基于One-Class学习的鲁棒音频真伪识别

梁子琪[1,2]，张旭龙[1]，王健宗[1]，肖京[1]
1. 平安科技（深圳）有限公司，广东 深圳 518063；
2. 中国科学技术大学先进技术研究院，安徽 合肥 230026

摘要
深度伪造技术对社会经济、政治稳定和社会安全构成了严重威胁，而深度伪造中，语音伪造技术被广泛应用于电话诈骗、舆论操控等危害性活动中。近年来，随着深度学习技术的应用，语音合成和语音转换技术飞速进步，已经能够生成以假乱真的语音，足以欺骗机器和人类。针对语音伪造技术的危害，目前已经有许多语音欺骗检测技术来提高说话人验证系统的可靠性。然而，现有方法往往依赖于已知攻击类型的先验知识，在面对未知攻击类型的先验知识时，其泛化能力受到限制。基于One-Class学习构建了一个语音欺骗检测系统，通过为真实语音建立严格的决策边界，将边界外的样本判定为伪造语音，从而增强了模型的泛化能力。此外，针对伪造语音数据稀缺的问题，引入具有更强通用性和鲁棒性的自监督模型Wav2vec2进行特征提取，进一步提高了模型在面对未知类型先验知识攻击时的识别准确率。实验结果表明，提出的方法在保证良好语音鉴伪性能的同时，减少了CM系统对下游ASV系统的潜在干扰，有效解决了伪造语音数据稀缺和模型泛化能力不足的问题。

关键词
伪造检测；One-Class学习；自监督学习

中图分类号：TP391　　文献标志码：A　　doi：10.11959/j.issn.2096-0271.2025031

Robust audio authenticity detection based on One-Class learning

LIANG Ziqi[1,2], ZHANG Xulong[1], WANG Jianzong[1], XIAO Jing[1]
1. Ping An Technology (Shenzhen) Co. , Ltd. , Shenzhen 518063, China
2. University of Science and Technology of China, Institute of Advanced Technology, Hefei 230026, China

Abstract
Deepfake technology poses a serious threat to social economy, political stability, and social security. Among them, voice forgery technology is widely used in harmful activities such as phone scams and public opinion manipulation. In recent years, with the application of deep learning technology, speech synthesis and voice conversion technology have advanced rapidly, and can generate fake voices which are enough to deceive both machines and humans. In response to the harm of voice forgery, many voice deception detection technologies have emerged to improve the reliability of speaker verification systems. However, existing methods often rely on prior knowledge of known attack types, and their generalization ability is limited when they faced with unknown attack types of prior knowledge. We built a voice

deception detection system based on One-Class learning, which enhanced the generalization ability of the model by establishing strict decision boundaries for real voices and judging samples outside the boundaries as fake voices. In addition, to address the scarcity of fake voice data, the more versatile and robust self-supervised model Wav2vec2 was introduced for feature extraction, further improving the recognition accuracy of the model when faced with unknown attacks. Experimental results show that the proposed method can not only ensure good voice anti-spoofing performance, but also reduce the potential interference of the CM system to the downstream ASV system, effectively solving the problems of scarce fake voice data and insufficient generalization ability of the model.

Key words
spoof detection, One-Class learning, self-supervised learning

0 引言

随着生物识别技术的快速发展，生物识别技术在现实生活中的应用日益广泛，基于生物特征的身份认证技术在保障网络数据安全和便捷身份验证方面发挥越来越重要的作用，指纹识别、人脸识别、声纹识别等技术不断涌现，如科大讯飞的声纹认证打卡和腾讯微信的声纹登录等，这些技术在为用户带来便利的同时，也引发了很多亟待解决的数据安全问题。

自动说话人验证（automatic speaker verification，ASV）系统因其非接触、易接受、成本低、伪造难等优点，已经被广泛应用在各种领域[1]。其利用声学特征和算法对输入人声进行模式识别和匹配，验证输入的说话人语音是否为合法用户的声音，在生物识别认证中有至关重要的作用。ASV 系统的两阶段流程如图 1 所示，说话人注册阶段通过离线训练和在

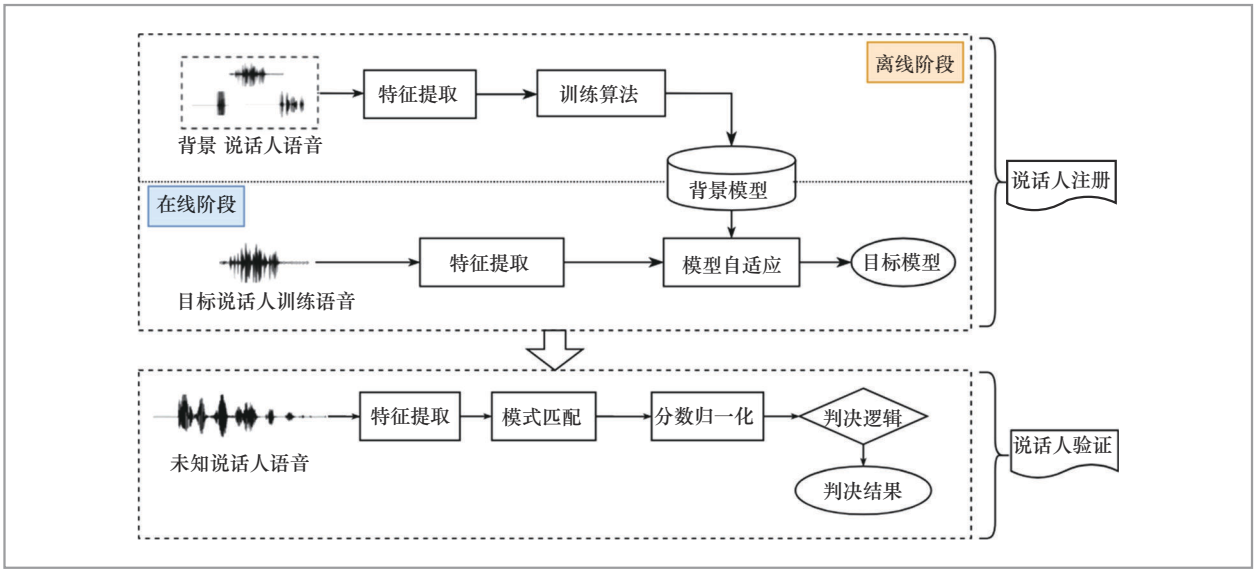


图1 ASV系统的两阶段流程

线模型自适应得到目标模型；说话人验证阶段将提取的未知说话人语音特征与模型相匹配，得到最终的判决结果。

然而，ASV系统很有可能被伪造音频欺骗，从而导致ASV系统误判，并且伪造语音不仅能够让ASV系统无法判断，甚至让人耳都无法判断音频的真伪。随着人工智能和语音欺骗技术的发展[2]，ASV系统在安全性方面遭到了严峻的挑战。检测输入ASV系统的语音真实性，防止欺骗语音通过ASV系统的验证以提高ASV系统的安全性，是近年来语音领域研究的一个热点[3-4]。目前常见的语音欺骗攻击方法包含语音转换（voice conversion，VC）[5]、文本到语音（text-to-speech，TTS）合成[6]、语音重放（预先录制的音频）[3]、语音模仿等，其中，合成语音攻击包括TTS合成和VC在内，对ASV系统的威胁越来越大[1]。

ASV系统防御欺骗性攻击的技术由于缺乏统一的策略评估标准，一直进展缓慢。自2015首届自动说话人验证欺骗与对策挑战赛的创建，学术界已经逐渐确立了一套通用的欺骗对策评估框架[7]。自此，为了提高自动说话人验证系统的欺骗鲁棒性，学术界出现了许多为抵御欺骗攻击而开发的独立的反欺骗模块。欺骗性语音不仅可以使ASV系统误判，甚至让人耳都无法判断声音的真伪[8]。受限于语音欺骗检测数据集，语音模仿需要较专业的模仿者录制，因此研究重点通常放在TTS、VC以及语音重放这3种语音欺骗检测方法。

目前使用最多的语音欺骗检测技术是一个与ASV系统独立的、互不关联的欺骗检测系统。一般情况下是将对策（countermeasure，CM）系统和ASV系统进行级联，系统是同时进行欺骗检测[9]和说话人验证的。具体来说，首先对输入欺骗检测系统中的语音样本进行安全性验证，通过欺骗检测系统，可以对输入语音进行判定，判定其是真实语音的样本还是伪造语音的样本。同时，音频输入ASV系统中进行说话人认证，判别说话人身份是目标说话人还是冒充者。典型的语音欺骗检测系统如**图2**所示。

当前的语音欺骗检测系统只能单独检测一种语音欺骗，如单独检测TTS与语音转换的欺骗攻击，或者是单独检测语音回放欺骗。因此如果语音的欺骗方法未知，那么需要开发一个对模型之前未学习过的攻击类型同样具有良好检测能力的CM系统。同时，为了验证说话人身份是目标说话人还是冒充者，以及评估设计的CM系统对ASV系统可靠性的影响，本文将ASV系统与CM系统级联，根据二者共同输出分数进行评估。自动说话人验证欺骗与对

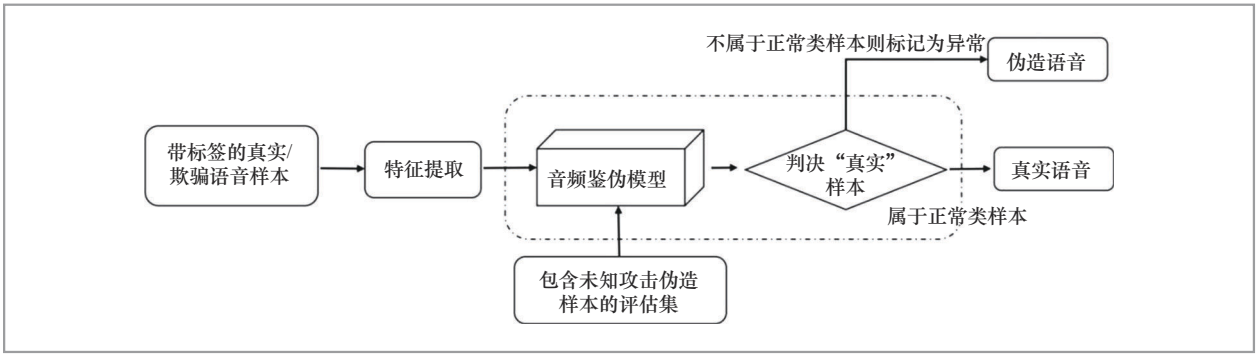


图2 典型的语音欺骗检测系统

策挑战赛一直为反欺骗攻击的说话人验证的研究提供数据集和评估指标，为了提高自动说话人验证系统的鲁棒性，本文开发了独立的欺骗检测模块CM系统来检测语音欺骗攻击。

笔者在2019年的自动说话人验证欺骗与对策挑战赛的基础上，专注于合成伪造语音攻击的反欺骗，即将真实语音与分别使用TTS合成和VC算法生成的语音区分开来[10]。传统语音欺骗检测方法把注意力集中在特征工程，通过设计合适的人工特征使欺骗检测算法取得较好的性能。使用较多的是梅尔频率倒谱系数（mel-frequency cepstral coefficient，MFCC）特征、耳蜗滤波器倒谱系数（cochlear filter cepstral coefficient，CFCC）特征以及瞬时频率（instantaneous frequency，IF）特征。此外，还有在一些提交方案中性能较良好的线性预测倒谱系数（linear prediction cepstral coefficient，LFCC）特征、常数Q变换（constant-Q transform，CQT）特征以及常数Q倒谱系数（constant Q cepstrum coefficient，CQCC）特征等。一些研究表明，语音信号高频部分蕴含的特征信息有助于伪造音频检测的任务。当使用LFCC特征作为前端提取的特征时，其能量较集中在语音信号的高频部分，可以使欺骗检测系统取得良好的性能，因此前端部分采用的是提取语音LFCC特征。同时，在Wu等[11]提出的Light CNN（convolutional neural network）基础上，本文参考图像伪造检测领域中的One-Class学习思想，引入了Am-Softmax分类器和OC-Softmax分类器来学习不同的决策分类边界，为真实语音样本训练出一个严格的决策边界，在此决策边界之外的非目标类数据则都被判决为伪造语音样本，并引入了ResNet作为特征提取器与Light CNN进行对比。实验表明，在2019年的自动说话人验证欺骗与对策挑战赛LA（logical access）数据集上，相较于前端提取LFCC特征加后端GMM（gaussian mixture model）分类器的基线模型，本文提出的方案取得了良好的实验效果，其中性能最好的单模型方案在等错误率（equal error rate，EER）和串联检测成本函数（tandem detection cost function，t-DCF）上分别降低了5.56%和0.145。

此外，面对现实中缺少带标签的伪造音频数据的低资源情况，使用通用性鲁棒性强的Wav2vec2特征代替前端语音LFCC特征，并改进了后端模型架构。实验结果表明，采用Wav2vec2特征代替LFCC特征的单模型方案相较于官网基线模型，性能均有提升。本文通过Optuna调参的融合模型实现了接近SOTA（state-of-the-art）模型的鉴伪能力和最优的模型鲁棒性，模型融合对下游ASV系统的干扰和影响都降到了最低。

综上，本文主要贡献如下。

（1）本文实验探究并分析了One-Class学习算法在语音欺骗检测任务中的适用性，并设计了前端提取语音LFCC特征与OC-Softmax分类器相结合的语音欺骗检测系统，与ASV系统级联后，评估CM系统鉴伪性能及其对ASV系统可靠性的影响。

（2）本文针对识别未知攻击类型的伪造音频的需求和目前缺少带标签的伪造音频数据的低资源现状，引入更具有通用性和鲁棒性的Wav2vec2特征。本文针对自监督预训练模型，改进了后端分类器的模型架构，并进行实验分析。相较于基准模型，本文提出的模型在2019年的自动说话人验证欺骗与对策挑战赛LA数据集上鉴伪率和准确率均有大幅度提升。

1 现有的语音欺骗检测方法

目前语音欺骗检测任务可以分为两个方向，比较典型的是传统的基于人工设计特征的机器学习方法，除此之外是基于深度学习搭建模型进行训练的方法。使用机器学习方法解决音频伪造检测问题，往往格外关注特征工程部分，需要事前得到用于语音欺骗检测的声学特征，然后使用性能良好的分类器，对从语音文件中提取的声学特征（如MFCC或Spectrogram）进行分类判决。在语音欺骗检测任务中，常见的基于传统机器学习的方法有基于GMM模型[12]或SVM（support vector machine）模型[13]的后端分类模型、概率线性判别分析（probabilistic linear discriminant analysis，PLDA）[14]、用于类似I-Vector嵌入式特征的常用打分策略等[15]。此外，还有将多个模型拼接的集成学习方法[16]。与传统机器学习更加关注特征工程不同的是，基于深度学习的语音欺骗检测系统直接将原始数据输入模型中自动提取学习特征。通常对于基于深度学习的语音欺骗检测模型，语音信号特征被输入神经网络以计算输入语音的嵌入向量。训练模型的目的是学习一个嵌入空间以更好地区分真实语音和伪造语音，嵌入层将进一步用于对输入音频属于真实语音的置信度进行打分。

1.1 基于GMM的语音欺骗检测方法

GMM通过将多个单个的高斯分布线性加权进行组合。GMM可以通过多维高斯分布来混合表示，近似表示出任意形状的分布。许多十分复杂的非线性问题可以把其建模为可通过多个GMM来拟合分布的问题。研究表明，在ASV系统中，利用GMM强大的数据拟合能力可以拟合目标说话人的身份模型[12]；在语音欺骗检测中，通过GMM对伪造语音样本进行聚类可以将聚类后得到的簇作为目标类或非目标类，或者将聚类得到的簇判定为同一类欺骗攻击类型，从而实现欺骗攻击类型的分类识别[17-19]。GMM通过上述方法分别拟合真实语音和欺骗语音两个模型，GMM概率密度函数如下。

$$P(x)=\sum_{i=1}^{C}\omega_i p\left(x|\mu_i,\boldsymbol{\Sigma}_i\right) \tag{1}$$

其中，C为高斯模型的个数，ω_i，μ_i，$\boldsymbol{\Sigma}_i$分别为每个高斯的权重、均值和协方差矩阵。假设一个语音样本的特征矢量矩阵为$\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T\}$，则该矩阵相对于GMM的对数似然得分为每个特征矢量$\boldsymbol{x}_i$相对于该模型（真实语音或欺骗语音）的对数似然得分之和。因此通过对所有特征向量得分取平均，可以得到最终特征矢量相对于GMM的似然得分。

1.2 基于SVM的语音欺骗检测方法

支持向量机作为一种经典二分类算法，非常适合于伪造语音检测任务[2]，因为从本质上来说，语音欺骗检测任务最终输出的标签只有“bonafide”和“spoof”两类，可看作一个二分类问题。SVM训练阶段寻找可以将真实语音和欺骗语音特征完全隔离的超平面，并且使真实类或虚假类的数据距离超平面的距离最大，同时使其各自类内距离最小[20]。

使用线性SVM作为语音欺骗检测系统，较典型的是Sergey等[15]使用具有线性核的SVM分类器，在归一化的I-Vector空间进行训练。由于SVM的分类效果和训

练速度较好，在自动说话人验证欺骗与对策挑战赛中已经被广泛作为分类器，并取得良好的效果。

1.3 基于深度学习的语音欺骗检测方法

语音欺骗检测方法除了上述基于传统机器学习的方法，还有基于深度学习的方法，其中较典型的有CNN（convolutional neural network）和RNN（recurrent neural network）[21]、以CNN和RNN的改进模型为主流的模型、使用端到端方法进行合成语音欺骗检测[22-23]，以及以生成对抗网络（generative adversarial network，GAN）为架构的深度神经模型。其中，GAN在区分复杂的非线性特征的同时，极大地提高了针对复杂样本的分类准确性。

例如，在2015年的自动说话人验证欺骗与对策挑战赛中，人们使用全连接神经网络（fully connected neural network，DNN）对提取的声学特征进行分类判别[24]。在2017年的自动说话人验证欺骗与对策挑战赛中人们使用CNN、BiRNN（bidirectional recurrent neural network）[25]、LSTM（long-short term memory）[26]等在比赛中均取得了优异的成绩，其证明了基于深度神经网络的方法在语音欺骗检测中的适用性。Zhang等[27]研究了用于反欺骗的深度学习模型，并证明CNN和RNN的组合可以提高系统的鲁棒性。Chen等[28]提出了一种采用具有大余弦损失的ResNet的方法，并应用频率掩模增强语音伪造检测系统。Gomez-Alanis等[29]采用轻量卷积门控RNN来改善语音欺骗攻击检测的长期依赖性。Wu等[30]提出了一种基于特征泛化的轻量化CNN系统，在伪造检测合成攻击方面优于其他单一系统。Aravind等[31]探索了使用ResNet的迁移学习方法。Tak等[32]将RawNet2结构引入，实现端到端的语音欺骗检测，并获得了不错的实验效果。Fu等[33]通过定义通用架构FastAudio对前端进行分类，用可学习的层替换了固定的滤波器组，以更好地适应欺骗检测任务。Jung等[34]利用图注意力层同时对时域和频域进行建模，进一步提高了欺骗检测系统的性能。虽然基于深度学习的语音欺骗检测已经取得了很大进展，但现有方法通常会在测试阶段泛化到看不见的欺骗攻击[35]。

2 基于One-Class学习的语音欺骗检测系统

现有的一些欺骗检测方法虽然能够做到语音真伪识别，但在模型推理阶段，面对未知攻击类型的伪造语音样本，其识别性能往往会大大降低，模型的泛化能力较差。这是因为大部分的语音欺骗检测方法把语音鉴伪问题建模为真实语音和伪造语音的二分类问题，即均假设了真实语音或伪造语音的训练集和测试集具有相似的分布。但现实情况是，随着语音合成技术的成熟和发展，越来越多未知攻击类型的合成欺骗语音涌现出来，这与之前研究中的假设相矛盾。

2.1 语音欺骗检测系统

针对以上问题，本文设计了一种语音欺骗检测系统作为CM系统，与ASV系统并联判断和使用。CM系统的输出分数是对输入语音是真实语音还是伪造合成语音进行决策，并联的ASV系统则是对输入的音频发音者是目标说话人还是冒充者进行验证判断，最终语音欺骗检测系统的性能

不是由单个子系统决定的，而是由组合的级联系统决定的。各子系统使用各自的评估指标进行训练，组合的级联系统可以使用串联检测成本函数进行性能评估。语音欺骗检测和说话人身份验证级联如**图3**所示，本文从语音中提取LFCC特征或者使用预训练模型提取自监督语音特征，然后分别通过CM系统和ASV系统进行语音真伪检测和说话人身份验证，最后根据CM分数和ASV共同判断。

语音欺骗检测系统的前端模块对输入语音波形进行一系列操作后得到LFCC特征，并将其用作下游分类任务的输入数据。LFCC特征提取流程如**图4**所示。

基于Light CNN构建模型中的特征学习模块，使用了最大特征映射MFM（max feature map）的操作。与传统使用大量隐藏层神经元的方法不同，传统方法是通过近似凸函数的分布来实现特征学习的，而MFM通过抑制少量神经元激活的原理，使MFM可以减少CNN模型的参数大小，加快模型训练的速度[22]。以MFM1/2为例，最大特征映射通过降低特征图的通道数来减少模型参数，MFM可以定义为式（2）。

$$\mathrm{MFM}_{ij}^{p}=\max(\mathrm{fm}_{ij}^{p},\ \mathrm{fm}_{ij}^{p+\frac{C}{2}}) \qquad (2)$$

式（2）中输入特征图fm量的格式是$H\times W\times C$，由高度、宽度以及通道数三者构成，FM是输出特征图，i和j分别表示时频域，p表示通道索引。

图5所示为使用Light CNN-9作为对前端LFCC特征进行进一步特征提取的模型，其中包含4层卷积层、4层最大池化层以及最大特征映射层等。Light CNN-9模型结构如**图6**所示。

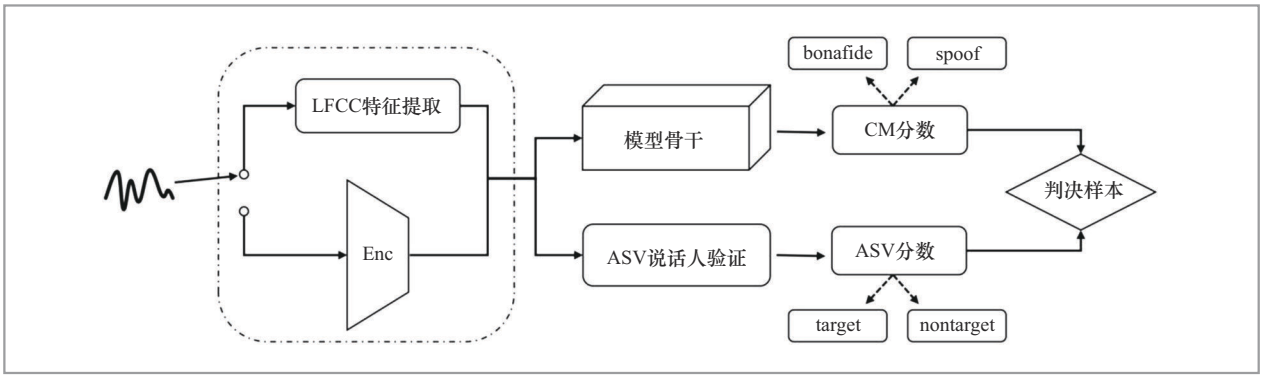


图3 语音欺骗检测和说话人身份验证级联

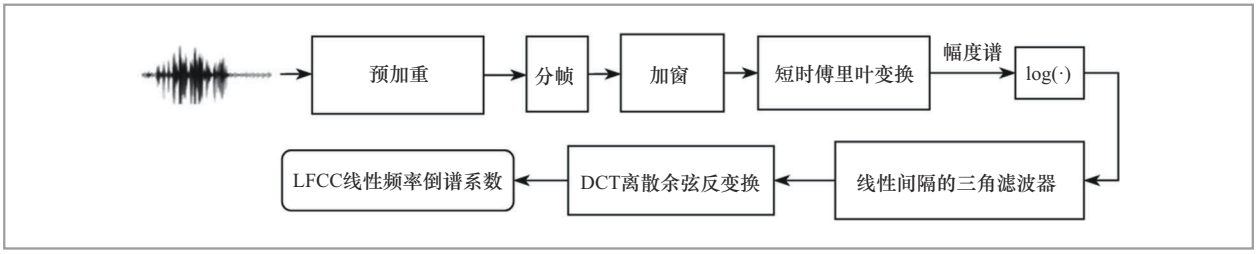


图4 LFCC特征提取流程

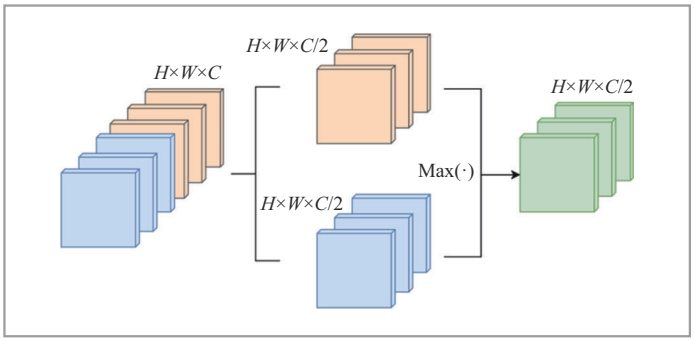


图5 MFM最大特征激活

本文将得到的LFCC特征输入后端分类模型中，将其输出作为CM系统的输出分数，得到CM系统的输出分数的同时，也得到了语音样本的真伪识别结果。ASV系统对语音进行说话人身份的验证，对发音者的身份进行鉴别。

2.2 One-Class分类器

语音欺骗检测任务在模型推理的过程中可能会遇到许多未知攻击类型的语音样本，这就导致了对于伪造样本来说，其训练集和测试集的数据特征分布极不匹配，训练集中包含的语音攻击类型往往有限，单纯以二分类问题看待语音鉴定伪问题在模型推理的时候，面对许多新样本往往性能会下降[23]。而这种单分类的任务场景实际上非常适合One-Class分类的思想。其只关注输入样本与已有样本的相似度或匹配程度，对于未知攻击类型的样本则统一排除。

传统语音欺骗检测任务使用的二分类损失函数的表示如下。

$$\mathcal{L}=-\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{\boldsymbol{\omega}_{y_i}^{\mathrm{T}}\times\boldsymbol{x}_i}}{e^{\boldsymbol{\omega}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i}+e^{\boldsymbol{\omega}_{(1-y_i)}^{\mathrm{T}}\times\boldsymbol{x}_i}} \tag{3}$$

其中，$\boldsymbol{x}_i$是第i个语音样本的嵌入向量，M表示计算样本数量，$y_i\in\{0,1\}$是对应的标签。后来通过引入角余量，在最大化目标类和非目标类之间距离的同时，最小化各类内数据间的距离，将决策边界变得更加紧凑。

$$\mathcal{L}_{\mathrm{AM-Softmax}}=-\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{\xi(\hat{\boldsymbol{\omega}}_{y_i}^{\mathrm{T}}\hat{\boldsymbol{x}}_i-m)}}{e^{\xi(\hat{\boldsymbol{\omega}}_{y_i}^{\mathrm{T}}\hat{\boldsymbol{x}}_i-m)}+e^{\xi\hat{\boldsymbol{\omega}}_{1-y_i}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}=$$

$$\frac{1}{M}\sum_{i=1}^{M}\log\left(1+e^{\xi(m-(\hat{\boldsymbol{\omega}}_{y_i}-\hat{\boldsymbol{\omega}}_{1-y_i}))^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right) \tag{4}$$

其中，ξ是比例因子，m是余弦相似度的余量，和$\hat{\boldsymbol{\omega}}$，$\hat{x}$分别为归一化$\boldsymbol{\omega}$和$\boldsymbol{x}$。根据上述的二分类损失函数，目标类和非目标类的嵌入向量都趋向于围绕两个相反的方向收敛，即分别为$\boldsymbol{\omega}_0-\boldsymbol{\omega}_1$和$\boldsymbol{\omega}_1-\boldsymbol{\omega}_0$，其中，$\boldsymbol{\omega}_0$是真实语音代表的目标类，$\boldsymbol{\omega}_1$是伪造语音代表的非目标类。式（4）中AM-Softmax函数，目标类和非目标类的嵌入特征都设置为相同的角余量大小m，m值越大，训练出的决策边界越紧凑。

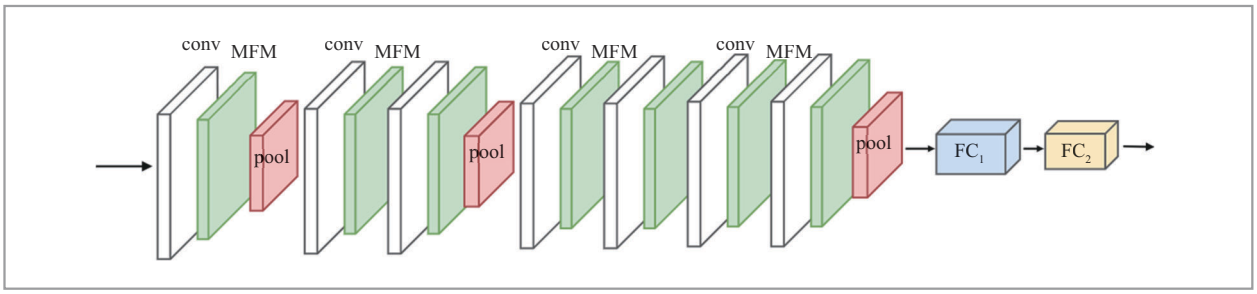


图6 Light CNN-9模型结构

在现实场景中，非目标类即伪造语音的特征分布与目标类的差异较大，若为二者训练相同的决策边界，往往会导致模型过拟合到训练集中的已知攻击类型，分类结果均判定为已知类型。因此，本文考虑为真实样本和伪造样本引入不同的角余量 m，训练不同的决策边界以更好地识别真实语音样本，隔离伪造语音样本，设计的损失函数如下。

$$\mathcal{L}_{oc}=\frac{1}{M}\sum_{i=1}^{M}\log\left(1+e^{\xi(m_{y_i}-\widehat{\omega}_0\cdot\hat{x}_i)(-1)^{y_i}}\right) \quad (5)$$

在OC-Softmax中只使用了一个权重 ω_0，其代表目标类嵌入向量的优化方向。二者与AM-Softmax类似，同样进行归一化。然后引入了两个不同的角余量 m_0、m_1，$m_0, m_1 \in [-1, 1]$，分别用于真实语音和伪造语音的分类，其中 $m_0 > m_1$。

如**图7**所示，OC-Softmax中只有一个表示目标类真实语音的收敛方向，同时用一个角度 θ_i 表示收敛方向和语音特征向量间的角度。当 $y_i=0$，即样本为目标类时，通过设定 m_0 来进行约束决策边界，使 θ_i 小于 $\arccos(m_0)$。$\arccos(m_0)$ 的值较小可以使目标类集中在权重向量 ω_0 周围，使真实语音样本的特征分布较紧凑集中。当 $y_i=1$ 时，设置 m_1 使 θ_i 大于 $\arccos(m_1)$。$\arccos(m_1)$ 的值较大可以将非目标类的数据远离收敛方向，以此实现为目标类和非目标类样本训练不同决策边界的目的。

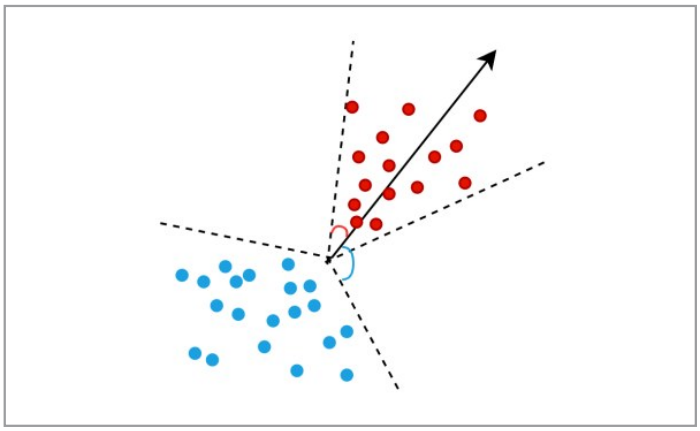


图7 OC-Softmax特征嵌入向量分布

2.3 基于Wav2vec2的欺骗检测系统结构

传统音频伪造检测系统的前端模块一种是使用DSP（digital signal processing）算法提取语音特征，另一种是使用可训练的DNN和反欺骗数据库以监督学习的方法进行训练。在由带标签的伪造音频和真实音频构成的训练集上训练过的CM策略系统，其在测试阶段面对未知攻击类型的语音样本，往往欺骗检测性能会产生不同程度的降低。针对这个问题，一方面可以通过对音频样本进行加噪声等方法增强模型的鲁棒性，另一方面可以通过训练一个有监督的DNN前端模块，然而这需要大量的带标签的欺骗语音样本，但这往往费时费力。人工设计的特征（如MFCC、LFCC和CQCC等）在语音识别任务中工作良好，但对于语音欺骗检测任务来说，缺乏标记数据，难分辨出未知类型的伪造语音。因此，在缺乏标记数据的语音任务中，预训练特征成为一种更好获得语音区分特征的方法。

为解决以上问题，本文使用Wav2vec2[24]特征代替传统的声学特征，Wav2vec2是通过无标签数据预训练获得的，具有更通用、鲁棒性更强的特征向量表示，因此其可以提高模型对未知语音欺骗攻击的泛化能力。Wav2vec2模型框架如**图8**所示。

自监督预训练模型可看作由卷积神经网络构成的编码器和Transformer组成，编码器从输入语音波形 $x_{1:T}$ 中提取特征向量 $z_{1:N}$，而Transformer则将特征向量 $z_{1:N}$ 转

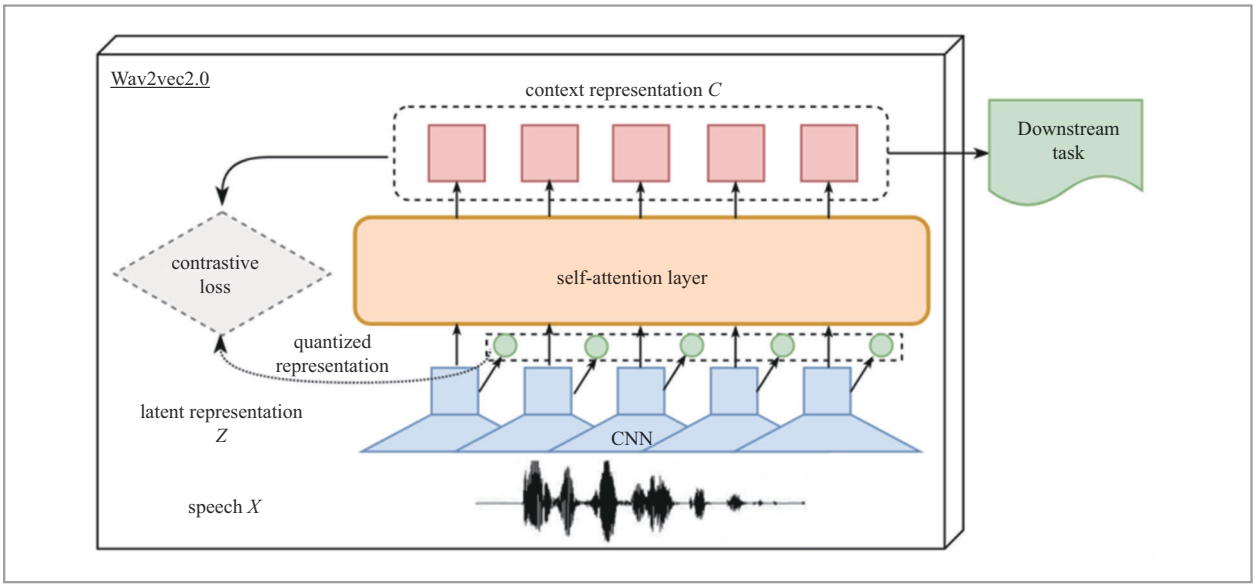


图8 Wav2vec2.0模型框架

换为包含序列全局信息的输出向量$\boldsymbol{\alpha}_{1:N}$。系统通过将自监督预训练模型的输出特征向量$\boldsymbol{\alpha}_{1:N}$输入欺骗检测系统后端，可以得到关于输入波形文件的策略分数。

图9展示了本文提出的基于Wav2vec2的语音欺骗检测系统。本文在LCNN的基础上添加了双向LSTM和一个全局平均池化层，用于和前端自监督预训练模型进行

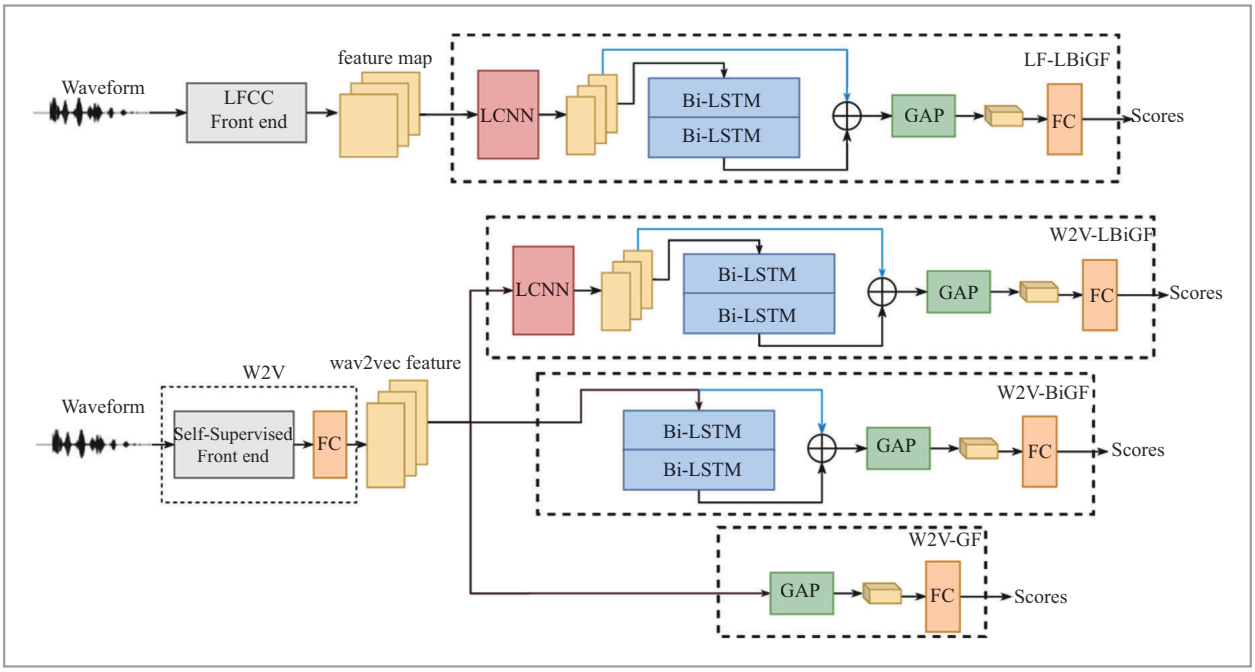


图9 基于Wav2vec2的语音欺骗检测系统

对比实验。当使用自监督预训练wav2vec2模型作为前端代替提取LFCC特征时，下游欺骗检测任务的模型进行了3种模型结构的实验：W2V-LBiGF、W2V-BiGF、W2V-GF。Wav2vec2.0后接一个全连接层调整模型的输出特征维度，将其调整到下游任务适合的输入特征维度上，通过全连接层的桥梁纽带作用，降低自监督预训练模型的输出维度，与语音欺骗检测任务的后端进行联合训练。之后，其将特征送入LCNN模块，LCNN中的每个卷积层都引入了最大特征激活图层MFM，将输入特征图分为两个部分，并通过竞争的方式保留值较大的一部分，通过MFM在保持一定性能的同时，还可以降低模型参数的大小，加快模型收敛速度。LCNN的输出特征向量输入两个双向长短时记忆网络层中，二者输入输出的特征图尺寸相同并在两个Bi-LSTM层上添加了一个残差连接。

基于自监督预训练模型的欺骗检测系统结构中，在模型的最后使用了全局平均池化代替其中部分的全连接层。全连接层的参数量很多，在训练的过程中，不可避免地会将一定的特征存储在全连接层的参数中，导致训练速度过慢。全局平均池化层可以将M个特征图降维成$1\times M$形状的特征图，若分类任务的类别数为N，再用N个1×1大小的卷积核将$1\times M$的特征图进行卷积运算，计算后得到$1\times N$的向量，以此可以等效一层全连接层。

W2V-BiGF在W2V-LBiGF的基础上删掉LCNN模块，前端自监督预训练模型Wav2vec2.0的输出向量直接输入双向长短时记忆网络层。W2V-GF则在W2V-BiGF模型的基础上，将Bi-LSTM模块删除，仅保留全局平均池化层（global average pooling，GAP）和全连接层（fully connected layer，FC）。

3 实验

3.1 数据集和预训练模型

本文使用2019年的自动说话人验证欺骗与对策挑战赛LA数据集作为实验的数据集，LA子集包括真实语音和伪造语音，伪造语音是由不同类型的TTS和VC欺骗攻击合成的。在LA数据集中，训练集和开发集包含真实语音和6种攻击类型（A01~A06）的伪造语音，6种攻击类型包括4种TTS算法合成和两种VC算法合成。

在评估集中，有11种未知攻击（A07~ A15、A17、A18），均是使用不同的TTS和VC算法生成的，由不同的TTS和VC攻击的组合构成。评估集另外还有两个特殊的攻击类型（A16、A19），二者与训练集中的两个攻击类型（A04、A06）使用相同的TTS/VC算法生成，但使用不同的数据进行训练。2019年的自动说话人验证欺骗与对策挑战赛LA数据集划分见**表1**。

自监督预训练模型Wav2vec2.0主要采用W2V-XLSR、W2V-Large1、W2V-Small，具体参数见**表2**。

3.2 训练设置

本文提取前60维的LFCC特征向量。

表1 2019年的自动说话人验证欺骗与对策挑战赛LA数据集

数据集	真实语音	伪造语音	
	音频数量/个	音频数量/个	攻击类型
训练集	2 580	22 800	A01-A06
验证集	2 548	22 296	A01-A06
测试集	7 355	63 882	A07-A19

表2 不同版本Wav2vec2.0预训练模型

Model	训练数据	参数量	输出维度
W2V-XLSR	LibriSpeech, CommonVoice, BABEL	317×10^6	1 024
W2V-Large1	Libri-Light	317×10^6	1 024
W2V-Small	Librispeech	95×10^6	768

音频采样率设置为16 kHz，帧长为320 ms，设置两帧起始点的时间差为帧移等于10 ms，帧跳数为160。

针对损失函数中的超参数，为AM-Softmax设置α=20，m=0.9。对于设计的OC-Softmax，本文设置α=20，m_0=0.9和m_1=0.2。使用Adam优化器，将β_1参数设置为0.9，将β_2参数设置为0.999，用来更新LCNN和ResNet模型中的权重。损失函数的参数使用SGD（stochastic gradient descent）优化器。一个批次大小设置为16，学习率设置为0.0001，epoch设置为50轮次，早停设置为5。

采用Wav2vec2预训练模型时，帧长设置改为20 ms，并采用512点的快速傅里叶变换FFT（fast fourier transform）运算，初始化学习率为3×10^{-4}，每隔10轮学习率减半。微调时小批量大小设置为8，学习率设为1×10^{-6}。微调时对Wav2vec2.0的隐藏特征不进行随机掩蔽。为了适配下游任务，前端自监督预训练模型后的全连接层输出维度设置为128。

3.3 评价指标

语音欺骗检测和身份验证级联系统的性能主要是通过等错误率（EER）和串联检测成本函数t-DCF构成。其中，t-DCF基于统计检测理论，语音欺骗检测和身份验证级联系统由ASV系统和CM系统并联组成，只需计算CM分数，并与2019年自动说话人验证欺骗与对策挑战赛组织者提供的ASV分数相结合。分数高表示真实样本，分数低表示欺骗攻击。CM系统的等错误率用来评估欺骗检测系统的性能，是衡量系统安全性和准确性的重要指标。EER是指当漏报率和误报率相等时的值，对应于CM欺骗检测系统的阈值τ_{EER}。

$$\mathrm{EER}=P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{EER}})=P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{EER}}) \tag{6}$$

其中，$P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{EER}})$表示当样本属于欺骗语音但CM系统判定分数大于设定的阈值τ_{EER}时的样本数量与所有欺骗语音样本总数相除的结果；$P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{EER}})$表示当样本属于真实语音样本但被CM系统判定分数小于等于设定的阈值τ_{EER}时的样本数量与所有真实语音样本总数相除的结果。

串联检测成本函数用来评估欺骗检测系统对ASV可靠性的影响，定义为式（7）。

$$\mathrm{t\_DCF}(\tau_{\mathrm{cm}})=C_0+C_1P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})+C_2P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) \tag{7}$$

其中，$P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})$和$P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})$分别代表当阈值为$\tau_{\mathrm{cm}}$时，CM系统的误拒率和误信率，二者的计算式如下。

$$P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})=\frac{N(\text{bonafide with CM score}\leqslant\tau_{\mathrm{cm}})}{N(\text{all bonafide})} \tag{8}$$

$$P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})=\frac{N(\text{spoof with CM score}>\tau_{\mathrm{cm}})}{N(\text{all spoof})} \tag{9}$$

t-DCF计算中C_0、C_1和C_2等参数的计算都依赖于ASV系统的错误率，其计算式如下。

$$C_0=\pi_{tar}C_{miss}P_{miss}^{asv}+\pi_{non}C_{fa}P_{fa}^{asv} \quad (10)$$

$$C_1=\pi_{tar}C_{miss}-(\pi_{tar}C_{miss}P_{miss}^{asv}+\pi_{non}C_{fa}P_{fa}^{asv}) \quad (11)$$

$$C_2=\pi_{spoof}C_{fa,spoof}P_{fa,spoof}^{asv} \quad (12)$$

其中，π_{tar}、π_{non}、π_{spoof}分别代表目标说话人先验概率、非目标说话人先验概率和欺骗攻击的先验概率；C_{miss}、C_{fa}、$C_{fa,spoof}$分别代表错误拒绝目标、错误接受非目标、错误接受的欺骗攻击的代价；P_{miss}^{asv}、P_{fa}^{asv}、$P_{fa,spoof}^{asv}$分别表示在指定ASV系统阈值时的ASV系统的目标错误拒绝率、非目标错误接受率、欺骗攻击错误接受率。

为了评估独立开发的语音欺骗检测系统的性能，本文记录了欺骗检测系统的输出分数，将其称为CM分数。CM分数用来衡量输入语音和真实语音之间的相似性。对于使用Softmax或AM-Softmax二分类分类器进行训练的CM系统，输出CM分数可以看作语音特征向量$\boldsymbol{x}_i$与权重向量$\boldsymbol{\omega}_0-\boldsymbol{\omega}_1$之间的余弦相似度，为真实语音和伪造语音同时训练具有相同角余量的决策边界，这也能从侧面衡量输入语音和真实语音之间的相似性，二者余弦相似度越低，二者的特征分布越相似。这种做法存在一个问题，如果为目标类和非目标类同时训练相同的决策边界，其会导致在模型推理阶段，CM系统在未知攻击类型的语音样本上性能较差。

本文针对上述问题设计了OC-Softmax单类分类器，其中语音欺骗检测系统的CM分数可以看作语音特征$\boldsymbol{x}_i$和真实语音代表的目标类ω_0之间的余弦相似度。这里只关注让模型学习真实语音样本的特征分布，得到针对目标类的决策边界。当模型推理时，如果给定语音样本的特征分布与系统之前学习过的语音特征分布相差较大，本文则将其判定为伪造语音。

3.4 实验结果与分析

本文为了验证语音LFCC特征相较于其他语音特征（如CQT、MFCC、Fbank特征等）的优越性，以及提出的OC-Softmax分类器和LCNN特征提取模块在处理语音欺骗检测任务中的有效性，对比实验分别选取了4种前端语音特征、3种分类损失函数（AM-Softmax、Sigmoid、OC-Softmax），特征提取模型同样选取了LCNN和ResNet两种网络进行了对比。

模型在验证集上的实验结果见**表3**，由**表3**可知，使用的3种损失函数，在LCNN和ResNet上训练的欺骗检测系统性能均高于基准模型。验证集上性能最好的LFCC-ResNet-OC-Softmax的实验方案，EER相对于基准模型降低了2.593%。在使用相同LCNN作为特征提取网络时，相对于传统的添加了角余量的二分类损失函数AM-Softmax，本文设计的OC-Softamx方案的EER降低了0.002。

对于测试集包含的所有攻击类型，并不都包含在训练集和验证集中，测试集还包含A07-A19等几种模型没有学习过的攻击类型。从**表4**可知，以LFCC特征为前端的4种方案均优于基线模型，EER最低的方案是LFCC-LCNN-OC-Softmax，其相较于基线模型和传统二分类损失函数的方法，EER分别降低了5.56%和0.143，其表示出设计的基于One-Class学习的损

表3 验证集实验结果

Dev Set	EER	t-DCF
Baseline	2.710%	0.066
LFCC-LCNN-AM-Softmax	0.393%	0.010
LFCC-LCNN-OC-Softmax	0.313%	0.008
LFCC-LCNN-Sigmoid	0.156%	0.004
LFCC-ResNet-OC-Softmax	**0.117%**	**0.003**

表4 测试集实验结果

Test Set	EER	t-DCF	Threshold
Baseline	8.090%	0.211	—
LFCC-LCNN-AM-Softmax	4.647%	0.082	0.807
LFCC-LCNN-OC-Softmax	**2.530%**	0.068	0.999
LFCC-LCNN-Sigmoid	2.869%	0.072	0.778
LFCC-ResNet-OC-Softmax	3.290%	**0.067**	0.999
MFCC-LCNN-OC-Softmax	6.783%	0.108	—
FBank-LCNN-OC-Softmax	3.116%	0.803	—
CQT-LCNN-OC-Softmax	2.931%	0.072	—
RawNet2[34]	**1.748%**	0.054	—
Ensemble	1.762%	**0.050**	—

失函数在模型推理阶段强大的泛化能力，在面对测试集中未知攻击类型的伪造语音仍具有一定的真伪识别能力。与此同时，本文还对比了不同前端语音特征对欺骗检测系统性能的影响，实验结果表明，相比于CQT、FBank和MFCC特征，LFCC特征的实验结果较好，最适合该任务。本文将LFCC-LCNN-OC-Softmax、LFCC-ResNet-OC-Softmax、CQT-LCNN-OC-Softmax，以及FBank-LCNN-OC-Softmax集成到一起，得到融合模型Ensemble，通过权重系数加权求和，并使用Optuna调参求得最优权重系数，最终实验结果与RawNet2结构的欺骗检测系统相比，不仅在EER指标上相接近，还降低了t-DCF值，其说明融合模型降低了CM系统对ASV系统可靠性的影响。

在整个级联系统中，对于并联的ASV系统是否影响说话人身份验证的性能，本文使用官方提供的评估指标t-DCF。t-DCF越低，则CM系统对ASV系统安全性和可靠性的影响越小。从**表3**可知，在t-DCF上，实验性能最好的是LFCC-ResNet-OC-Softmax，相对于使用LCNN特征作为特征提取网络，在保证一定的欺骗检测能力的同时，其t-DCF降低了约0.001，但从整体上看，同样使用OC-Softmax损失函数训练的欺骗检测系统，使用LCNN和ResNet二者在t-DCF上相差不大，但在EER上LCNN（2.530%）相对于ResNet（3.290%）改善了0.76%。

从**表4**所示实验结果来看，LFCC-LCNN-OC-Softmax方案的CM欺骗检测系统的效果最好；LFCC+ResNet+OC-Softmax方案的CM欺骗检测系统，在与ASV系统级联使用时，对ASV系统可靠性的干扰最小。在验证集上，模型推理阶段面对的是已知攻击类型即模型学习过的伪造语音类型，使用设计的OC-Softmax损失函数训练，ResNet-18网络提取特征的性能最好。但面对测试集中的未知攻击类型的伪造语音，同样使用设计的OC-Softmax损失函数训练，但是使用LCNN提取特征比ResNet-18的效果更好，这也重复展现了LCNN在模型推理阶段，面对未知攻击类型伪造语音的良好的泛化能力。

从随机一种攻击类型和真实类型数据中选择500组数据，使用t-SNE降维算法，将模型输出的特征嵌入向量可视化，分别展示ASVspoof2019 LA测试集数据经过1轮训练和50轮训练之后的训练效果。

图10所示为经过多轮学习之后，真实语音和伪造语音已经有了大致的区分，除了测试集中极个别未知攻击类型的伪造语音，由于其使用伪造算法，模型分类较困难，总体上设计的LCNN特征提取和OC-Softmax损失函数训练相结合的方案，欺骗检测能力性能较良好，并且面对大多数未知攻击的特征表示，也显示出了良好的泛化能力。

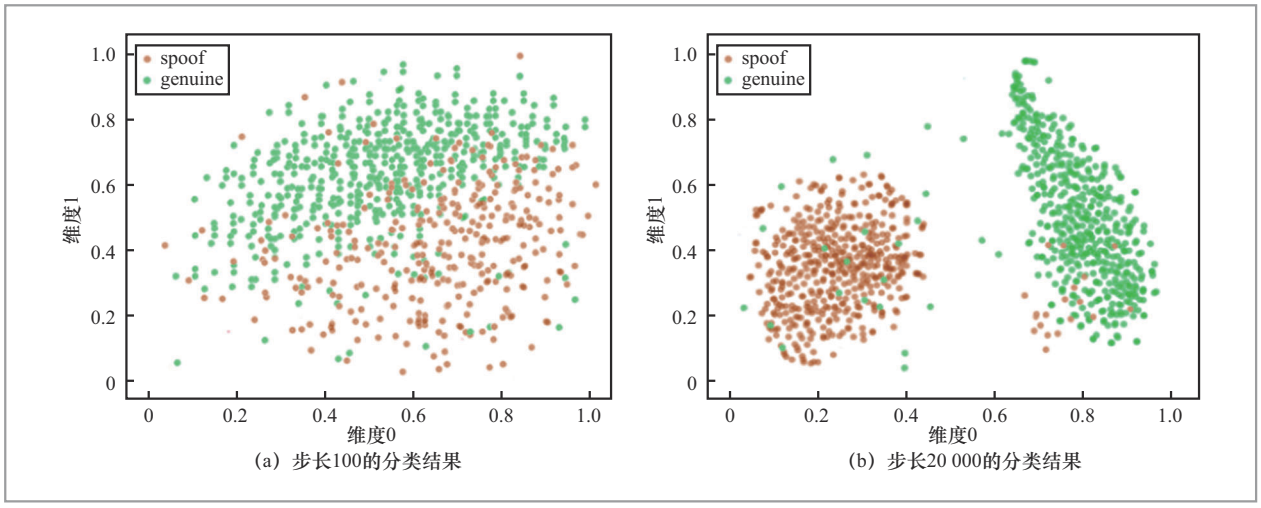


图10 2019年的自动说话人验证欺骗与对策挑战赛LA测试集数据降维分析

通过上述实验已经验证了语音LFCC特征在音频伪造检测任务中的优越性，但是面对显示场景，往往更多的是面对未知攻击类型的伪造音频，此时模型泛化能力至关重要。为了探究自监督预训练模型对提升模型泛化性的影响，本文在Wav2vec2预训练模型的基础上进行了多组实验，并与基线模型进行了对比，实验方案设置见**表5**。

本文所有系统均先在2019年的自动说话人验证欺骗与对策挑战赛toy数据集上进行实验，并使用官网提供的计算函数计算两个系统级联后的t-DCF和设计的CM系统的EER。CM系统在toy数据集上的效果如下。

在探究Wav2vec2是否能够应用到语音欺骗检测任务中时，由于计算资源有限，先在一个小数据集上进行实验并观察效果，实验结果见**表6**，相较于前端使用LFCC特征的模型，个别实验方案相较于基线模型有所提升，说明使用自监督预训练模型Wav2vec2进行语音真伪识别的任务具有可行性。

3.5 消融实验

为了保证实验的公平性，前端使用LFCC特征作为基线模型，后端部分与其他实验方案一样，也进行了改进，添加了BiLSTM和GAP模块。实验结果见**表7**，当前端使用同一种预训练模型但没有采用

表5 实验方案设置

CM系统	Front-end	LCNN	Bi-LSTM	GAP
LF-LBiGF(baseline)	LFCC	√	√	√
W2V-XLSR-LBiGF		√	√	√
W2V-XLSR-BiGF	Wav2vec-XLSR-56	×	√	√
W2V-XLSR-GF		×	×	√
W2V-large1-LBiGF	Wav2vec-large	√	√	√
W2V-small-LBiGF	Wav2vec-small	√	√	√

表6 在LA_Toy数据集上的实验效果

CM系统/toy	EER	min t-DCF	Threshold	Train
LF-LBiGF(baseline)	6.110%	0.067	6.426	—
W2V-XLSR-LBiGF	0.746%	0.014	3.307	frozen
	6.110%	0.067	10.194	fine-tune
W2V-XLSR-BiGF	4.991%	0.037	-0.913	frozen
	1.493%	0.029	4.849	fine-tune
W2V-XLSR-GF	13.713%	0.450	-1.747	frozen
	1.493%	0.029	4.419	fine-tune
W2V-large1-LBiGF	6.110%	0.165	-0.434	frozen
W2V-small-LBiGF	1.119%	0.022	0.388	frozen

表7 在LA数据集上的实验效果

CM系统/LA	EER	min t-DCF	Threshold	Train
LF-LBiGF(baseline)	3.182%	0.092	4.829	—
W2V-XLSR-LBiGF	3.705%	0.114	2.074	frozen
	2.086%	0.047	7.087	fine-tune
W2V-XLSR-BiGF	7.812%	0.213	-0.998	frozen
	1.263%	**0.032**	6.291	fine-tune
W2V-XLSR-GF	17.442%	0.482	-1.714	frozen
	1.428%	0.041	3.849	fine-tune
W2V-large1-LBiGF	18.090%	0.602	0.354	frozen
	33.966%	0.912	-4.961	fine-tune
W2V-Small-LBiGF	20.196%	0.500	-2.981	frozen
	3.744%	0.098	1.010	fine-tune
RawNet2[34]	1.748%	0.054	—	—
FastAudio[35]	1.655%	0.049	—	—
AASIST	**1.342%**	0.037	—	—
Ensemble_P	1.469%	**0.033**	—	—

微调策略时，以W2V-XLSR为例，后端使用LCNN-BiLSTM（LBiGF）的结构进一步提取特征，并在模型后面接全局平均池化层，然后使用全连接层调整输出维度的组合，在EER和t-DCF上，相较于后端删除LCNN模块的BiGF模型和同时删除LCNN和BiLSTM两个模块的GF模型，EER分别降低了4.107%和13.737%，t-DCF分别降低了0.099和0.368，即模型W2V-XLSR-LBiGF在没有微调时的效果是最好的。与此同时，将LFCC-LCNN-OC-Softmax、W2V-XLSR-LBiGF、W2V-XLSR-BiGF、W2V-XLSR-GF、以及W2V-Small-LBiGF模型通过加权平均融合成Ensemble_P模型，同样使用Optuna调参工具获得最佳权重系数。与RawNet2，FastAudio和AASIST相比，融合模型在EER上低于以FastAudio和以RawNet2为结构的欺骗检测系统，并接近目前的SOTA方案AASISIT，显示出了良

好的语音鉴伪能力。其在t-DCF上取得了最好的效果，表明Ensemble_P在保持良好语音鉴伪能力的同时，大幅减少了CM系统可能对ASV系统造成的干扰，提升了模型的鲁棒性。

本文将前端Wav2vec2模型在2019年的自动说话人验证欺骗与对策挑战赛LA数据集上进行微调后，单模型的实验结果与之前不同，后端删除了LCNN的实验方案，即W2V-XLSR-GF模型以及W2V-XLSR-BiGF模型的EER分别为1.428%和1.263%，t-DCF分别为0.041和0.032，相较于W2V-XLSR-LBiGF均有大幅提升，这表明当前端Wav2vec2模型在目标任务数据集上微调后，后端模型往往不需要很复杂的结构和很深的网络，如实验所示采用GAP和FC组合的后端结构已经在测试集上取得了不错的效果，且预训练模型微调后提升了欺骗检测效果。

从**表7**中的W2V-XLSR-LBiGF模型的实验结果可以看到，使用微调策略后，串联检测成本函数t-DCF的值大幅降低；实验结果表明，针对前端使用自监督模型，微调能有效提升CM系统语音欺骗检测能力，并在保证一定鉴伪能力的同时，不影响ASV系统的安全性和可靠性。所有实验方案中，除了使用W2V-Large1预训练模型的实验，在进行系统微调后的模型性能均优于冻结预训练参数的方案，并且由于微调后可以简化后端分类器模型的结构，减少了训练参数，欺骗检测模型的收敛速度也更快。单模型情况下XLSR-BiGF-finetune方案的CM欺骗检测系统效果最好，同时对级联的ASV系统可靠性影响最小。针对前端自监督预训练模型的选择部分，采用微调后，由表7可知，实验性能最好的是W2V-XLSR-BiGF模型，其EER为1.263%，t-DCF为0.032。

从模型结构来看，W2V-XLSR-BiGF模型在后端相较于W2V-XLSR-LBiGF减少了LCNN模块，从侧面证明了之前的结论，即当前端使用自监督预训练模型时，后端往往不需要复杂或较深的网络结构。与W2V-XLSR-GF相比，最优模型较之增加了Bi-LSTM模块，同样说明了具有上下文时序信息的Bi-LSTM模型在语音欺骗检测任务中的适用性，并且表现出不错的性能。

综合实验结果来看，本文提出的前端基于自监督Wav2vec2预训练模型，后端采用LCNN、Bi-LSTM和全局平均池化模块组合的语音欺骗检测系统，展现出了不错的效果，这也证明了自监督预训练模型在语音欺骗检测任务中的可行性。本文通过模型融合的手段，集成模型Ensemble_P优于FastAudio和RawNet2，在EER上接近AASIST，在t-DCF上实现最优，证明了模型具有良好的语音鉴伪能力，同时与对比实验相比，系统可靠性最高。

4 结束语

本文对语音欺骗检测任务展开讨论，针对语音欺骗检测模型面对未知攻击类型的伪造语音模型泛化能力较差的问题，提出了OC-Softmax单类分类器，使模型只关注真实语音的特征分布，为真实语音和伪造语音训练不同的决策边界，实验结果表明，相较于基线模型和其他二分类损失函数，模型性能有明显提升。

面对现实工作中往往缺少带标签的伪造语音样本的问题，本文提出了基于自监督预训练模型Wav2vec2的语音欺骗检测系统，旨在利用通用性、鲁棒性更强的Wav2vec特征向量代替之前的LFCC特

征。其前端直接从语音波形文件获取适用于分类的特征向量。本文在证明了Wav2vec2应用在语音鉴伪领域的可行性的基础上，探究了适合前端预训练模型的最优的后端分类模型结构，以及欺骗检测性能最好的前后端配置。本文将以LFCC特征为前端的高分模型和以Wav2vec2为前端的高分模型加权平均，并通过Optuna调参工具寻找最优的权重系数。结果表明，融合模型Ensemble_P实现了接近SOTA的鉴伪能力，并降低了对ASV系统安全性和可靠性的影响，提升了模型的鲁棒性，同时提出的方案也解决了低资源伪造音频和模型泛化能力弱两个问题。

参考文献：

[1] 张奥运. 基于密集神经网络的语音欺骗检测[D]. 广州: 广东技术师范大学, 2021.
ZHANG A Y. Voice deception detection based on dense neural network[D]. Guangzhou: Guangdong Polytechnic Normal University, 2021.

[2] 陈群, 陈肇强, 侯博议, 等. 人工智能风险分析技术研究进展[J]. 大数据, 2020, 6(1): 47-59.
CHEN Q, CHEN Z Q, HOU B Y, et al. Research progress on risk analysis for artificial intelligence[J]. Big Data Research, 2020, 6(1): 47-59.

[3] 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 X W, LI J K, SUN M, et al. Speech anti-spoofing: the state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2020, 35(5): 807-823.

[4] 甘海林. 基于高斯概率特征的双路BiLSTM和DCNN在语音欺骗检测中的研究[D]. 南昌: 江西师范大学, 2021.
GAN H L. Research on dual BiLSTM and DCNN based on Gaussian probability feature in voice fraud detection[D]. Nanchang: Jiangxi Normal University, 2021.

[5] 李鹏程, 张旭龙, 王健宗, 等. 面向非平行语料的语音转换技术综述[J]. 大数据, 2024, 10(3): 65-81.
LI P C, ZHANG X L, WANG J Z, et al. A survey of voice conversion based on non-parallel data[J]. Big Data Research, 2024, 10(3): 65-81.

[6] 唐浩彬, 张旭龙, 王健宗, 等. 表现性语音合成综述[J]. 大数据, 2023, 9(6): 53-71.
TANG H B, ZHANG X L, WANG J Z, et al. A survey of expressive speech synthesis[J]. Big Data Research, 2023, 9(6): 53-71.

[7] 徐嘉, 简志华, 金宏辉, 等. 采用恒Q调制包络的合成语音伪装检测方法[J]. 电信科学, 2023, 39(11): 107-115.
XU J, JIAN Z H, JIN H H, et al. A method of synthetic speech spoofing detection using constant Q modulation envelope[J]. Telecommunications Science, 2023, 39(11): 107-115.

[8] 邱泽宇, 屈丹, 张连海. 基于WaveNet的端到端语音合成方法[J]. 计算机应用, 2019, 39(5): 1325-1329.
QIU Z Y, QU D, ZHANG L H. End-to-end speech synthesis based on WaveNet[J]. Journal of Computer Applications, 2019, 39(5): 1325-1329.

[9] 曹娟, 朱勇椿, 亓鹏, 等. 数字内容生成、检测与取证技术综述[J]. 大数据, 2023, 9(5): 150-173.
CAO J, ZHU Y C, QI P, et al. A survey on digital content generation, detection, and forensics techniques[J]. Big Data Research, 2023, 9(5): 150-173.

[10] TODISCO M, WANG X, VESTMAN V, et al. ASVspoof 2019: Future horizons in spoofed and fake audio detection[J]. arXiv preprint, arXiv: 1904.05441, 2019.

[11] WU Z Z, DAS R K, YANG J C, et al. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks[C]//Proceedings of the Interspeech 2020. [S.l.: s.n.], 2020: 1101-1105.

[12] HANILÇI C, KINNUNEN T, SAHIDULLAH M, et al. Classifiers for synthetic speech detection: a comparison[C]//Proceedings of the Interspeech 2015. [S.l.: s.n.] 2015: 2057-2061.

[13] HASSAN F, JAVED A. Voice spoofing countermeasure for synthetic speech detection[C]//Proceedings of the 2021 International Conference on Artificial Intelligence (ICAI). Piscataway: IEEE Press, 2021: 209-212.

[14] TIAN X H, XIAO X, CHNG E S, et al. Spoofing speech detection using temporal convolutional neural network[C]//Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Piscataway: IEEE Press, 2016: 1-6.

[15] NOVOSELOV S, KOZLOV A, LAVRENTYEVA G, et al. STC anti-spoofing systems for the ASVspoof 2015 challenge[C]//Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2016: 5475-5479.

[16] WANG X, YAMAGISHI J. Estimating the confidence of speech spoofing countermeasure[C]//Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6372-6376.

[17] LIN G Y, LUO W Q, LUO D, et al. One-class neural network with directed statistics pooling for spoofing speech detection[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 2581-2593.

[18] TAK H, PATINO J, TODISCO M, et al. End-to-end anti-spoofing with RawNet2[C]//Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 6369-6373.

[19] REN Y Q, PENG H P, LI L X, et al. Generalized voice spoofing detection via integral knowledge amalgamation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2461-2475.

[20] HASSAN F, JAVED A. Voice spoofing countermeasure for synthetic speech detection[C]//Proceedings of the 2021 International Conference on Artificial Intelligence (ICAI). Piscataway: IEEE Press, 2021: 209-212.

[21] ZHANG C L, YU C Z, HANSEN J H L. An investigation of deep-learning frameworks for speaker verification anti-spoofing[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11 (4): 684-694.

[22] 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.

[23] LIU P F, ZHANG Z C, YANG Y C. End-

to-end spoofing speech detection and knowledge distillation under noisy conditions[C]//Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2021: 1-7.

[24] ZEN H G, SENIOR A, SCHUSTER M. Statistical parametric speech synthesis using deep neural networks[C]//Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2013: 7962-7966.

[25] KINNUNEN T, SAHIDULLAH M, DELGADO H, et al. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection[C]//Interspeech 2017. International Speech Communication Association, 2017: 2-6.

[26] LI J K, ZHANG X W, SUN M, et al. Attention-based LSTM algorithm for audio replay detection in noisy environments[J]. Applied Sciences, 2019, 9(8): 1539.

[27] ZHANG C L, YU C Z, HANSEN J H L. An investigation of deep-learning frameworks for speaker verification antispoofing[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(4): 684-694.

[28] CHEN T X, KUMAR A, NAGARSHETH P, et al. Generalization of audio deepfake detection[C]//Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2020). [S. l.: s. n.] 2020: 132-137.

[29] 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 Interspeech 2019. [S.l.: s. n.] 2019: 1068-1072.

[30] WU Z Z, DAS R K, YANG J C, et al. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks[C]//Proceedings of the Interspeech 2020. [S.l.: s.n.] 2020: 1101-1105.

[31] RAHUL T P, ARAVIND P R, RANJITH C, et al. Audio spoofing verification using deep convolutional neural networks by transfer learning[EB]. arXiv preprint, 2008, arXiv: 2008.03464.

[32] TAK H, PATINO J, TODISCO M, et al. End-to-end anti-spoofing with RawNet2[C]//Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 6369-6373.

[33] FU Q, TENG Z, WHITE J, et al. Fastaudio: a learnable audio front-end for spoof speech detection[C]//Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscatawa IEEE Press, 2022: 3693-3697.

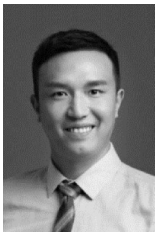
[34] JUNG J W, HEO H S, TAK H, et al. AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks[C]//Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6367-6371.

[35] KAMBLE M R, SAILOR H B, PATIL H A, et al. Advances in anti-spoofing: from the perspective of ASVspoof challenges[J]. APSIPA Transactions on Signal and Information Processing, 2020, 9(1).

作者简介

梁子琪（2000– ），男，中国科学技术大学先进技术研究院硕士生，平安科技（深圳）有限公司算法工程师，主要研究方向为语音合成、语音转换等。

张旭龙（1988– ），男，博士，平安科技（深圳）有限公司高级算法研究员，清华大学深圳研究院以及中国科学技术大学先进技术研究院校外导师，IEEE、中国自动化学会以及中国计算机学会会员，联邦数据与联邦智能专业委员会委员，2023年入选上海市东方英才计划青年项目，主要研究方向为语音合成、语音转换、音频驱动虚拟人生成、音乐信息检索、机器学习和深度学习方法在人工智能领域应用。



王健宗（1983– ），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理，智能金融前沿技术研究院院长，美国佛罗里达大学人工智能博士后，美国莱斯大学和华中科技大学联合培养博士，中国计算机学会资深会员，中国计算机学会大数据专家委员会委员，中国自动化学会联邦数据和联邦智能专业委员会副主任，主要研究方向为大模型、联邦学习和深度学习等。

肖京（1972– ），男，博士，美国卡耐基梅隆大学博士，国家特聘专家，国家新一代普惠金融人工智能开放创新平台技术负责人、深圳市政协委员、深圳市决策咨询委员会委员，兼中国计算机学会深圳分部副主席、广东省人工智能与机器人学会副理事长、深圳市人工智能行业协会会长、深圳市人工智能学会副理事长，清华大学、上海交通大学、同济大学等客座教授，先后在爱普生美国研究院及美国微软公司担任高级研发管理职务，现任平安集团首席科学家，负责人工智能技术研发及在金融、医疗、智慧城市等领域的应用，带领团队树立了多项传统行业智能化经营的标杆，主要研究方向为人工智能与大数据分析挖掘相关领域。已发表学术论文249篇，美国授权专利101项，中国发明专利155项，参与及承担国家级项目8项。凭借在技术创新及应用的杰出贡献，先后获得2018年中国专利奖、2019年吴文俊人工智能杰出贡献奖、2020年吴文俊人工智能科技进步一等奖、2020年上海市科技进步奖一等奖、2020年中国人工智能十大风云人物、2021年深圳市五一劳动奖章、2022年深圳市最美科技工作者等荣誉。

收稿日期: 2023-09-26

通信作者: 王健宗，jzwang@188.com

基金项目: 广东省重点领域研发计划“新一代人工智能”重大专项(No.2021B0101400003)

Foundation Item: The Key Research and Development Program of Guangdong Province“New Generation Artificial Intelligence”Major Special Project (No. 2021B0101400003)