多阶 GMM-ResNet 融合在语音伪造检测中的研究

曹明明, 雷震春, 杨印根, 周 勇

江西师范大学计算机信息工程学院 南昌 中国 330022

摘要 近年来, 自动说话人识别技术取得了显著进步, 但同时也容易受到合成或转换语音的伪造攻击, 语音伪造检测系统致力于解决这一问题。本文根据不同阶数 GMM 中高斯分量之间的相关性和 ResNet 模型中不同层次残差块输出的特征信息, 提出了一种多阶 GMM-ResNet 融合模型进行语音伪造检测。该模型主要包含两部分: 多阶对数高斯概率(Log Gaussian Probability, LGP)特征融合和多尺度特征聚合 ResNet(Multi-Scale Feature Aggregation ResNet, MFA-ResNet)。GMM 描述了语音特征在其空间的分布情况, 不同阶数的 GMM 则具有不同描述能力来形成对特征分布的平滑近似。此外, 根据不同阶数 GMM 计算出来的 LGP 特征也就在不同阶上捕获语音信息。多阶 LGP 特征融合将基于不同阶数的 GMM 得到的三种不同阶 LGP 特征进行加权融合, 从而促进不同阶 LGP 特征之间的信息交换。另一方面, 神经网络模型中第一层或中间层获得的特征信息对于分类任务也是非常有用的。基于这一经验, MFA-ResNet 模块通过对每个 ResNet 块输出的特征进行聚合, 充分融合网络内不同层级的特征信息, 从而提高网络的特征提取能力。在 ASVspoof 2019 逻辑访问场景下, LFCC+多阶 GMM-ResNet 融合系统的 min t-DCF 和 EER 分别为 0.0353 和 1.16%, 比基线系统 LFCC+GMM 分别相对降低了 83.3%和 85.7%。在 ASVspoof 2021 逻辑访问场景下, LFCC+多阶 GMM-ResNet 融合系统的 min t-DCF 和 EER 分别为 0.2459 和 2.50%, 比基线系统 LFCC+GMM 分别相对降低了 57.3%和 87.1%, 比基线系统 LFCC+LCNN 分别相对降低了 28.6%和 73.0%。与目前最先进模型相比, 本文模型也非常具有竞争力。

关键词 多阶 GMM-ResNet 融合; 多阶对数高斯概率特征融合; 多尺度特征聚合; 语音伪造检测

中图法分类号 TP391; TP183 DOI 号 10.19363/J.cnki.cn10-1380/tn.2025.03.08

Research on Multi-order GMM-ResNet Fusion for Speech Deepfake Detection

CAO Mingming, LEI Zhenchun, YANG Yingen, ZHOU Yong

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China

Abstract Automatic speaker verification technology has made remarkable progress in recent years, but it is also vulnerable to deepfake attacks by synthesized or converted speech. Therefore, the speech deepfake detection systems have been developed to address this issue. In this paper, we propose a fusion model that combines multi-order GMMs with ResNet for speech deepfake detection. The model leverages the correlation between Gaussian components in different order GMMs and the feature maps of all residual blocks in the ResNet model. The multi-order GMM-ResNet fusion model mainly consists of two parts: multi-order Log Gaussian Probability (LGP) Feature fusion and Multi-scale Feature Aggregation ResNet (MFA-ResNet). The conventional GMM describes the distribution of speech features in the feature space, and different order GMMs have different descriptive abilities to form smooth approximations to the feature distribution. Additionally, the multi-order LGP features are based on the different order GMMs, which also capture the speech information at different scales. The multi-order LGP feature fusion module weights three order LGP features from different order GMMs and facilitates information exchange between them. On the other hand, the feature information obtained in the first or intermediate layers in neural network model is also very useful for classification tasks. Based on this experience, the MFA-ResNet module aggregates all outputs from ResNet blocks, and all feature maps can also contribute towards the accurate speech embedding extraction. On the ASVspoof 2019 logical access task, the LFCC+multi-order GMM-ResNet fusion system achieves a minimum t-DCF of 0.0353 and an EER of 1.16%, which relatively reduces by 83.3% and 85.7% compared with the LFCC+GMM baseline. On the ASVspoof 2021 logical access task, the LFCC+multi-order GMM-ResNet fusion system achieves a minimum t-DCF of 0.2459 and an EER of 2.50%, which relatively reduces by 57.3% and 87.1% compared with the LFCC+GMM baseline, and relatively reduces by 28.6% and 73.0% compared with the LFCC+ LCNN baseline. Compared with current state-of-the-art models, the proposed model is competitive.

Key words multi-order GMM-ResNet fusion; multi-order log-gaussian probability feature fusion; multi-scale feature aggregation; speech deepfake detection

---

通讯作者: 雷震春, 博士, 副教授, Email: zhenchun.lei@hotmail.com。

本课题得到国家自然科学基金(No. 62067004), 江西省教育厅科学技术研究项目(No. GJJ2200331)资助。

收稿日期: 2023-08-03; 修改日期: 2023-11-28; 定稿日期: 2025-02-12

1 引言

近年来, 随着深度学习模型的广泛应用, 自动说话人确认(Automatic Speaker Verification, ASV)[1-2]技术也得到了快速发展, 并展现出良好的性能。与人脸识别、指纹识别、虹膜识别等生物识别方法类似, ASV 技术也被大规模应用于各种服务场景中[3-4]。同时, 针对 ASV 系统的语音伪造攻击[5-6]也越来越频繁, 设计有效的语音伪造检测模型对于提高 ASV 系统的可靠性变得至关重要[7]。

目前, 人们对语音伪造检测模型的研究主要从两方面进行: 前端声学特征和后端分类模型。在前端声学特征方面, 研究者提出许多新的声学特征, 以提高特征的判别能力。Todisco 等人[8]使用常数 Q 变换(Constant Q Transform, CQT)代替傅里叶变换, 提出了常数 Q 倒谱系数(Constant Q Cepstral Coefficients, CQCC), 它具有在低频段频率分辨率高, 高频段时间分辨率高的特点, 被广泛应用在语音伪造检测系统中。Sahidullah 等人[9]提出了基于滤波器的线性频率倒谱系数(Linear Frequency Cepstral Coefficients, LFCC), 该系数通过使用线性频率分布的滤波器组代替传统的 Mel 分布的滤波器组, 与传统的梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)相比更专注于高频特征。相比特定人工设计的声学特征, 使用深度神经网络直接从原始波形中提取通用特征表示的方法也取得了非常大的进展。Tak 等人[10]首次在语音伪造检测中应用 RawNet2 深度神经网络, RawNet2 强大的表征能力可以直接从原始波形中提取更高级别、更通用的特征表示, 几乎完全避免了对人工设计特征的依赖。近年来, 随着无监督预训练模型的发展, 研究者也开始尝试将预训练模型的架构和方法应用到语音伪造检测当中。Yang 等人[11]使用 wav2vec 预训练模型作为特征提取器, 直接运行在原始波形上, 从而获得更鲁棒的语音特征表示。在 ASVspoof 2021 逻辑访问场景(Logical Access, LA)和语音深度伪造(Speech Deepfake, DF)检测中, Tak 等人[12]使用改进版本 wav2vec 2.0 作为特征提取器对原始波形进行特征提取, 将获得的通用特征输入到后端模型分类器中, 分类性能得到了显著的提升。

在后端的分类模型中, 经典的高斯混合模型(Gaussian Mixture Model, GMM)[13]通常作为挑战赛的基线系统。随着深度学习技术的快速发展, 越来越多的神经网络模型被应用于语音伪造检测, 特别是卷积神经网络(Convolutional Neural Networks, CNN)在语音伪造检测任务中表现出卓越的性能。如 Lavrentyeva 等人[14]提出在卷积层后引入最大特征图(Max feature map, MFM)函数的轻量卷积神经网络(Light Convolutional Neural Networks, LCNN), ASVspoof 2017 挑战赛和 ASVspoof 2019 挑战赛逻辑访问(Logical access, LA)场景下最佳单系统都基于该网络模型。He 等人[15]为了解决训练深层神经网络过程中出现网络退化和梯度消失等问题, 提出了残差网络(Residual Networks, ResNet)模型, 且广泛应用于语音伪造检测。Alzantot 等人[16]受 ResNet 在许多分类任务中取得成功的启发, 提出了用于语音伪造检测的 ResNet, 针对 MFCC, CQCC 和频谱图三种不同输入特征构建出三种不同的 ResNet 变体进行实验, 实验表明 ResNet 在语音伪造检测中同样显示出先进的性能。Kwak 等人[17]提出 ResMax 检测模型, 该模型把 ResNet 残差结构和 LCNN 中的 MFM 函数相结合, 在减少参数量的同时还提升了模型的性能。Li 等人[18]提出了基于 Res2Net 网络结构的语音伪造攻击检测模型, 通过修改 ResNet 块获得多尺度特征表示, 从而提高系统的泛化能力。Lai 等人[19]提出了基于挤压-激励网络(Squeeze-Excitation Network, SE-Net)[20]和 ResNet 的检测模型 ASSERT, 并引入统计池化方法解决语音伪造攻击问题。Wang[21]等人通过全局和时频特征图两个层面的注意力机制为不同的特征赋予不同的注意力权重, 提出了全局-时频注意力网络模型, 并使用 A-softmax 损失函数替换 softmax 损失函数进一步扩大了真伪语音的区分性。

在经典的 GMM 模型中, GMM 独立的累计所有帧的分数, 并未考虑每个高斯分量对最终分数的贡献程度, 且忽略了语音特征帧之间的局部关系。2022 年, Lei 等人[22]提出 GMM-ResNet 模型, 综合考虑了特征帧在所有 GMM 分量上的得分分布情况和帧之间的相互联系, 在 ASVspoof 2019 数据集上取得良好的性能。在训练 GMM 过程中, 人们通常使用二叉分裂方式来估计出模型的最终参数。二叉分裂过程从单阶 GMM 开始, 分裂成 2 阶 GMM, …, 直到选取的 512 阶 GMM, 且每次进行分裂之后, 都需要对模型参数进行重新估计。在二叉分裂过程中, 不同阶数的 GMM 包含的高斯分量之间存在一定的联系。因此, 本文提出多阶 LGP 特征融合模型, 通过加权融合的方式促进不同阶数 LGP 特征之间的信息交换。

以前的研究[23]表明, 深度神经网络模型中的低级特征信息也有助于说话人嵌入的提取。基于这一经验, Zhang 等人[24]提出 MFA-Conformer 模型, 通过将每个 Conformer 块的输出特征图进行串联, 然后进行层标准化处理, 以进一步改善结果。Jung 等人[25]

将 ECAPA-TDNN 和 RawNet2 模型进行混合提出 RawNet3 说话人识别模型, 并把第一层和第二层输出的总和输入到第三层中, 然后将三层输出的特征进行聚合也取得了具有竞争力的效果。Juan 等人[26]通过使用预训练模型 wav2vec2 作为特征提取器, 对来自每个 Transfomer 层的特征进行时间归一化之后, 将每层特征进行权重求和得到最终输入到后端分类模型的特征表示, 展现了其在 ASVspoof 2021 挑战赛 LA 和 DF 任务中的良好检测性能。受这一启发, 本文提出一种多尺度特征聚合 ResNet(Multi-scale Feature Aggregation ResNet, MFA-ResNet) 网络模型, MFA-ResNet 通过采用 ResNet 强大的网络特征提取能力, 将每个残差块的输出特征图进行拼接, 以在最终池化之前聚合多个层级特征表示提高模型的检测性能。

本文提出的多阶 GMM-ResNet 融合模型的主要贡献如下: 1) 在原对数高斯概率特征的基础上进行了扩展, 提出了多阶高斯概率特征融合, 目的是促进不同阶数 LGP 特征之间的信息交换。2) 提出了 MFA-ResNet 模型用于语音伪造检测, 该模型结构通过对每个 ResNet 残差块输出的特征进行聚合得到更充分的特征信息。3) 将多阶对数高斯概率特征融合与多尺度特征聚合 ResNet 模型进行整合, 提出了多阶 GMM-ResNet 融合模型进行语音伪造检测。

2 相关工作

2.1 高斯混合模型

GMM 是一种经典的概率式聚类模型, 它采用多维概率密度函数对语音特征进行建模。GMM 由 K 个具有不同权重和不同参数的单高斯概率密度函数线性加权组合而成, 可以拟合样本空间中任意形状的数据分布。其概率密度函数如公式(1)所示,

$$P(x)=\sum_{i=1}^{K} w_i p_i(x) \tag{1}$$

其中, K 是高斯分布个数, x 是大小为 D 的样本向量, w_i 为第 i 个高斯分量的权重, 且满足 $\sum_{i=1}^{K} w_i=1$, $p_i(x)$ 是单高斯概率密度函数, μ_i 为均值向量, Σ_i 为协方差矩阵, 如公式(2)所示,

$$p_i(x)=\frac{1}{2\Pi^{D/2}\left|\Sigma_i\right|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \tag{2}$$

在历届 ASVspoof 挑战赛中, 基线系统通常包含两个 GMM, 它们分别在所有真实语音和伪造语音特征集上训练得到。在评估过程中, 基线系统计算测试语音在这两个 GMM 上的对数似然值之差, 作为判断测试语音是真实语音还是伪造语音的依据, 如公式(3)所示,

$$score(X)=logp(X|\lambda_g)-logp(X|\lambda_s) \tag{3}$$

式中, λ_g 和 λ_s 分别表示真实语音和伪造语音上 GMM 参数, $score(X)$ 表示测试语音 X 的对数似然比分数。

2.2 对数高斯概率特征

对于语音帧特征 x, GMM 对 K 个高斯分量概率密度值进行累加, 却没有考虑各个高斯分量具体得分分布情况。由于真实语音和伪造语音在特征空间上的分布存在差异, 所以它们在各个高斯分量上的得分分布也会不同。因此, Lei 等人[22]认为这种得分分布信息对语音伪造检测是有用的, 并基于 LFCC 特征在每个高斯分量上的得分分布信息构建对数高斯概率特征。

对于原始语音帧特征 x, 经过计算后新的对数高斯概率特征 y 的大小等于 GMM 的阶数, 其中分量 y_i 表示如公式(4)所示,

$$y_i=\ln p_i(x) \tag{4}$$

然后, 对 y_i 进行均值方差归一化, 得到最终的对数高斯概率特征 f_i, 如公式(5)所示,

$$f_i=\frac{y_i-mean_{y_i}}{std_{y_i}} \tag{5}$$

其中, $mean_{y_i}$ 和 std_{y_i} 分别为整个训练集上 y_i 的均值和方差。

3 多阶 GMM-ResNet 融合模型

3.1 多阶 GMM-ResNet 融合模型框架

本文提出的多阶 GMM-ResNet 融合模型进行语音伪造检测的流程图如图 1 所示, 主要分为多阶 LGP 特征融合和多尺度特征聚合 ResNet(MFA-ResNet)模型两个部分。在多阶 LGP 特征融合部分, 首先将真实语音特征和伪造语音特征放在一起训练 128、256 和 512 阶的 GMM 模型, 并根据这三个 GMM 计算出相应的 LGP 特征, 然后将这三种不同阶数的 LGP 特征进行加权融合。

在多尺度特征聚合 ResNet 模型部分, 首先将多阶数 LGP 特征融合得到的三种不同维数的特征分别输入到三路基于 1D-CNN 的 MFA-ResNet 模型中。然后将得到的特征图进行自适应最大池化操作。最后将每路得到的特征进行拼接输入到全连接线性层, 通过 Softmax 函数输出真实语音和伪造语音标签的概率分布, 从而进行语音伪造检测。

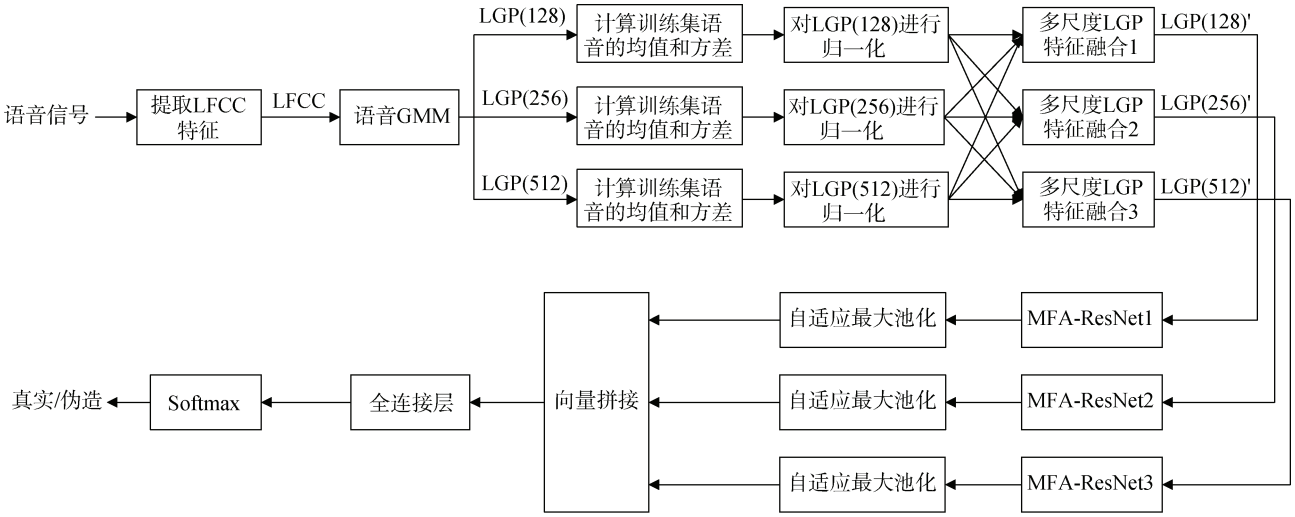

图 1　多阶 GMM-ResNet 融合模型流程图

Figure 1　Flowchart of the Multi-order GMM-ResNet fusion model

3.2　多阶 LGP 特征融合

GMM 描述了语音特征在其空间的分布情况, 不同阶数的 GMM 则具有不同描述能力, 且根据不同阶数 GMM 计算出来的 LGP 特征也就在不同阶上反映出语音所包含的信息。人们训练 GMM 模型时通常采用二叉分裂和最大期望(Expectation-Maximization, EM)算法进行反复迭代求解, 直至最后收敛。在二叉分裂过程中, GMM 最初只包含 1 个高斯分量, 然后在这个高斯分量的基础上, 通过加上和减去一个小的偏移值生成 2 个高斯分量, 并采用 EM 算法对其进行重新估计。以此类推, 2 个高斯分量分裂为 4 个, 4 个分裂为 8 个, …, 一直到目标数量为止(本文采用 512 阶 GMM)。根据二叉分裂算法, 分裂过程中不同阶数的 GMM 包含的高斯分量也具有一定的关联, 依此计算的 LGP 特征也会存在相关性。因此, 本文对二叉分裂过程中不同阶数 GMM 得到的不同阶 LGP 进行加权融合, 进一步提高语音伪造检测的效果。多阶 LGP 特征融合模型图如图 2 所示。

本文根据二叉分裂过程中 128 阶、256 阶和 512 阶 GMM 分别计算 LGP 特征, 然后将这三种不同阶数的 LGP 特征通过加权方式进行融合。计算公式如下:

$$X_1' = w_{11}X_1 + w_{12}f_{12}(X_2) + w_{13}f_{13}(X_3) \tag{6}$$

$$X_2' = w_{21}f_{21}(X_1) + w_{22}X_2 + w_{23}f_{23}(X_3) \tag{7}$$

$$X_3' = w_{31}f_{31}(X_1) + w_{32}f_{32}(X_2) + w_{33}X_3 \tag{8}$$

$$w_{i1} + w_{i2} + w_{i3} = 1 \tag{9}$$

其中, X_1, X_2, X_3 分别表示 128, 256, 512 阶的高斯概率特征, X_1', X_2', X_3' 分别表示多阶高斯概率特征融合之后的结果, f 代表卷积操作, w 表示可训练的权重参数。由于三个阶数的特征具有不同的通道数, 我们使用卷积层 f 进行扩充或压缩, 其中卷积核大小为 1, 步长为 1。在多阶 LGP 特征融合过程中, 不同阶数的特征对总体的贡献度不一样, 本文采用加权融合的方式进行训练, 且系数加权和为 1。经多阶 LGP 特征融合后, 系统仍然得到 128, 256, 512 三种不同阶数的特征, 并作为后续分类器的输入。

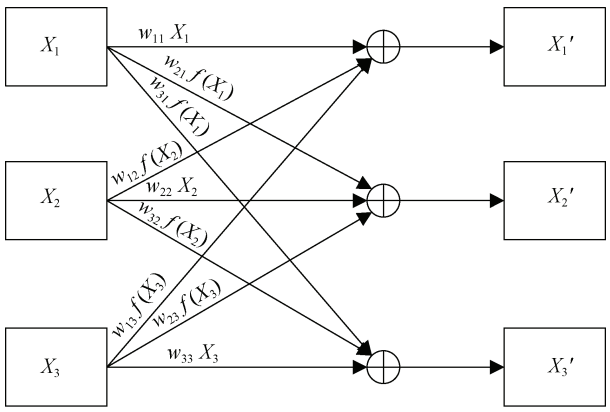

图 2　多阶 LGP 特征融合模块

Figure 2　The multi-order LGP feature fusion module

3.3　多尺度聚合 ResNet(MFA-ResNet)模型

基于 CNN 的网络架构已经在语音伪造检测研究领域展示出了其有效性[14,16], 通过卷积操作使得其在局部特征提取方面具有天然的优势。但随着网络层数的加深, 训练过程中会出现梯度爆炸或梯度消失, 甚至网络退化的问题。本文采用残差结构的思想建立残差网络, 提高模型训练过程的稳定性。残差块的结构如图 3 所示, 每个残差块包含两个一维卷积

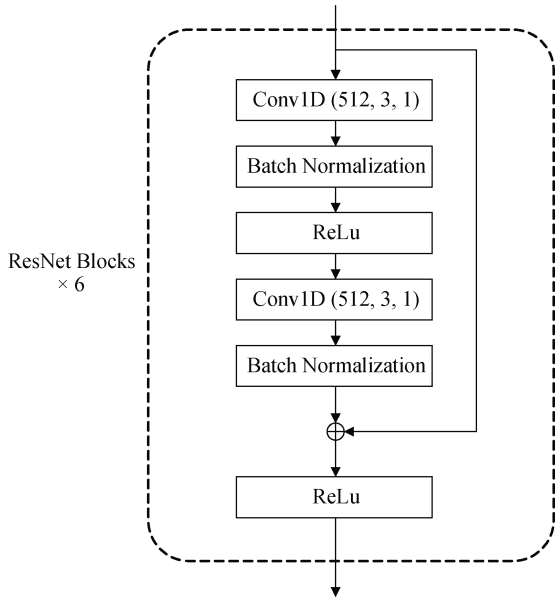

图 3　残差网络结构图

Figure 3　The architecture of the ResNet Block

层, 并且在两个卷积层中间依次引入批处理归一化层(Batch Normalization, BN)和修正线性单元(Rectified Linear Unit, ReLU)。它们的作用是加快特征在深度神经网络中的收敛速度和训练效率, 得到更好的训练效果。然后再将输入特征与经过两个卷积块运算并进行批处理归一化得到的特征图进行逐元素相加。最后, 通过 ReLu 激活函数得到最终的残差块输出。

以前的研究[23-26]表明, 在训练深度神经网络时, 第一层或中间层获得的特征信息对于分类任务也是非常有用的。例如在说话人确认系统中, 多尺度聚合的方法及其变体在说话人确认任务中带来了明显的性能提升。这些方法采用多尺度聚合的方式提取说话人嵌入, 充分融合网络内不同层级的特征信息。受这一思想的启发, 本文提出 MFA-ResNet 网络模型。首先采用拼接的方式将每个残差块中输出的特征信息进行融合, 以获得更充分的特征信息, 其次将它们输入到批处理归一化层, 如公式(10)所示,

$$X = BatchNorm1d\left(Concat\left(x_1, x_2, \ldots, x_L\right)\right) \quad (10)$$

其中, $Concat$ 表示拼接操作, x_i 表示第 i 个 block 块的输出特征值, L 表示 block 块的个数。

然后将上述操作提取的特征使用自适应最大池化在时间维度上进行压缩, 获得对应维度的最具表示性的特征向量。最后将该特征输入到全连接线性层, 通过 Softmax 函数输出真实语音和伪造语音标签的概率分布进行分类。MFA-ResNet 模型结构图如图 4 所示。本文通过基于 1-D CNN 的 MFA-ResNet 模型对 LGP 特征进行建模, 不仅考虑了 GMM 分量上语音特征帧的分数分布情况, 还充分考虑了帧之间的局部关系。

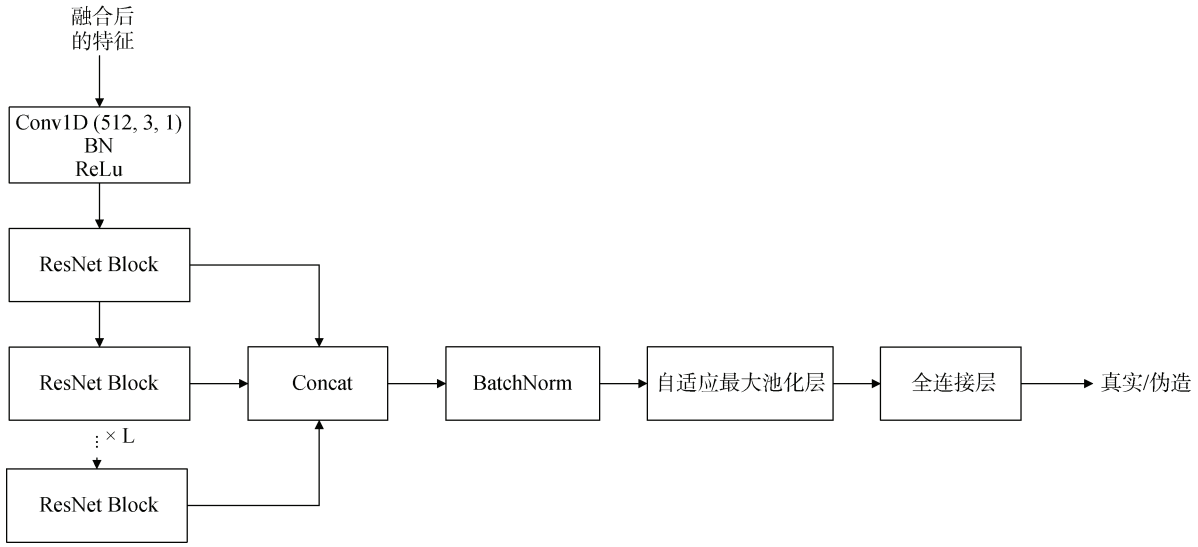

图 4　MFA-ResNet 模型结构

Figure 4　The architecture of the MFA-ResNet

3.4　两步训练

多步训练方案[27]通常被用来解决模型过拟合问题, 本文通过采用两步训练的方式提高模型的鲁棒性。在第一步中, 独立预训练多阶高斯概率特征融合模块和 MFA-ResNet 模块, 将带有 Softmax 输出的全连接层临时添加到每一路径的自适应最大池化层之后, 并使用交叉熵损失函数进行训练。在第二步中, 移除临时添加的全连接层, 然后冻结多阶高斯概率特征融合模块和 MFA-ResNet 模块的模型参数来训练全连接层分类器。

3.5 数据增强

在神经网络训练中, 数据增强(Data Augmentation, DA)[28-32]能够减少过拟合, 从而提高模型的泛化能力。目前在语音伪造领域进行数据增强的方法主要有两种: ①利用外部数据集进行数据增强, 例如添加混响和背景噪声[29,31]等; ②直接对原始波形进行操作, 如编解码[32], FIR 滤波[30], RawBoost[28]等。本文采用了 RawBoost 数据增强方法, 它专为电话场景而设计, 通过添加不同的干扰噪声对训练数据进行不同程度的增强。例如, 线性和非线性卷积噪声, 与脉冲信号相关的加性噪声和与稳态信号无关的加性噪声, 以及它们之间的组合。

RawBoost 数据增强采用与原论文中相同的实验参数配置。本文在经过多次实验验证之后, 使用线性和非线性卷积噪声、与脉冲信号相关的加性噪声以及稳态信号无关的加性噪声三者的组合更适合本文所提出的系统模型。本文将 RawBoost 所得到的增强数据集与原始训练数据集进行混合作为新的训练数据集, 这样训练数据集就被扩充了 1 倍。

4 实验设置

4.1 数据集

本文实验在 ASVspoof 2019 和 ASVspoof 2021 挑战赛 LA 场景数据集上进行。根据挑战赛规则, 模型在 ASVspoof 2019 训练集上进行训练, 且在 ASVspoof 2019 开发集、评估集和 ASVspoof 2021 评估集(eval 集)上进行测试。ASVspoof 2019 LA 和 ASVspoof 2021 LA 数据集的具体数据分布情况如表 1 所示:

ASVspoof 2019 LA 数据集取自于 VCTK 基本语音库, 由真实语音和通过语音合成(Text to Speech, TTS)以及语音转换(Voice Conversion, VC)生成的伪造语音组成。其中, 训练集和开发集中的伪造语音是由编号 A01-A06 算法生成的, 包含 2 种语音转换和 4 种语音合成算法; 评估集中的伪造语音是由编号 A07-A19 算法生成的, 包含 6 种语音转换和 7 种语音合成算法, 且其中有两种已知算法在训练集中出现过, 其余为未知算法。ASVspoof 2021 LA 评估集包含的伪造攻击算法和 ASVspoof 2019 评估集出现的一样, 并进一步将语音样本使用某些特定的编码器在公共交换电话网络(Public Switched Telephone Network, PSTN)或基于 IP 的语音传输(Voice over Internet Protocol, VoIP)网络上进行了传输处理, 加大了伪造语音检测的难度。

表 1 ASVspoof 2019 与 ASVspoof 2021 LA 数据集分布

Table 1 The distribution of ASVspoof 2019 and ASVspoof 2021 LA dataset

	真实语音	伪造语音
2019 训练集	2580	22800
2019 开发集	2548	22296
2019 评估集	7355	63882
2021 评估集(eval 集)	14816	133360

4.2 评价指标

实验采用 ASVspoof 2021 挑战赛主办方提供的最小串联检测代价函数(Minimum tandem Detection cost Function, Min t-DCF)[13]作为主要指标和等错误率(Equal Error Rate, EER)[13]作为次要指标来评估模型的性能。t-DCF 和 EER 的值越低, 语音伪造检测模型的性能越好。

4.2.1 串联检测代价函数

为了评估语音伪造检测模型与 ASV 系统一起使用时的综合性能, ASVspoof 2019 年引入了以 ASV 为中心的串联检测代价函数作为主要评价指标。与 ASVspoof 2019 给出的公式不同的是, ASVspoof 2021 保留了参数 C_0, 为了方便计算, 通常使用串联检测代价函数的最小归一化形式表示, 定义如公式(11)所示:

$$\min \mathrm{t}-DCF=\min_{\tau_{cm}}\left\{\frac{C_0+C_1P_{miss}^{cm}\left(\tau_{cm}\right)+C_2P_{fa}^{cm}\left(\tau_{cm}\right)}{C_0+min\left\{C_1,C_2\right\}}\right\} \tag{11}$$

其中, $P_{miss}^{cm}\left(\tau_{cm}\right)$ 和 $P_{fa}^{cm}\left(\tau_{cm}\right)$ 分别表示 CM 系统在阈值 τ_{cm} 处的未命中率和误报率; C_0, C_1, C_2 都取决于 $t-DCF$ 参数和 ASV 系统的错误率。它们的表示如下所示:

$$C_0=\pi_{tar}C_{miss}P_{miss}^{asv}+\pi_{non}C_{fa}P_{fa}^{asv} \tag{12}$$

$$C_1=\pi_{tar}C_{miss}-\left(\pi_{tar}C_{miss}P_{miss}^{asv}+\pi_{non}C_{fa}P_{fa}^{asv}\right) \tag{13}$$

$$C_2=\pi_{spoof}C_{fa,spoof}P_{fa,spoof}^{asv} \tag{14}$$

其中, π_{tar}, π_{non}, π_{spoof} 分别表示目标说话人、非目标说话人和伪造语音的先验概率(非负且总和为 1)。C_{miss}, C_{fa}, $C_{fa,spoof}$ 分别表示漏检目标用户、错误接收非目标用户说话人和错误接受伪造语音的代价。P_{miss}^{asv}, P_{fa}^{asv}, $P_{fa,spoof}^{asv}$ 表示在 ASV 系统阈值下未命中率、误报率和伪造攻击错误接受率。

4.2.2 等错误率

等错误率(Equal Error Rate, EER)表示的是错误拒绝率(False Rejection Rate, FRR)等于错误接受率

(False Acceptance Rate, FAR)时所对应的取值, 常用于评估语音伪造检测模型的性能。其中, FRR 表示真实语音被模型检测为伪造语音的比例, FAR 表示伪造语音被模型检测为真实语音的比例。

4.3 参数设置及训练方式

实验使用 LFCC 作为语音伪造检测模型的原始声学特征, 特征提取器采用 ASVspoof 2021 挑战赛主办方提供的基线系统 LFCC-LCNN 的实现方式。在特征提取过程中, 首先对语音信号进行分帧, 帧长为 20 ms、帧移为 10 ms。然后将分帧的每一帧信号使用汉明窗进行加窗, 并做 1024 点傅里叶变换之后, 输入到线性三角滤波器组进行处理, 其中滤波器的个数为 20, 最后进行对数运算和离散余弦变换得到倒谱系数。通过提取线性频率倒谱系数的一阶差分和二阶差分系数, 并与原系数进行拼接最终得到 60 维 LFCC 特征向量。语音 LFCC 特征都沿时间轴保留 400 帧的固定长度, 若长度大于 400 帧直接进行截取, 小于 400 帧则重复补齐。

在进行训练高斯混合模型的过程中, 本文采用将 ASVspoof 2019 训练集的真实语音和伪造语音混合在一起进行训练, 使用 MSR Identity Toolbox[33]工具箱训练具有 128 阶, 256 阶, 512 阶三种高斯混合模型, 其中 EM 迭代次数为 30。

实验使用 Pytorch 深度学习框架实现神经网络模型, 在带有 GTX 3090 GPU 的服务器进行模型训练。训练过程中, 损失函数采用交叉熵损失函数, 优化器使用 Adam, 初始学习率为 0.0001, 当连续 5 个周期损失值没有改善时, 动态调整学习率为原来的 0.1 倍。训练模型的迭代次数为 100, 批处理大小设置为 32。

5 实验结果及分析

实验首先分别验证多阶 LGP 特征融合和 MFA-ResNet 模型的有效性, 然后对多阶 GMM-ResNet 融合模型进行了实验, 并与当前主流模型的实验结果进行比较。

5.1 多阶 LGP 特征融合有效性实验

为验证多阶 LGP 特征融合的有效性, 本文设计了 7 个对比实验: 3 个实验分别是 128、256、512 阶 LGP 特征与单路 ResNet 模型组合; 2 个实验是三种 LGP 特征独立输入到三路 ResNet 模型且分别采用用单步和两步训练; 2 个实验是三种 LGP 特征经过多阶加权融合之后输入到三路 ResNet 模型且分别采用单步和两步训练。实验结果如表 2 所示。

表 2 多阶 LGP 特征融合+ResNet 模型在 ASVspoof 2019 和 ASVspoof 2021 LA 数据集上实验结果比较

Table 2 Comparison of experimental results of multi-order LGP feature fusion + ResNet on ASVspoof 2019 and ASVspoof 2021 LA datasets

特征阶数	训练方式	数据增强	ASVspoof 2019 开发集		ASVspoof 2019 评估集		ASVspoof 2021 评估集	
			min t-DCF	EER/%	min t-DCF	EER/%	min t-DCF	EER/%
128	-	-	0.0090	0.31	0.1121	3.92	0.4188	10.02
256	-	-	0.0076	0.31	0.0894	3.20	0.3661	8.19
512	-	-	0.0071	0.34	0.0838	3.06	0.3800	8.49
独立三路	-	-	0.0070	0.27	0.0954	3.45	0.3802	8.77
	两步	-	0.0062	0.24	0.0902	3.18	0.3715	8.26
多阶特征融合	-	-	**0.0060**	0.31	0.0838	3.05	0.3617	7.90
	两步	-	0.0073	**0.24**	0.0778	2.72	0.3560	7.64
128	-	✓	0.0120	0.39	0.0572	1.96	0.2711	3.64
256	-	✓	0.0162	0.48	0.0415	1.49	0.2592	3.17
512	-	✓	0.0125	0.39	0.0413	1.47	0.2568	3.03
独立三路	-	✓	0.0121	0.39	0.0384	1.40	0.2501	2.86
	两步	✓	0.0112	0.35	0.0383	1.27	0.2506	2.79
多阶特征融合	-	✓	0.0086	0.27	0.0402	1.41	0.2496	2.72
	两步	✓	0.0093	0.31	**0.0354**	**1.24**	**0.2488**	**2.62**

从表 2 可以看出, 在这三种单阶 GMM-ResNet 模型中, 128 阶特征结果最差, 256 阶特征次之, 512 阶特征最好。这反映出了高斯分量越多, GMM 对语音特征的拟合效果越好, 得到的 LGP 特征对语音伪造检测的效果也越好。本文提出的多阶 LGP 特征融合模型明显优于三种单阶 GMM-ResNet 模型。如果将三种 LGP 特征不经过特征融合独立输入到三路 ResNet 模型中, 模型的性能会降低。由此可见, 加

权特征融合的作用很重要, 这也反映出每种特征在融合的过程中所做出的贡献是不一样的。此外, 两步训练方式可以进一步提高模型的检测性能。使用数据增强的情况下, 与单阶 512 阶 LGP 特征相比, 多阶 LGP 特征融合(两步)在 ASVspoof 2019 LA 评估集上的 min t-DCF 和 EER 分别相对降低了 14.29%和 15.65%; 在 ASVspoof 2021 LA 评估集上的 min t-DCF 和 EER 分别相对降低了 3.12%和 13.53%。

此外, 从表 2 还可以看出, 在开发集上的性能指标都远优于评估集上的性能指标。产生这种结果的原因在于, ASVspoof 2019 评估集上包含 11 种训练集和开发集没有的未知攻击算法; ASVspoof 2021 评估集在使用和 ASVspoof 2019 评估集相同攻击算的同时, 还另外使用了某些特定的编码器并通过各种电话系统进行了传输处理, 加大了检测的难度。因此为了提高系统的泛化性能, 在进行模型训练过程对训练集进行一定的数据增强是有必要的。

5.2 MFA-ResNet 模型有效性实验

为了验证 MFA-ResNet 模型有效性, 本文暂时去除模型中的多阶 LGP 融合模块, 并使用单路 ResNet 模型进行对比实验。本文使用 512 阶的 LGP 作为模型的输入, 分别输入到 ResNet 模型和 MFA-ResNet 模型当中进行实验。表 3 显示了 ResNet 模型和 MFA-ResNet 模型分别在 ASVspoof 2019 评估集与 ASVspoof 2021 评估集上的实验结果。

表 3 ResNet 与 MFA-ResNet 在 ASVspoof 2019 和 ASVspoof 2021 LA 数据集上的实验结果比较

Table 3 Comparison of experimental results between ResNet and MFA-ResNet on ASVspoof 2019 and ASVspoof 2021 LA datasets

模型	数据增强	ASVspoof 2019 开发集		ASVspoof 2019 评估集		ASVspoof 2021 评估集	
		min t-DCF	EER/%	min t-DCF	EER/%	min t-DCF	EER/%
ResNet	-	**0.0071**	0.34	0.0838	3.06	0.3800	8.49
MFA-ResNet	-	0.0074	**0.27**	0.0793	2.88	0.3623	7.61
ResNet	✓	0.0125	0.39	0.0413	1.47	0.2568	3.03
MFA-ResNet	✓	0.0107	0.35	**0.0410**	**1.33**	**0.2515**	**2.78**

从表 3 中实验结果可以看出, 使用多尺度特征聚合得到的 MFA-ResNet 模型相对于 ResNet 模型都有明显的性能提升。不使用数据增强的情况下, 与 ResNet 模型相比, MFA-ResNet 模型在 ASVspoof 2019 LA 评估集上的 min t-DCF 和 EER 分别相对降低了 5.37%和 5.88%; 在 ASVspoof 2021 LA 评估集上的 min t-DCF 和 EER 分别相对降低了 4.66%和 10.37%。

5.3 多阶 GMM-ResNet 融合模型实验

5.3.1 ASVspoof 2019 LA 场景下的实验

多阶 GMM-ResNet 融合模型整合了多阶 LGP 特征融合和 MFA-ResNet 模型。表 4 显示了 ASVspoof 2019 LA 场景下基线系统、512 阶 GMM-ResNet 模型和多阶 GMM-ResNet 融合模型的实验结果。在使用数据增强的情况下, 多阶 GMM-ResNet 融合模型在评估集上 min t-DCF 和 EER 分别为 0.0353 和 1.16%, 相对 512 阶 GMM-ResNet 模型分别降低了 14.5%和 21.1%; 相对基线系统 LFCC+GMM 分别降低了 83.3%和 85.7%。

表 5 显示了多阶 GMM-ResNet 融合模型与目前主流的单一语音伪造检测模型在 ASVspoof 2019 LA 数据集上的实验结果比较。表 5 包含了具有代表性的前端特征和模型结构, 且各个模型性能指标 min t-DCF 和 EER 均来自原论文给出的结果, 其中 LFCC+LCNN 系统为 ASVspoof 2019 挑战赛 LA 场景下最佳单模型系统。从表 5 中可以看出, 只有 FFT+SENet、Raw waveform+AASIST 和 Wav2vec 2.0+VIB 三个系统的性能优于多阶 GMM-ResNet 融合模型。Wav2vec 2.0+VIB 系统是目前已知最好的单一系统, 该系统使用最近流行的无监督预训练模型 wav2vec2.0 作为特征提取器, 然后采用微调的方式得到最终的语音特征向量。虽然使用 wav2vec2.0[37]预训练模

表 4 多阶 GMM-ResNet 融合模型在 ASVspoof 2019 LA 数据集上的实验结果

Table 4 Experimental results of the multi-order GMM-ResNet fusion model on the ASVspoof 2019 LA dataset

特征	模型	数据增强	min t-DCF	EER/%
LFCC	GMM[16]	-	0.2116	8.09
CQCC	GMM[16]	-	0.2366	9.57
LFCC	GMM-ResNet	-	0.0838	3.06
	多阶 GMM-ResNet 融合	-	0.0713	2.49
	GMM-ResNet	✓	0.0413	1.47
	多阶 GMM-ResNet 融合	✓	**0.0353**	**1.16**

表 5 多阶 GMM-ResNet 融合模型与其他主流模型在 ASVspoof 2019 LA 数据集上的实验结果比较

Table 5 Comparison of experimental results between the multi-order GMM-ResNet fusion model and other mainstream models on the ASVspoof 2019 LA dataset

特征	模型	min t-DCF	EER/%
LFCC	LCNN[14]	0.1000	5.06
LFCC[34]	ResNet18-OC-softmax	0.0590	2.19
LFB[29]	ResNet18-LMCL-FM	0.0520	1.81
CQT[18]	MCG-Res2Net50	0.0520	1.78
Raw waveform[35]	Res-TSSDNet	0.0481	1.64
FFT[36]	SENet	0.0368	1.14
Raw waveform[37]	AASIST	0.0275	0.83
Wav2vec 2.0[38]	VIB	**0.0107**	**0.40**
LFCC	多阶 GMM-ResNet 融合	0.0353	1.16

型可以获得更具有说话人特性的特征表示, 但这种方式在前期需要使用大量的样本进行训练, 而本文只在官方提供的数据集中进行模型训练, 且挑战赛的规则不允许使用其他数据。这也说明本文所提模型的结果在 ASVspoof 2019 逻辑访问场景下具有一定竞争力。

5.3.2 ASVspoof 2021 LA 场景下的实验

表 6 显示了 ASVspoof 2021 LA 场景下基线系统、512 阶 GMM-ResNet 模型和多阶 GMM-ResNet 融合模型的实验结果。具体来说, 在不使用数据增强的情况下, 本文多阶 GMM-ResNet 融合模型在评估集上的 min t-DCF 指标分别比基线系统 LFCC+LCNN 的性能相对升高了 2.4%, 但是 EER 相对降低了 19.98%。在使用数据增强的情况下, 多阶 GMM-ResNet 融合模型在评估集上的 min t-DCF 和 EER 指标比 512 阶 GMM-ResNet 模型的性能分别相对降低了 4.2%和 17.5%; 比基线系统 LFCC+GMM 的性能分别相对降低了 57.3%和 87.1%;且比基线系统 LFCC+LCNN 的性能分别相对降低了 28.6%和 73.0%, 这说明本文采用的 RawBoost 数据增强方法非常适合本文所提出的模型, 在检测性能上提升有显著提升。

表 7 显示了多阶 GMM-ResNet 融合模型与目前主流的系统在 ASVspoof 2021 LA 数据集上的性能比较, 其中对比系统的性能指标结果均来自原论文, 且 MSTFT+LCNN 系统为 ASVspoof 2021 挑战赛 LA 场景下最佳单模型系统。RawNet2 模型和 AASIST 模型同样采用 RawBoost 数据增强的方式, 其中 AASIST 系统在评估集上优于本文所提的系统, 但此系统使用了无监督预训练模型 wav2vec2.0[12]作为特征提取器, 与挑战赛的规则不一致。多阶 GMM-ResNet 融合模型的实验结果在目前已发表文献中仅次于 LCNN 和 AASIST 系统, 这也说明了本文所提模型在 ASVspoof 2021 LA 场景下同样具有竞争力。

表 6 多阶 GMM-ResNet 融合模型在 ASVspoof 2021 LA 数据集上的实验结果

Table 6 Experimental results of the multi-order GMM-ResNet fusion model on the ASVspoof 2021 LA dataset

特征	模型	数据增强	min t-DCF	EER/%
CQCC	GMM[13]	-	0.4794	15.62
LFCC	GMM[13]	-	0.5758	19.30
LFCC	LCNN[13]	-	0.3445	9.26
Raw	RawNet[13]	-	0.4257	9.50
LFCC	GMM-ResNet	-	0.3800	8.49
	多阶 GMM-ResNet 融合	-	0.3527	7.41
	GMM-ResNet	✓	0.2568	3.03
	多阶 GMM-ResNet 融合	✓	**0.2459**	**2.50**

表 7 多阶 GMM-ResNet 融合模型与其他主流模型在 ASVspoof 2021 LA 数据集的实验结果比较

Table 7 Comparison of experimental results between the multi-order GMM-ResNet fusion model and other mainstream models in the ASVspoof 2021 LA dataset

特征	模型	数据增强	min t-DCF	EER/%
Raw[28]	RawNet2	✓	0.3099	5.31
Wav2vec 2.0[38]	VIB	-	-	4.92
Wav2vec 2.0[26]	W2V2	✓	-	3.54
MSTFT	LCNN[30]	✓	-	2.21
Wav2vec 2.0[12]	AASIST	✓	**0.2066**	**0.82**
LFCC	多阶 GMM-ResNet 融合	✓	0.2459	2.50

6 结论

本文提出了一种多阶 GMM-ResNet 融合模型进行语音伪造检测, 该模型主要包含了多阶 LGP 特征融合和 MFA-ResNet 模型两部分。首先, 本文对不同阶 LGP 特征进行了多阶特征融合。由于不同阶 LGP 特征之间包含的信息具有相关性, 因此本文选取了 128 阶, 256 阶和 512 阶三种阶数的 LGP 特征进行加权融合, 促进不同阶特征之间的信息交换。此外, 本文提出 MFA-ResNet 模型, 将每个 ResNet Block 块输出的特征进行聚合得到更全面的特征信息。最后, 本文提出多阶 GMM-ResNet 融合模型, 它将多阶 LGP 特征融合和 MFA-ResNet 整合在一起。实验在 ASVspoof 2019 和 ASVspoof 2021 挑战赛提供的 LA 数据集上进行, 实验结果验证了多阶 GMM-ResNet 融合模型的有效性。

参考文献

[1] Snyder D, Garcia-Romero D, Sell G, et al. X-Vectors: Robust DNN Embeddings for Speaker Recognition[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 5329-5333.

[2] Chen C, Rong Y F, Ji C Q, et al. Speaker Verification Method Based on Deep Information Divergence Maximization[J]. *Journal on Communications*, 2021, 42(7): 231-237.
(陈晨, 彤娅峰, 季超群, 等. 基于深层信息散度最大化的说话人确认方法[J]. *通信学报*, 2021, 42(7): 231-237.)

[3] Lee K A, Ma B, Li H. Speaker Verification Makes Its Debut in Smartphone[J]. *IEEE signal processing society speech and language technical committee newsletter*, 2013.

[4] Jelil S, Shrivastava A, Das R K, et al. SpeechMarker: A Voice Based Multi-Level Attendance Application[C]. *The 20th Annual Conference of the International Speech Communication Association*, 2019: 3665-3666.

[5] Tao J H, Fu R B, Yi J Y, et al. Development and Challenge of Speech Forgery and Detection[J]. *Journal of Cyber Security*, 2020, 5(2): 28-38.
(陶建华, 傅睿博, 易江燕, 等. 语音伪造与鉴伪的发展与挑战[J]. *信息安全学报*, 2020, 5(2): 28-38.)

[6] Sisman B, Yamagishi J, King S, et al. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning[J]. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, 2020, 29: 132-157.

[7] Wu Z Z, Evans N, Kinnunen T, et al. Spoofing and Countermeasures for Speaker Verification: A Survey[J]. *Speech Communication*, 2015, 66: 130-153.

[8] 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.

[9] Sahidullah M, Kinnunen T, Hanilçi C. A Comparison of Features for Synthetic Speech Detection[C]. *Interspeech 2015*, 2015: 2087-2091.

[10] 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.

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

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

[14] Lavrentyeva G, Novoselov S, Tseren A, et al. STC Antispoofing Systems for the ASVspoof 2019 Challenge[C]. *Interspeech 2019*, 2019: 1033-1037.

[15] 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.

[16] Alzantot M, Wang Z Q, Srivastava M B. Deep Residual Neural Networks for Audio Spoofing Detection[C]. *Interspeech 2019*, 2019: 1078-1082.

[17] Kwak I Y, Kwag S, Lee J, et al. ResMax: Detecting Voice Spoofing Attacks with Residual Network and Max Feature Map[C]. *2020 25th International Conference on Pattern Recognition*, 2021: 4837-4844.

[18] Li X, Wu X X, Lu H, et al. Channel-Wise Gated Res2Net: Towards Robust Detection of Synthetic Speech Attacks[C]. *Interspeech 2021*, 2021: 4314-4318.

[19] Lai C I, Chen N X, Villalba J, et al. ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks[C]. *Interspeech 2019*, 2019: 1013-1017.

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

[21] Wang C L, Yi J Y, Tao J H, et al. Global and Temporal-Frequency Attention Based Network in Audio Deepfake Detection[J]. *Journal of Computer Research and Development*, 2021, 58(7): 1466-1475.
(王成龙, 易江燕, 陶建华, 等. 基于全局-时频注意力网络的语音伪造检测[J]. *计算机研究与发展*, 2021, 58(7): 1466-1475.)

[22] Lei Z C, Yan H, Liu C H, 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*, 2022: 6377-6381.

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

[24] Zhang Y, Lv Z Q, Wu H B, et al. MFA-Conformer: Multi-Scale Feature Aggregation Conformer for Automatic Speaker Verification[C]. *Interspeech 2022*, 2022: 306-310.

[25] Jung J W, Kim Y, Heo H S, et al. Pushing the Limits of Raw Waveform Speaker Recognition[C]. *Interspeech 2022*, 2022: 2228-2232.

[26] 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.

[27] Jung J W, Heo H S, Yang I H, et al. Avoiding Speaker Overfitting in End-to-End DNNS Using Raw Waveform for Text-Independent Speaker Verification[C]. *Interspeech 2018*, 2018: 3583-3587.

[28] 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.

[29] Chen T X, Kumar A, Nagarsheth P, et al. Generalization of Audio Deepfake Detection[C]. *The Speaker and Language Recognition Workshop*, 2020: 132-137.

[30] Tomilov A, Svishchev A, Volkova M, et al. STC Antispoofing

Systems for the ASVspoof 2021 Challenge[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 61-67.

[31] Chen T X, Khoury E, Phatak K, et al. Pindrop Labs' Submission to the ASVspoof 2021 Challenge[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 89-93.

[32] Das R K. Known-Unknown Data Augmentation Strategies for Detection of Logical Access, Physical Access and Speech Deepfake Attacks: ASVspoof 2021[C]. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021: 29-36.

[33] Sadjadi S O, Slaney M, Heck L. MSR Identity Toolbox v1.0: A matlab toolbox for speaker recognition research[J]. *Speech and Language Processing Technical Committee Newsletter*, 2013, 1(4): 1-32.

[34] Zhang Y, Jiang F, Duan Z Y. One-Class Learning towards Synthetic Voice Spoofing Detection[J]. *IEEE Signal Processing Letters*, 2021, 28: 937-941.

[35] Hua G, Teoh A B J, Zhang H J. Towards End-to-End Synthetic Speech Detection[J]. *IEEE Signal Processing Letters*, 2021, 28: 1265-1269.

[36] Zhang Y X, Wang W C, Zhang P Y. The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System[C]. *Interspeech 2021*, 2021: 4279-4283.

[37] 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.

[38] Eom Y, Lee Y, Um J S, et al. Anti-Spoofing Using Transfer Learning with Variational Information Bottleneck[C]. *Interspeech 2022*, 2022: 3568-3572.

曹明明 于 2020 年在徐州工程学院计算机科学与技术专业获得工学学士学位。现在江西师范大学计算机信息工程学院计算机技术专业攻读硕士学位, 研究领域为语音伪造检测。研究兴趣包括说话人识别、语音信号处理。Email:cmm0807@jxnu.edu.cn

雷震春 于 2006 年在浙江大学计算机科学与技术专业获得博士学位。现任江西师范大学副教授, 硕士生导师。CCF 会员。研究领域为说话人识别、语音信号处理。Email: zhenchun.lei@hotmail.com

杨印根 江西师范大学教授, 硕士生导师。研究领域为说话人识别、智能信息处理。Email: yyg1999@sina.com

周勇 江西师范大学副教授, 硕士生导师。CCF 会员。研究领域为智能信息处理、机器学习。Email: zhouyong@jxnu.edu.cn