要研究方向为人工智能、人力资源等。