研究与开发

采用自监督对比学习的合成伪造语音检测方法

杨曼，简志华，梁承涵
（杭州电子科技大学通信工程学院，浙江 杭州 310018）

摘　要：为了消除训练数据集中真实语音和伪造语音的样本数量不平衡对合成伪造语音检测系统性能的影响，并进一步提高系统的检测准确率，提出了一种基于自监督对比学习的合成语音检测方法。所提方法将经过音高变换后的样本视为负样本，通过训练神经网络使锚点样本特征与负样本特征不同，从而促使网络提取对于音高变换敏感的特征，再采用深度残差网络作为后端分类器来判决语音真伪。实验结果表明，与传统手工设计的声学特征方法、基于深度学习的伪造语音检测系统以及基于端到端的伪造语音检测系统相比，所提方法显著降低了系统的等错误率。由于自监督对比学习的合成伪造语音检测方法可以训练网络提取对音高变换敏感的特征，并且不受数据集中真伪语音数量不平衡的影响，因此显著提高了合成伪造语音检测的准确率。

关键词：伪造语音检测；合成语音检测；自监督对比学习；深度残差网络；音高变换
中图分类号：TN912
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024236

A method of synthetic spoofing speech detection using self-supervised contrastive learning

YANG Man, JIAN Zhihua, LIANG Chenghan
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: In order to eliminate the impact of the imbalance of the sample size of bonafide speech and fake speech in the training dataset on the performance of synthetic speech detection system and further improve the accuracy of synthetic speech detection, a method of synthetic speech detection was proposed based on self-supervised contrastive learning. In this method, the samples after pitch transformation were regarded as negative samples, and the neural network was trained to make the anchor sample features different from the negative sample features, so that the network could extract the features sensitive to pitch transformation. And the deep residual network was used as the back-end classifier to judge the authenticity of the speech. Experimental results show that, compared with the traditional hand-crafted acoustic features, the deep learning-based and the end-to-end spoofing speech detection systems, the proposed method significantly reduces the equal error rate of the system. The synthetic forged speech detection method based

收稿日期：2024−07−30；**修回日期：**2024−10−07
通信作者：简志华，jianzh@hdu.edu.cn
基金项目：国家自然科学基金资助项目（No.61201301，No.61772166）
Foundation Items: The National Natural Science Foundation of China (No.61201301, No.61772166)

on self-supervised contrastive learning can train the network to extract features sensitive to pitch transformation and will not affect the accuracy of synthetic speech detection because of the imbalance of bonafide and fake speech in the dataset, so the accuracy of synthetic forged speech detection is significantly improved.

Key words: spoofing speech detection, synthesized speech detection, self-supervised contrastive learning, deep residual network, pitch transformation

0　引言

随着智能语音助手应用市场规模的快速增长，用于身份安全认证的声纹识别技术也得到了越来越广泛的应用[1]。声纹识别是一种根据语音特征来识别说话人身份的技术，它包括说话人辨识（speaker identification，SI）和说话人确认（speaker verification，SV）两方面。其中，自动说话人确认（automatic speaker verification，ASV）通过输入语音并与语音助手中已注册的语音进行匹配来确认说话人身份[2]。然而，随着语音深度伪造[3]（deepfake）技术的发展，生成的语音越来越逼近真实的自然语音，这对包括智能语音助手在内的各类声纹识别设备的安全构成越来越大的威胁。常见的伪造语音方法有语音合成、语音转换、语音重放和语音模仿等。深度学习技术的发展以及端到端语音合成系统性能的不断提升，减少了合成过程中的人工干预以及对语言学相关背景知识的需求，让语音合成变得越来越便捷，合成语音质量也越来越高，利用合成语音对ASV系统进行恶意攻击成为威胁声纹安全认证的主要因素。因此，对合成语音检测（synthetic speech detection，SSD）[4]展开研究具有非常重要的意义。

SSD的基本原理主要是根据合成语音与自然语音之间的特征差异，并利用分类器进行真伪判决。SSD系统由前端特征提取模块和后端分类器模块两部分组成，前端模块分析语音信号并提取具有区分性的声学特征，后端模块则通过分类器判断输入语音是真实语音还是合成的伪造语音。传统的SSD系统前端模块通常采用手工设计的声学特征，主要有梅尔频率倒谱系数[5]（Mel-frequency cepstral coefficient，MFCC）、线性频率倒谱系数[6]（linear frequency cepstral coefficient，LFCC）和常数Q倒谱系数[7]（constant Q cepstral coefficient，CQCC）等。MFCC的提取使用梅尔滤波器组，该滤波器组的设计是根据人耳对声音的感知特性而来，即人耳对于频率的感知不是线性的，而是在低频段较为敏感，在高频段逐渐不太敏感。因此，MFCC特征比较符合人耳对声音的感知。CQCC与MFCC基本相似，不同之处在于CQCC使用了常数Q滤波器组，该滤波器组在频率上的间隔也是非线性的，与MFCC一样，也是考虑人耳对声音的感知，但CQCC使用的常数Q滤波器组在低频部分的频率间隔相对较小，这使CQCC在低频范围内能够提供更高的频谱分辨率，因此，CQCC比MFCC在低频部分能够更好地捕捉到细微的频率变化。LFCC提取时则是直接在线性频率尺度上进行滤波，没有先转换到梅尔频率刻度上，这样做的好处是能够更均匀地覆盖整个频率范围，特别是在高频部分，由于频率间隔相对较小，采用线性频率刻度能够更好地捕捉到声音的细微变化，因此，LFCC特征比MFCC和CQCC特征具有更高的区分度和准确率。然而，以上这些特征参数主要反映的是声学信息，只考虑了与频率有关的特征信息，难以较为全面地揭示语音中的丰富信息。

近年来，基于深度学习的伪造语音检测系统逐渐成为主流[8]，如卷积神经网络[9]（convolutional neural network，CNN）、递归神经网络[10]

（recurrent neural network，RNN）、长短期记忆[11]（long short-term memory，LSTM）网络等。这些深度学习模型通常需要大量标记数据进行训练，而获取大规模的伪造语音数据并进行标注是一项成本高昂且耗时的工作。此外，在伪造语音检测任务中，公开数据集中真实语音和伪造语音的样本数量往往存在明显的不平衡现象，这会导致模型倾向于更多地关注样本数量较多的类别，而忽略样本数量较少的类别，从而影响了系统的检测性能。传统的基于深度神经网络的方法需要预先手工设计特征，再通过神经网络学习这些特征的高级表示，并据此进行分类判决。为了简化系统结构，学术界设计了端到端的伪造语音检测系统，与传统的基于手工设计特征和深度学习模型的方法不同，端到端系统能够直接从原始数据中学习特征和训练分类器，省去了手工设计特征等复杂流程，使得系统更加简洁和高效。然而，端到端的伪造语音检测系统与传统的基于深度神经网络的检测系统一样，都需要大量标记数据进行训练，才能获得良好的性能。

为了更全面地提取语音中的特征信息，并使用非标记数据训练模型，本文提出了一种基于自监督学习[12]的合成伪造语音检测方法。自监督学习利用借口任务从大量未标记数据中挖掘丰富的信息对模型进行训练，从而学习到对下游任务具有价值的特征。该借口任务使用模型自动生成的伪标签预训练模型，将某一样本的特征采用两种不同的变换后分别得到锚点样本和正样本，其余样本的特征则视为负样本。由于锚点样本和正样本从同一样本变换得到，没有改变语义信息，因此被称作正样本对，相对于正样本对，其余样本的特征则都为锚点样本的负样本对。自监督对比学习（self-supervised contrastive learning，SSCL）通过比较正样本对和负样本对的相似性或者差异性来学习特征表示，其核心思想是训练模型使锚点样本特征接近正样本的特征，同时远离负样本的特征，从而使模型可以获取更加鲁棒的特征。由于合成语音存在韵律泄露的情况，音高或者音量等韵律元素在语音合成过程中会遭到破坏甚至丢失，合成语音在音高、音量等方面与自然语音有所不同，存在合成的痕迹。音高对于辨别该语音是否为合成语音至关重要，因此该方法选择音高变换来生成负样本。在SSD中，对比学习模型学习语音信号之间的相似性或者差异性，可以更好地区分真实语音和伪造语音，因此，自监督对比学习非常适用于合成伪造语音检测。然而，在合成伪造语音检测的相关研究中，常用的自监督对比学习语音表示方法，由于在训练时未提前标注大规模的伪造语音数据，因此，通常在实验中将相邻的语音帧作为正样本，并通过对来自同一批次的数据进行随机的语音增强以获得负样本。这种正负样本的选择方法未得到理论上的论证，无法保证所采样得到的负样本的质量，因此可能会出现与预期效果相反的正负样本对，导致训练效果不理想[13]。与普通的自监督对比学习不同，本文所提算法将经过音高变换后的样本视为负样本，通过训练神经网络来使锚点样本特征与负样本特征相区分，从而促使网络提取出对于特定变换敏感的特征，进而提高合成语音检测的准确性。

1 自监督模型

本文提出的SSD方法采用动量对比（momentum contrast，MoCo）[14]作为对比学习模型。MoCo原本的设计思想是通过训练网络使锚点样本的特征与经过特定变换后的正样本特征相似，从而促使网络能够提取出对于特定变换鲁棒的特征。本文对MoCo模型进行了改进，将经过特定变换的样本视为负样本，并训练神经网络使锚点样本特征与负样本特征不同，从而使网络提取出对于特定变换敏感的特征。

MoCo对比学习网络预训练流程如图1所示，每个批次的预训练过程分为以下4个步骤。

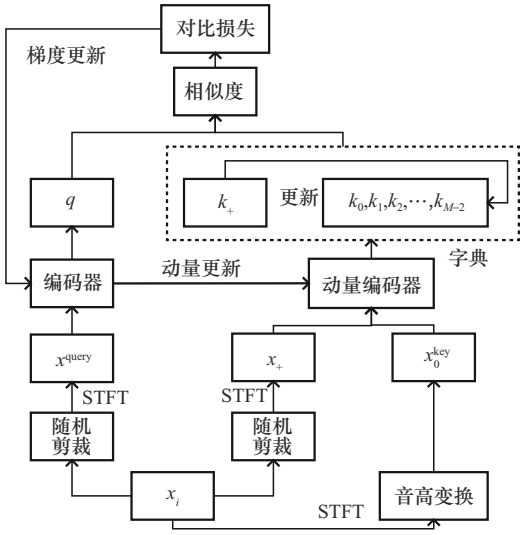

图1　MoCo对比学习网络预训练流程

步骤1　从该训练批次的N个样本中任选一语音样本x_i并做两次随机剪裁，通过短时傅里叶变换[15]（short-time Fourier transform，STFT）生成对应的语谱图，即得到锚点样本x^{query}和正样本x_+，同时对该语音样本x_i做音高变换处理后应用STFT生成负样本x_0^{key}。然后再将该训练批次中其余$N-1$个语音样本用同样的方法，即经音高变换后应用STFT生成语谱图，得到$N-1$个负样本$\left\{x_n^{\text{key}}, n=1,2,\cdots,N-1\right\}$。

步骤2　将语音样本x_i生成的锚点样本x^{query}和正样本x_+分别输入编码器和动量编码器，得到固定大小的特征表示$f_\varphi(x^{\text{query}})$和$f_\phi(x_+)$，语音样本$x_i$生成的负样本$x_0^{\text{key}}$和其余$N-1$个样本生成的负样本输入动量编码器后得到固定大小的特征表示$f_\phi\left(x_0^{\text{key}}\right)$和$\left\{f_\phi\left(x_n^{\text{key}}\right), n=1,\cdots,N-1\right\}$。

步骤3　通过编码器和动量编码器生成的特征表示分别经过投影，生成查询$q=g\left(f_\varphi\left(x^{\text{query}}\right)\right)$和键值$k_+=g\left(f_\phi(x_+)\right)$、$k_0=g\left(f_\phi\left(x_0^{\text{key}}\right)\right)$、$\left\{k_n=g\left(f_\phi\left(x_n^{\text{key}}\right)\right), n=1,\cdots,N-1\right\}$。

步骤4　计算对比损失梯度更新编码器，同时动量更新动量编码器。

在步骤1的数据预处理阶段，将同一训练批次中锚点样本对应的语音样本和其余语音样本都进行音高变换以获得负样本，由于在预训练过程中模型会拉近锚点样本和正样本之间的距离，同时推远锚点样本与负样本之间的距离，因此，对比学习模型在预训练结束后可以获得对音高变换敏感的特征表示。

在步骤2的编码器阶段，编码器和动量编码器具有相同的网络结构，即卷积循环神经网络[16]（convolutional recurrent neural network，CRNN）。在网络训练过程中，通过计算损失函数的梯度来更新编码器参数，其中动量编码器的更新方式为动量更新，其数学表达式为：

$$\Theta_\phi \leftarrow \delta\cdot\Theta_\phi+(1-\beta)\cdot\Theta_\varphi \tag{1}$$

其中，$\delta\in[0,1)$，为动量系数。

在步骤3的动态字典更新阶段，初始动态字典是由训练之初多个训练批次的语音样本特征组成的，当动态字典中的样本特征个数达到队列大小M后，动态字典开始使用维护队列来存储动量编码器提取到的特征。在下一训练批次的样本特征被输入动态字典中之前，先把最早训练批次的样本特征从动态字典中移除，从而使动态字典保持一个恒定大小为M的队列。动态字典中的k_+为q的正对，$(k_0,k_1,k_2,\cdots,k_{M-2})$为$q$的负对，即每个特征在预训练阶段存在一个正对和$M-1$个负对。

在步骤4的计算损失函数阶段，通过计算信息最大化的归一化交叉熵[17]（normalized cross entropy for info-max，infoNCE）来更新编码器的参数。首先计算q_i, $i\in\{1,\cdots,N\}$和$k_{i,j}$, $j\in\{+,0,1,\cdots,M-2\}$之间的相似度，其中N表示每个训练批次中的语音样本个数并且$N\ll M$，并采用余弦相似度进行计算，即：

$$s_{i,j}=\frac{q_i^{\mathrm{T}}k_{i,j}}{\tau\left\|q_i\right\|\left\|k_{i,j}\right\|} \tag{2}$$

其中，τ表示超参数。然后，再计算infoNCE：

$$L=\frac{1}{N}\sum_{i=1}^{N}\left(-\log\frac{\exp(s_{i,+})}{\sum_{j=0}^{M-2}\exp(s_{i,j})}\right) \tag{3}$$

2 合成伪造语音检测

2.1 后端分类模型

由于检测系统的前端模型采用了自监督模型，为了减少系统整体的复杂度，后端模型采用层数较少的深度残差网络ResNet18作为分类器的基础网络架构[18]。该网络主要由1个卷积层、4个残差块、1个平均池化层和1个全连接层组成。每个残差块包含4个卷积层、2个ReLU激活函数和3个批归一化操作。ResNet18将通过自监督模型提取的特征向量作为输入，经过卷积、归一化、池化等操作后，将最后一个全连接层之前的输出作为语音的嵌入向量。

ResNet18的损失函数采用单分类损失函数（one classification softmax， OC-Softmax）[19]。OC-Softmax设置两个边界得到一个紧凑的嵌入空间，其中真实语音的嵌入具有紧凑的边界，而伪造语音与真实语音之间保持一定的距离，在该函数中真实语音被视为目标类，伪造语音则为非目标类[20]。OC-Softmax通过压缩真实语音的表示，并加入角度余量来分离嵌入空间中的伪造语音，OC-Softmax的表达式为：

$$L_{\mathrm{OCS}}=\frac{1}{D}\sum_{i=1}^{D}\log\left(1+\mathrm{e}^{\alpha(m_{b_i}-\boldsymbol{\omega}_0\boldsymbol{a}_i)(-1)^{b_i}}\right) \tag{4}$$

其中，$\boldsymbol{a}_i$为每个语音的嵌入向量，$b_i\in\{0,1\}$表示第i个语音样本的标签，$b_i=0$表示该语音样本为目标类，$b_i=1$表示该语音样本为非目标类，D为一个批次中的样本数量，α是一个缩放因子，权重向量$\boldsymbol{\omega}_0$表示目标类嵌入的优化方向。两个边距$m_0,m_1\in[-1,1]$并且满足$m_0>m_1$，分别用于压缩真实语音的嵌入和合成语音的嵌入，以限制$\boldsymbol{\omega}_0$和$\boldsymbol{a}_i$之间的角度θ_i。当$b_i=0$时，m_0用于迫使$\theta_i<\arccos m_0$；当$b_i=1$时，m_1用于迫使$\theta_i>\arccos m_1$。一个相对较小的$\arccos m_0$使目标类集中在权重向量$\boldsymbol{\omega}_0$附近，而一个相对较大的$\arccos m_1$使非目标类远离权重向量$\boldsymbol{\omega}_0$。

2.2 检测流程

在检测系统的前端模块，首先对预训练数据集中的每个语音样本进行随机剪裁，并截取时长为5 s的语音信号，然后对剪裁后的语音样本进行音高变换处理。对变换后的语音样本做STFT处理得到频谱图，并输入前端自监督对比学习模型，在经过CRNN处理后最终得到1×256维的特征表示。自监督对比学习模型在预训练过程结束后，得到自监督对比学习的模型参数并保持不变，用于提取评估数据集的特征向量。

在后端模型训练时，首先将训练集和开发集的特征向量作为网络输入，输出每个语音样本的嵌入向量，再使用OC-Softmax损失函数调整优化网络模型的参数，使得训练所得的权重向量与真实语音样本的嵌入向量尽可能接近，与欺骗语音样本的嵌入向量尽可能分离。训练过程结束后获得网络模型的参数，然后将测试集的特征向量输入该网络中，输出每个语音样本的嵌入向量。最终计算测试语音样本的嵌入向量和权重向量的余弦相似度，并将其作为评价得分，用于判决语音信号的真伪。合成伪造语音检测系统流程如图2所示。

3 实验与结果

3.1 数据集

为了预训练检测系统的前端模型，实验使用的语料库是ASVspoof 2015[21]挑战赛和ASVspoof 2017挑战赛中的逻辑访问（logical access，LA）场景数据集，它们都包括训练集、开发集和评估集，每个子集都包括真实语音样本和欺骗语音样本。实验将ASVspoof 2015 LA和ASVspoof 2017 LA各子集共20 162个真实语音样本组成预训练数据集，用于训练自监督模型CRNN。

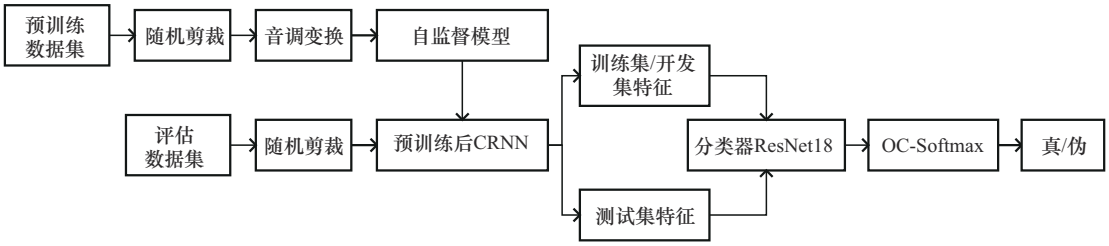

图2　合成伪造语音检测系统流程

对于检测系统的后端模型，实验使用的语料库是 ASVspoof 2021 挑战赛[22]中的 LA 场景数据集，它包括训练集、开发集和测试集。由于挑战赛官方未更新训练集与开发集，继续沿用 ASVspoof 2019 LA 数据集中的训练集和开发集，而它的测试集在 ASVspoof 2019 LA 测试集的基础上采用多种编解码器进行处理，并用传输设备模拟现实场景下的传输处理，增加了语音数据的复杂性和多样性。实验选取训练集和开发集作为训练语音样本，选用测试集的语音样本进行性能评估。实验数据集信息见表 1。

表 1　实验数据集信息

数据集	语料库	子集	真实语音/条	欺骗语音/条	总语音/条
预训练数据集	ASVspoof 2015 LA	—	16 651	—	16 651
	ASVspoof 2017 LA	—	3 511	—	3 511
评估数据集	ASVspoof 2021 LA	训练集	2 580	22 800	25 380
		开发集	2 548	22 296	24 844
		测试集	—	—	181 556

3.2　实验参数设置

针对检测系统的前端模块，自监督对比学习模型通过 Adam 优化网络参数进行预训练，学习率设置为 0.01，权重衰减为 10^{-4}，并且每 25 个 epoch 衰减 20%，实验共训练 150 个 epoch。针对检测系统的后端模块，同样使用 Adam 优化后端网络模型参数，其中，参数 $\beta_1=0.9$、$\beta_2=0.999$，用于更新 ResNet 模型中的权重。在实验初始化时，设置 OC-Softmax 损失函数的尺度因子为 $\alpha=20$、$m_0=0.9$、$m_1=0.2$，利用随机梯度下降优化器来更新参数，其中 Batch 大小为 64，学习率初始设置为 0.000 3，并且每 5 个 epoch 衰减 50%，共训练 100 个 epoch。

3.3　性能评价指标

实验采用等错误率（equal error rate，EER）指标来评价伪造语音检测系统的性能。伪造语音检测作为二分类任务，当检测系统将伪造语音错误分类成真实语音时为错误接受，当检测系统将真实语音错误分类成伪造语音时为错误拒绝。给定检测系统的检测分数和阈值 λ，错误接受率（false accept rate，FAR）和错误拒绝率（false rejection rate，FRR）的计算表达式分别为：

$$P_{\mathrm{FAR}}(\lambda)=\frac{\text{得分大于阈值}\lambda\text{的伪造语音数量}}{\text{伪造语音总数}} \tag{5}$$

$$P_{\mathrm{FRR}}(\lambda)=\frac{\text{得分小于或等于阈值}\lambda\text{的伪造语音数量}}{\text{真实语音总数}} \tag{6}$$

其中，$P_{\mathrm{FAR}}(\lambda)$和$P_{\mathrm{FRR}}(\lambda)$分别是$\lambda$的单调递减函数和单调递增函数，$P_{\mathrm{FAR}}(\lambda)$和$P_{\mathrm{FRR}}(\lambda)$相等时的值称为等错误率 EER，即 $\mathrm{EER}=P_{\mathrm{FAR}}(\lambda)=P_{\mathrm{FRR}}(\lambda)$，EER 越小则代表伪造语音检测系统的性能越好。

3.4　系统性能测试

在对前端模型进行训练时，音高变换的动态范围过大会导致特征表示仅对非自然伪影敏感，而音高变换过小又不能为对比学习产生有效的负样本。因此，为了找到最优的音高变换范围，实验比较了音高变换范围分别为 1～5 半音时前端模型的性能。实验使用 SoX（sound eXchange）[23]工具进行音高变换，其具体方式为 SoX 中 pitch 功

能接受参数用于指定音高变换范围，即当pitch接受参数为$[-h,h]$时，意味着将语音随机地降低或提高h个半音，实验通过改变pitch接受参数来调整音高变换范围。实验分别将预训练数据集应用5种音高变换范围后，训练前端的对比学习模型，然后将评估数据集中各子集输入前端对比学习模型提取特征向量，高斯混合模型（Gaussian mixture model，GMM）作为后端分类器进行合成伪造语音检测。将评估数据集中训练集和开发集的特征向量输入GMM模型进行训练，计算评估数据集中开发集的EER值。不同音高变换范围下特征向量的EER对比见表2。实验结果表明，将音高变换范围限制为3个半音时，取得了较低的EER值，对评估数据集中开发集的合成伪造语音检测性能最优，因此，后续实验均在音高变换范围为3个半音的基础上开展。

表2　不同音高变换范围下特征向量的EER对比

音高变换范围	EER
[−1,1]	10.32%
[−2,2]	8.26%
[−3,3]	7.21%
[−4,4]	7.98%
[−5,5]	12.37%

在确定音高变换范围后，为了验证后端分类器对于前端模型性能的影响，实验采用不同的后端分类模型对前端对比学习模型的性能进行比较分析，包括GMM、五折交叉验证（5-fold cross-validation）[24]和ResNet18。五折交叉验证将评估数据集中的训练集和开发集合并后分为5个子集，其中4个子集用于训练，1个子集用于测试，过程重复5次，每次使用不同的子集作为测试集，最终取5次测试EER的平均值作为模型的性能指标。GMM和ResNet18都是将评估数据集中训练集和开发集的特征向量输入后端分类模型进行训练，计算评估数据集中开发集的EER值。不同后端分类模型下EER的对比见表3。实验结果表明，在不同的后端分类模型上，前端模型提取出的特征都取得了较低的EER，而其中ResNet18模型取得了最低的EER，因此，后续实验的后端分类器模型都采用ResNet18模型。

表3　不同后端分类模型下EER的对比

分类器	EER
GMM	7.21%
五折交叉验证	5.67%
ResNet18	2.31%

在确定音高变换最佳范围和最佳后端分类器模型后，实验对评估数据集中语音样本采用不同手工提取的特征输入ResNet18后端分类器并进行比较分析，几种不同手工提取特征的合成伪造语音检测的EER对比见表4。实验结果表明，自监督对比学习比MFCC、LFCC和CQCC这3种手工特征性能更加优异，在EER数值上分别降低了61.73%、75.84%和57.74%。这是因为自监督对比学习模型可以更好地提取出语音信号中丰富的信息，其不同于手工提取的特征只考虑与频率有关的声学信息。

表4　几种不同手工特征的合成伪造语音检测的EER对比

前端特征	EER
LFCC	13.45%
MFCC	8.57%
CQCC	7.69%
SSCL+ResNet18	3.25%

此外，实验将基于自监督学习的合成伪造语音检测方法与基于深度学习的合成伪造语音检测方法进行了对比分析，其中，ECAPA-TDNN模型引入了一些增强功能来改进原始时延神经网络（time delay neural network，TDNN），首先利用SE（squeeze-and-excitation）块重新缩放卷积块中具有全局上下文信息的特征图的通道，并通过跳跃连接聚合多级特征，再利用Res2Net架构通过分组卷积进一步提高了性能，同时减少了参数总数。mGMM-MobileNet模型对于来自不同数据

增强方法的训练集数据采用不同的GMM拟合其特征分布，接着用对数高斯概率特征输入不同的分支网络。实验在评估数据集上开展，与几种基于深度学习的合成伪造语音检测方法的EER对比见表5。实验结果表明，基于自监督学习的合成伪造语音检测方法的性能明显优于基于深度学习的合成伪造语音检测方法的性能，SSCL+ResNet18系统相比于ECAPA-TDNN、mGMM-MobileNet和GMM-LCNN在EER上分别降低了40.47%、20.73%和10.22%。这是因为基于深度学习的合成伪造语音检测方法在面对真实语音和伪造语音的样本数量存在明显的不平衡数据集时，模型会倾向于更多地关注样本数量较多的类别，而忽略样本数量较少的类别，从而影响了系统的检测性能，而自监督对比学习利用借口任务从大量未标记数据中挖掘丰富的信息对模型进行训练，不依赖数据集标签数量是否平衡，因此基于自监督学习的合成伪造语音检测方法的性能优于基于深度学习的合成伪造语音检测方法的性能。

表5　与几种基于深度学习的合成伪造语音检测方法的EER对比

检测系统	EER
ECAPA-TDNN[25]	5.46%
mGMM-MobileNet[26]	4.10%
GMM-LCNN[27]	3.62%
SSCL+ResNet18	**3.25%**

为了进一步验证SSCL+ResNet18系统的性能，实验将SSCL+ResNet18系统与几种端到端检测系统的性能比较，其中RawNet2+RawBoost方法采用RawBoost对原始波形进行数据增强，在面对由未知编码和传输条件引起的干扰变化时提高了欺骗检测的准确性，同时降低了计算成本。实验在评估数据集上开展，与几种端到端检测系统合成伪造语音检测方法的EER对比见表6。由表6可以得出，SSCL+ResNet18系统优于RawNet2+RawBoost和GMM+LCNN这两个端到端的检测系统，SSCL+ResNet18系统相比于RawNet2+RawBoost和GMM+LCNN在EER上分别降低了38.79%和10.22%。这是由于端到端的伪造语音检测系统和传统的基于深度神经网络的检测系统一样，都需要大量的标记数据进行训练，才能获得良好的性能，而数据集中存在真实语音和伪造语音的样本数量存在明显的不平衡现象，导致端到端的伪造语音检测系统的性能不佳。

表6　与几种端到端检测系统合成伪造语音检测方法的EER对比

检测系统	EER
RawNet2+RawBoost[28]	5.31%
GMM+LCNN[29]	3.62%
SSCL+ResNet18	**3.25%**

为了进一步验证SSCL+ResNet18的性能以及泛化能力，将其与ASVspoof 2021挑战赛中的基线系统[30]进行了对比分析，与ASVspoof 2021基线系统的EER对比见表7。其中ASVspoof 2021挑战赛中的基线系统是采用所有攻击方式进行实验，其中包括语音合成攻击、语音转换攻击和语音深度伪造攻击。因此，SSCL+ResNet18在实验时也采用了数据集中所有的攻击方式。实验结果显示，SSCL+ResNet18的性能优于ASVspoof 2021挑战赛中的基线系统，并且在面对未知伪造语音的攻击时，SSCL+ResNet18依然有较低的EER值，这表明SSCL+ResNet18有良好的泛化能力。

表7　与ASVspoof 2021基线系统的EER对比

检测系统	EER
CQCC-GMM	15.62%
LFCC-GMM	19.30%
LFCC-LCNN	9.26%
RawNet2[31]	9.50%
SSCL+ResNet18	**5.64%**

4　结束语

本文提出了一种基于自监督对比学习的合成

伪造语音检测方法SSCL+ResNet18，该方法利用对比学习模型学习语音信号之间的相似性或者差异性，可以更好地区分真实语音和合成伪造语音，有效地提升了对合成伪造语音检测的准确率。该方法将经过音高变换后的样本视为负样本，训练神经网络使锚点样本特征与负样本特征不同，促使网络提取出对合成伪造语音敏感的特征，从而提高合成语音检测的准确性。实验结果表明，本文方法能够有效地检测出合成伪造语音，极大地提升了合成伪造语音检测系统的性能。

参考文献：

[1] 杨震, 王天朗, 郭海燕, 等. 跨域注意力特征融合的说话人确认方法[J]. 通信学报, 2023, 44(8):89-98.
YANG Z, WANG T L , GUO H Y, et al. Speaker verification method based on cross-domain attentive feature fusion[J]. Journal on Communications, 2023, 44(8): 89-98.

[2] 徐嘉, 简志华, 金宏辉, 等. 采用局部相位量化的合成语音检测方法[J]. 电信科学, 2024, 40(2): 63-71.
XU J, JIAN Z H, JIN H H, et al. A method for synthetic speech detection using local phase quantization[J]. Telecommunications Science, 2024, 40(2): 63-71.

[3] GARG D, GILL R. Deepfake generation and detection: an exploratory study[C]//Proceedings of 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). Piscataway: IEEE Press, 2023: 888-893.

[4] 金宏辉, 简志华, 杨曼, 等. 采用圆周局部三值模式纹理特征的合成语音检测方法[J]. 电信科学, 2023, 39(6): 85-95.
JIN H H, JIAN Z H, YANG M, et al. Synthetic speech detection method using texture feature based on circumferential local ternary pattern[J]. Telecommunications Science, 2023, 39 (6): 85-95.

[5] WINURSITO A, HIDAYAT R, BEJO A. Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition[C]//Proceedings of 2018 International Conference on Information and Communications Technology (ICOIACT). Piscataway: IEEE Press, 2018: 379-383.

[6] MON K Z, GALAJIT K, MAWALIM C O, et al. Spoof detection using voice contribution on LFCC features and ResNet-34[C]//Proceedings of 2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). Piscataway: IEEE Press, 2023:1-6.

[7] YANG J, DAS R K, LI H. Extended constant-Q cepstral coefficients for detection of spoofing attacks[C]//Proceedings of 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway：IEEE Press, 2018：1024-1029.

[8] 任延珍, 刘晨雨, 刘武洋, 等. 语音伪造及检测技术研究综述[J]. 信号处理, 2021, 37(12): 2412-2439.
REN Y Z, LIU C Y, LIU W Y, et al. A survey on speech forgery and detection[J]. Journal of Signal Processing, 2021, 37(12): 2412-2439.

[9] 陈暄, 吴吉义. 基于优化卷积神经网络的车辆特征识别算法研究[J]. 电信科学, 2023, 39(10): 101-111.
CHEN X, WU J Y. Research on vehicle feature recognition algorithm based on optimized convolutional neural network[J]. Telecommunications Science, 2023, 39(10): 101-111.

[10] FUJIMOTO M, KAWAI H. Comparative evaluations of various factored deep convolutional RNN architectures for noise robust speech recognition[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 4829-4833.

[11] ALAMSYAH R D, SUYANTO H. Speech gender classification using bidirectional long short term memory[C]//Proceedings of 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). Piscataway: IEEE Press, 2020: 646-649.

[12] ZHANG J, WANG T, WANG J, et al. A study of contrastive self-supervised learning generalization based on augmented data[C]//Proceedings of 2023 38th Youth Academic Annual Conference of Chinese Association of Automation (YAC). Piscataway: IEEE Press, 2023: 659-664.

[13] 张文林, 刘雪鹏, 牛铜, 等. 基于正样本对比与掩蔽重建的自监督语音表示学习[J]. 通信学报, 2022, 43(7): 163-171.
ZHANG W L, LIU X P, NIU T, et al. Self-supervised speech representation learning based on positive sample comparison and masking reconstruction[J]. Journal on Communications, 2022, 43(7): 163-171.

[14] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9726-9735.

[15] ELFAKI A, ASNAWI A L, JUSOH A Z, et al. Using the short-time Fourier transform and ResNet to diagnose depression from speech data[C]//Proceedings of 2021 IEEE International Conference on Computing (ICOCO). Piscataway: IEEE Press, 2021: 372-376.

[16] LIU A, LI J, YE H. A Prediction model combining convolutional neural network and LSTM neural network[C]//Proceedings of 2023 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS). Piscataway: IEEE Press, 2023: 318-321.

[17] WU Z, XIONG Y, YU S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of

2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3733-3742.

[18] MONTEIRO J, ALAM J, FALK T H. Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers[J]. Computer Speech & Language, 2020(63): 101096.

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

[20] ZHAO J, BAI X, CHEN Y, et al. Speech spoofing detection based on one-class residual attention network[C]//Proceedings of 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP). Piscataway: IEEE Press, 2023: 329-333.

[21] WU Z, KINNYNEN T, EVANS N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge[C]//Proceedings of Interspeech 2015. [S.l.:s.n.], 2015: 462-467.

[22] BHUKYA K, RAJ A, RAJA D N. Audio deepfakes: feature extraction and model evaluation for detection[C]//Proceedings of 5th International Conference for Emerging Technology (INCET). Piscataway: IEEE Press, 2024: 1-6.

[23] MATHEW L R, ANSELAM A S, PILLAI S S. Analysis of LD-CELP coder output with sound eXchange and Praat software[C]//Proceedings of IEEE International Conference on Advanced Communications, Control and Computing Technologies. Piscataway: IEEE Press, 2014: 1281-1285.

[24] SATRIA A, SITOMPUL O S, MAWENGKANG H. 5-fold cross validation on supporting K-nearest neighbour accuration of making consimilar symptoms disease classification[C]//Proceedings of International Conference on Computer Science and Engineering (IC2SE). Piscataway: IEEE Press, 2021: 1-5.

[25] MARTÍN-DOÑAS M J, ÁLVAREZ A. The vicomtech audio deepfake detection system based on Wav2vec2 for the 2022 ADD challenge[C]//Proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022:9241-9245.

[26] 温燕 . 基于多分支卷积神经网络的合成与转换语音检测研究[D]. 南昌: 江西师范大学, 2024.
WEN Y. Research on synthetic and converted speech detection based on multi-branch convolutional neural network[D]. Nanchang: Jiangxi Normal University, 2024.

[27] DAS R K. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021[C]//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Paris：International Speech Communication Association, 2021: 29–36.

[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]//Proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6382-6386.

[29] LEI Z, YAN H, LIU C, et al. GMM-ResNet2: ensemble of group resnet networks for synthetic speech detection[C]//Proceedings of ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 12101-12105.

[30] 余惠 . 基于改进的 Transformer 模型的合成语音伪造检测研究[D]. 南昌: 江西师范大学, 2023.
YU H. Research on synthetic speech deepfake detection based on improved transformer[D]. Nanchang: Jiangxi Normal University, 2023.

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

[作者简介]

杨曼（2000– ），女，杭州电子科技大学通信工程学院硕士生，主要研究方向为伪造语音检测。

简志华（1978– ），男，博士，杭州电子科技大学通信工程学院副教授、硕士生导师，主要研究方向为伪造语音检测、语音中的隐私保护、语音转换与生成等。

梁承涵（2001– ），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为伪造语音检测与声纹鉴伪。