一种采用激活函数的具有噪声鲁棒性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：在现实应用场景中，攻击者在伪造语音中加入加性噪声或者混响等干扰，会导致经纯净语音训练得到的检测系统性能急剧下降，为此，通过设计一种激活函数代替残差网络中的跳跃连接，实现了具有噪声鲁棒性的合成语音检测系统。通过分析不同激活函数对残差块跳跃连接的影响，将输入特征划分为非显著特征、显著特征和无法判断特征，提出了一个新的激活函数，并通过方差增长的算法来寻找激活函数的最优参数。实验结果表明，与现有方法相比，不仅显著降低了系统的等错误率，而且对噪声干扰具有很好的鲁棒性。

关键词：伪造语音检测；合成语音检测；激活函数；噪声鲁棒性
中图分类号：TN912
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2026038

A noise-robust synthetic spoofing speech detection method using activation function

Yang Man, Jian Zhihua, Liang Chenghan
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: In real-world application scenarios, attackers often add additive noise or reverberation and other interferences to the spoofing speech, which will cause the performance of the detection system trained with clean speech to drop sharply. Therefore, an activation function was designed to replace the skip connection in the residual network, thereby proposing a synthetic speech detection system with noise-robust. After analyzing the influence of different activation functions on the skip connection of the residual block, the input features were divided into non-significant features, significant features and undetermined features, and a novel activation function was proposed. The optimal parameters of the activation function were determined through the method of variance growth. Experimental results show that compared with existing methods, the method proposed not only significantly reduces the equal error rate of the system, but also has good robustness to noise interference.

Key words: spoofing speech detection, synthetic speech detection, activation function, noise-robust

收稿日期：2025−06−19；**修回日期：**2025−09−08
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)

0　引言

在社会需求不断增加的背景下，语音生成技术进入了一个迅猛发展的阶段，从最初的简单语音处理到如今深度学习、自然语音处理技术的应用，语音生成技术使得机器能够更加精准地合成人类语言[1]。然而语音生成技术的发展在带来显著便利的同时，也暴露了许多潜在的安全风险[2-3]。例如，语音生成技术使得伪造语音能够绕过自动说话人验证（automatic speaker verification，ASV）系统进行身份冒充和非法交易，给社会带来极大的危害[4]。目前常见的语音伪造方式包括语音模仿、语音合成、语音转换以及重放攻击[5]。随着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[6]，语音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逐渐减少，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[7]。因此，为有效提高ASV系统的可靠性，对合成语音检测（synthetic speech detection，SSD）展开研究具有非常重要的意义[8]。

SSD系统一般由前端特征提取模块和后端分类器模块两部分组成[9]，其中前端模块通过分析语音信号来提取具有区分性的声学特征，后端模块通过分类器判断输入的语音特征是真实语音还是合成的伪造语音[10]。SSD系统后端模块的研究历程大致上可以分为两个阶段：一是传统的通过机器学习方法对前端提取的手工特征进行分类，二是随着深度学习的发展，通过神经网络算法对前端生成特征进行深度特征提取和高维度表征，随后进行分类判决。传统的SSD后端模块有基于高斯混合模型[11]（Gaussian mixture model，GMM）的分类器和基于支持向量机[12]（support vector machine，SVM）的分类器。将真实语音和合成语音分别作为真实类和伪造类用来训练GMM，通过对比不同类别的对数似然比值并结合预先设定的分类阈值，实现真伪语音判决。但是GMM在面对少量训练数据时，会试图拟合尽可能多的模型参数，以便尽可能准确地描述数据分布。这通常会导致模型过于复杂，并且训练数据中的噪声会干扰GMM拟合过程，导致估计不准确从而影响聚类效果。随着机器学习的发展，SVM通常能通过最大化分类边界的间隔来避免过拟合，从而具有很好的泛化能力。即使训练数据量较小，SVM往往能够提取数据中的主要结构，并且更容易捕捉到数据的总体趋势，因此SVM可以有效地寻找决策边界，防止过拟合。然而SVM对于数据中的噪声非常敏感，特别是较强的噪声干扰，因此SVM作为SSD后端模型在现实有噪声的环境中会降低分类性能。

随着深度学习的发展，基于深度神经网络的分类器逐渐变成SSD研究的主流方向，尤其是以卷积神经网络（convolutional neural network，CNN）为核心的网络架构。基于CNN分类器的主要工作原理是通过多批次训练不断提升神经网络的学习能力，获得更具有区分性的特征表示，即令同类样本之间的特征表示更接近而异类样本之间的特征表示更远离，因此神经网络可以获得更好的分类效果。大多数研究在面对大量训练数据时会为了提升检测准确率而不断地增加CNN的网络层数，但这种做法会导致梯度消失等一系列后果，进而导致该网络架构的低层网络无法接收到训练信息而不能及时更新[13]。因此He等[14]提出了残差网络（residual network，ResNet），该网络在CNN的基础上在网络层之间引入跳跃连接，使得网络参数能够更高效地向较低层传递更新信息。基于这样的改进，神经网络梯度消失问题得到了有效改善，相比于CNN取得了更好的检测效果。Alzantot等[15]在SSD系统前端将3种不同的前端特征进行特征融合，后端分类模型使用深度ResNet的网络架构，该模型通过分数融合不同

的特征使得深度ResNet可以更好地学习高级的特征表示，提升了检测准确率。Gao等[16]公开了基于Res2Net的后端模型，该模型将ResNet中的卷积分解成多个子模块，因此可以沿着通道维度将输入特征分割为多个组进行跨组特征聚合，这样获得的特征增加了多个特征尺度，提升了检测准确率和模型的泛化能力。Parasu等[17]提出了基于Light-ResNet的模型结构，该模型在ResNet的基础上大大降低了网络层数以及复杂度，减少了参数，防止过拟合的同时实现网络轻量化。然而ResNet在面对含噪语音输入时，噪声会通过残差块中的跳跃连接直接传递到输出端，对语音检测系统造成直接干扰，从而降低检测准确性。

在实际伪造语音检测场景中，攻击者会在攻击过程中向伪造语音加入加性噪声或者混响等干扰，导致纯净语音训练的伪造语音检测系统往往达不到预期的检测效果。为了增强检测系统的噪声鲁棒性，针对检测系统中的后端模块，在前期研究基础上[18]，提出了一种具有噪声鲁棒性的伪造语音检测方法SSCL-actResNet18。它通过分析不同激活函数对残差块跳跃连接的影响后将输入特征划分为非显著特征、显著特征和无法判断特征，并据此设计了一个新的激活函数实现跳跃连接，再通过方差增长的算法来寻找激活函数的最优参数，显著增强了检测系统的噪声鲁棒性。

1 网络架构actResNet18

1.1 ResNet18残差块中噪声传播计算式

残差块结构对比如图1所示。ResNet18模块中的残差块如图1（a）所示，ResNet18以CNN为基础在网络层之间引入跳跃连接，使得网络参数能够更高效地向较低层传递更新信息。这一改进有效缓解了网络层数过多时出现的梯度消失问题，取得了更好的检测效果。

由此可得，残差块输出为：

$$y_l=\mathcal{F}\left(x_l,W_l\right)+x_l=W_l\odot x_l+x_l \tag{1}$$

其中，x_l和y_l分别为第l个残差块的输入与输出，W_l表示第l个残差块的权重和偏差，$\mathcal{F}$是权重层的函数，它包括两个卷积层、ReLU激活和批量归一化，⊙为哈达玛积操作。

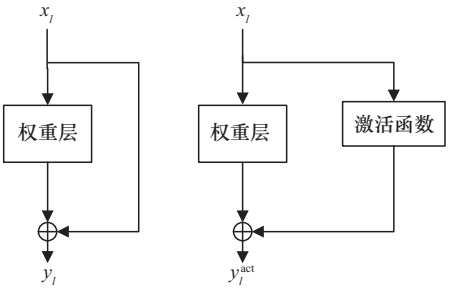

图1 残差块结构对比

在含噪语音检测过程中，残差块输入为$x_l=\hat{x}_l+\epsilon_l$，其中$\hat{x}_l$和ϵ_l分别为输入信号中的纯净语音和噪声。噪声会通过跳跃连接直接传递到输出端，对语音检测系统造成直接干扰，从而降低检测准确性，此时残差块输出为：

$$\hat{y}_l=W_l\odot\hat{x}_l+W_l\odot\epsilon_l+\hat{x}_l+\epsilon_l \tag{2}$$

则第l个残差块含噪语音和纯净语音输出信号之差值χ_l为：

$$\chi_l=\left|y_l-\hat{y}_l\right|=\left|W_l\odot\epsilon_l+\epsilon_l\right|=\left|\left(W_l+1\right)\odot\epsilon_l\right| \tag{3}$$

由式（3）可知，χ_l随着残差块中W_l的增加而增加。为了解决这一问题，提出了一种改进残差块结构，如图1（b）所示。通过在跳跃连接中添加激活函数来强化反伪造特征的提取能力，同时有效抑制跳跃连接中噪声的传播，此时残差块输出为：

$$y_l^{\mathrm{act}}=\mathcal{F}\left(x_l,W_l\right)+\mathrm{act}\left(x_l\right) \tag{4}$$

其中，$\mathrm{act}(\cdot)$为激活函数，则此时第l个残差块含噪语音和纯净语音输出信号的差值χ_l^{act}为：

$$\chi_l^{\mathrm{act}}=\left|W_l\odot\epsilon_l+\mathrm{act}\left(\hat{x}_l+\epsilon_l\right)-\mathrm{act}\left(\hat{x}_l\right)\right| \tag{5}$$

由式（5）可知，有两种情况的$\mathrm{act}(\cdot)$函数会降低噪声对χ_l^{act}的影响。

情况一：保持有噪语音x_l不变，$\mathrm{act}(\cdot)$函数增

加纯净语音特征$\hat{x}_l$和噪声特征ϵ_l之比。在这种情况下$\mathrm{act}(\hat{x}_l+\epsilon_l)\approx\mathrm{act}(\hat{x}_l)$，则：

$$\chi_l^{\mathrm{act}}\approx|W_l\odot\epsilon_l| \tag{6}$$

此时χ_l^{act}小于χ_l，因此$\mathrm{act}(\cdot)$函数通过增加纯净语音特征来降低噪声的影响。

情况二：保持纯净语音特征$\hat{x}_l$和噪声特征ϵ_l之比不变，$\mathrm{act}(\cdot)$降低跳跃连接中输入特征x_l和纯净语音特征$\hat{x}_l$。在极限情况下$\mathrm{act}(\hat{x}_l+\epsilon_l)=\mathrm{act}(\hat{x}_l)=0$，则：

$$\chi_l^{\mathrm{act}}=|W_l\odot\epsilon_l| \tag{7}$$

此时χ_l^{act}远小于χ_l，因此$\mathrm{act}(\cdot)$通过去除纯净语音特征来降低噪声的影响。

当有噪语音输入中纯净语音较大时，应用情况一中的激活函数。当有噪语音输入中噪声较大时，应用情况二中的激活函数。因此面对不同的有噪语音输入时，设置一个清晰的边界以灵活应用不同的激活函数$\mathrm{act}(\cdot)$可以降低噪声对残差块输出的影响。

1.2　ReLU代替跳跃连接可行性分析

由于ReLU激活函数的特性可以丢弃负输入，在传统的跳跃连接上放置一个ReLU激活函数后，第l个残差块含噪语音和纯净语音输出信号之差值χ_l^R为：

$$\chi_l^R=\left|W_l\odot\epsilon_l+\sigma(\hat{x}_l+\epsilon_l)-\sigma(\hat{x}_l)\right| \tag{8}$$

其中，$\sigma(\cdot)$为ReLU激活函数。

根据式（8），可以通过以下4种情况比较χ_l^R和χ_l。

（1）当$\hat{x}_l+\epsilon_l\geqslant0,\hat{x}_l\geqslant0$时，$\sigma(\hat{x}_l+\epsilon_l)-\sigma(\hat{x}_l)=\epsilon_l$，$\chi_l^R=|W_l\odot\epsilon_l+\epsilon_l|$，此时$\chi_l^R=\chi_l$。

（2）当$\hat{x}_l+\epsilon_l\geqslant0,\hat{x}_l<0$时，$\sigma(\hat{x}_l+\epsilon_l)-\sigma(\hat{x}_l)=\hat{x}_l+\epsilon_l$，$\chi_l^R=|W_l\odot\epsilon_l+\hat{x}_l+\epsilon_l|$，此时$\chi_l^R<\chi_l$。

（3）当$\hat{x}_l+\epsilon_l<0,\hat{x}_l\geqslant0$时，$\sigma(\hat{x}_l+\epsilon_l)-\sigma(\hat{x}_l)=-\hat{x}_l$，$\chi_l^R=|W_l\odot\epsilon_l-\hat{x}_l|$，此时$\chi_l^R<\chi_l$。

（4）当$\hat{x}_l+\epsilon_l<0,\hat{x}_l<0$时，$\sigma(\hat{x}_l+\epsilon_l)-\sigma(\hat{x}_l)=0$，$\chi_l^R=|W_l\odot\epsilon_l+\epsilon_l|$，此时$\chi_l^R<\chi_l$。

4种情况下均使得$\chi_l^R<\chi_l$，在传统跳跃连接上放置一个ReLU激活函数以保留正值作为有用特征并去除负值作为噪声。特别是在输入为负值时，满足激活函数情况二即抑制噪声，因此可以减少噪声对残差块输出的影响，进而提高检测准确率。

1.3　最大池化层代替跳跃连接可行性分析

假设一个形状为$[h\times w]$的感受野S，包含噪声ϵ_l和纯净语音$\hat{x}_{i,j}^S$。由于最大池化函数的特性会将每个元素增加到该感受野中的最大值，该感受野再经过最大池化激活函数后包含同样的值即$x_{i,j}^S=\hat{x}_{i,j}^S\pm|\epsilon_l|$，其中$i$、$j$分别对应大小为$h\times w$的感受野中的坐标。假设$r_{i,j}=|\epsilon_l|/\hat{x}_{i,j}^S$，则传统残差块跳跃连接中纯净语音和噪声的差值可表示为：

$$D(i,j)=\left|\hat{x}_{i,j}^S\right|-\left|r_{i,j}\cdot\hat{x}_{i,j}^S\right| \tag{9}$$

则经过最大池化残差块跳跃连接后，纯净语音和噪声的差值可表示为：

$$D^M(i,j)=\left|\hat{x}_{u,v}^S\right|-\left|r_{i,j}\cdot\hat{x}_{u,v}^S\right| \tag{10}$$

其中，u、v分别对应大小为$h\times w$感受野中的坐标，$\hat{x}_{u,v}^S$为在感受野S内的最大有噪语音$x_{u,v}^S$中的纯净语音。因此感受野S内的平均差值为：

$$\Delta\bar{D}=\frac{1}{hw}\cdot\sum_{(i,j)\in S}\left[D^M(i,j)-D(i,j)\right]=\frac{1}{hw}\cdot\sum_{(i,j)\in S}\left(1-r_{i,j}\right)\left(\left|\hat{x}_{u,v}^S\right|-\left|\hat{x}_{i,j}^S\right|\right) \tag{11}$$

则由式（11）可知，当$\Delta\bar{D}>0$时说明最大池化激活函数增加了纯净语音特征$\hat{x}_l$和噪声特征ϵ_l之比，满足情况一。则可以通过以下3种情况分析$\Delta\bar{D}$。

（1）所有$\hat{x}_{u,v}^S\geqslant0$，此时$\left(\left|\hat{x}_{u,v}^S\right|-\left|\hat{x}_{i,j}^S\right|\right)\geqslant0$，则$\Delta\bar{D}>0$。

（2）所有$\hat{x}_{u,v}^S\leqslant0$，此时$\left(\left|\hat{x}_{u,v}^S\right|-\left|\hat{x}_{i,j}^S\right|\right)<0$，则$\Delta\bar{D}<0$。

（3）一些 $\hat{x}_{u,v}^{S}\geqslant 0$ 而另一些 $\hat{x}_{u,v}^{S}\leqslant 0$，此时 $\left(\left|\hat{x}_{u,v}^{S}\right|-\left|\hat{x}_{i,j}^{S}\right|\right)$ 无法判断大小。

分析3种情况可知，当输入是正值时，最大池化激活函数可以增加纯语音来降低噪声的影响，满足情况一。此时可以减少噪声对残差块输出的影响，进而提高检测准确率。

1.4 构建激活函数

由上文分析可知，所需激活函数应该满足以下3个要求：（1）非显著特征被移除；（2）显著特征被加强；（3）无法判定特征被抑制。然而ReLU激活函数将正值作为有用特征保留并将负值作为噪声特征被移除，这一特性只满足要求（1）。最大池化激活函数当输入为正值时，将最大元素视为重要特征并增强，这一特性只满足要求（2）。为此，提出一个新的激活函数，旨在抑制噪声特征并增强重要特征，该函数可表示为：

$$z_l=\mathbf{act}_l\odot x_l \tag{12}$$

其中，$\mathbf{act}_l$ 是第 l 层与输入特征 x_l 维度相同的权重矩阵，$\mathbf{act}_l^{i,j}$ 表示 $x_l^{i,j}$ 对 $z_l^{i,j}$ 的增强或者抑制程度。由于合成语音检测系统的前端提取的特征为 1×256 维，则令 R_u^{med} 为 $x_l^{i,j}$ 周围 $[1\times w]$ 感受野中的中位数。在构建激活函数 $\mathbf{act}_l^{i,j}$ 时，设计了以下规则。

（1）如果 $x_l^{i,j}<0$，则类似于ReLU激活函数的操作将 $x_l^{i,j}$ 视为噪声被移除，此时令 $\mathbf{act}_l^{i,j}=0$。

（2）如果 $0<x_l^{i,j}<R_u^{\mathrm{med}}$，则类似于最大池化激活函数的操作将 $x_l^{i,j}$ 视为无法判定特征应被抑制，此时令 $\mathbf{act}_l^{i,j}<1$，则使用指数函数作为激活函数，即：

$$\mathbf{act}_l^{i,j}=\mathrm{e}^{\mathrm{cur}\left(x_l^{i,j}-R_u^{\mathrm{med}}\right)} \tag{13}$$

其中，cur为曲率函数。

（3）如果 $x_l^{i,j}>R_u^{\mathrm{med}}$，则类似于最大池化激活函数的操作将 $x_l^{i,j}$ 视为重要特征应被增强，此时令 $\mathbf{act}_l^{i,j}>1$ 并且快速增加，则使用一元二次函数作为激活函数，即：

$$\mathbf{act}_l^{i,j}=\frac{1}{\left(R_u^{\mathrm{med}}\right)^2}\left(x_l^{i,j}\right)^2 \tag{14}$$

综上所述，激活函数 $\mathbf{act}_l^{i,j}$ 可表示为分段函数：

$$\mathbf{act}_l^{i,j}=\begin{cases}0, & x_l^{i,j}<0\\ \mathrm{e}^{\mathrm{cur}\left(x_l^{i,j}-R_u^{\mathrm{med}}\right)}, & 0<x_l^{i,j}<R_u^{\mathrm{med}}\\ \left(R_u^{\mathrm{med}}\right)^{-2}\left(x_l^{i,j}\right)^2, & x_l^{i,j}>R_u^{\mathrm{med}}\end{cases} \tag{15}$$

为了使得提出的残差网络后端模型actResNet18可以去除输入特征中的噪声特征，目标转化为最大化激活函数输出 z_l 的方差并且使得 z_l 小于权重层输出 $\mathcal{F}(x_l)$，再通过拉格朗日法[21]生成cur的候选解，并根据方差和约束条件筛选最优解。

2 合成伪造语音检测

检测系统的前端采用前期工作中提出的基于自监督对比学习的SSCL模型[18]，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短时傅里叶变换（short-time Fourier transform，STFT）处理得到频谱图，并输入前端自监督对比学习模型，在经过循环卷积神经网络处理后最终得到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。

2.1 后端模型

为了减少噪声对合成伪造语音检测准确性的影响，提出了基于激活函数 $\mathbf{act}_l^{i,j}$ 的actResNet18后端模型，actResNet18后端模型整体框架如图2所示。

在检测系统的后端模块，首先，actResNet18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，将特征输入基于激活函数的残差块，该残差块不仅保持了跳跃

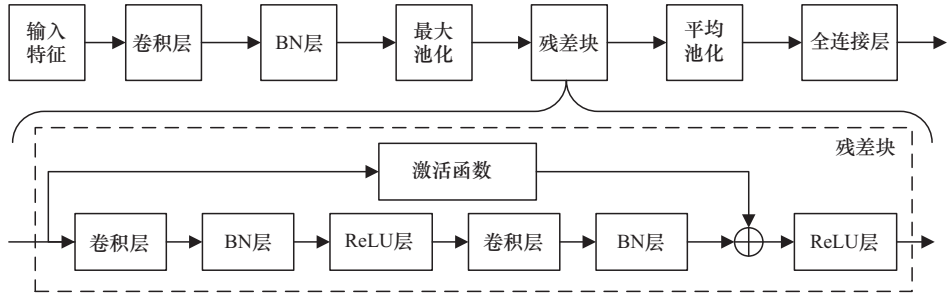

图2　actResNet18后端模型整体框架

连接，具有网络不会出现梯度消失和梯度爆炸的优点，而且可以降低噪声特征通过跳跃连接对残差块输出的影响；然后，将全连接层之前的输出作为语音的嵌入向量；最后，将嵌入向量输入全连接层映射为真实/合成语音的概率，从而完成真伪判决。

2.2　合成语音检测系统

合成语音检测系统由基于SSCL的前端模块和基于actResNet18的后端模块组成，前端模块在预训练结束后提取语音的自监督特征，并将该特征输入后端模块以判断该语音为合成语音还是真实语音，合成语音检测系统SSCL-actResNet18流程如图3所示。

前端模型训练时，首先，从任一训练批次的M个样本中任选一语音样本x_i并做两次随机剪裁，再通过短时傅里叶变换生成对应语谱图，得到锚点样本x^{query}和正样本x_+，同时对x_i做音高变换处理后应用短时傅里叶变换生成负样本x_0^{key}。然后，再将该训练批次中其余$M-1$个语音样本用同样的方法，即经音高变换后应用短时傅里叶变换生成语谱图，得到$M-1$个负样本$\left\{x_m^{\text{key}}, m=1,2,\cdots,M-1\right\}$。然后将$x_i$生成的$x^{\text{query}}$和$x_+$分别输入编码器和动量编码器，得到固定大小的特征表示$f_\varphi\left(x^{\text{query}}\right)$和$f_\phi(x_+)$，$x_i$生成的$x_0^{\text{key}}$和其余$M-1$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$\left\{f_\phi\left(x_m^{\text{key}}\right), m=1,\cdots,M-1\right\}$，并经过投影生成查询$q=g\left(f_\varphi\left(x^{\text{query}}\right)\right)$和键值$k_+=g\left(f_\phi(x_+)\right)$、$k_0=g\left(f_\phi\left(x_0^{\text{key}}\right)\right)$、$\left\{k_m=g\left(f_\phi\left(x_m^{\text{key}}\right)\right), m=1,\cdots,M-1\right\}$。最后，计算对比损失梯度更新编码器，同时动量更新动量编码器。

后端模型训练时，首先，将训练集和开发集的特征向量作为actResNet18输入网络，输出每个语音样本的特征向量，通过最大化激活函数输出的方差并且使得小于权重层输出生成cur的候选解，并根据方差和约束条件筛选最优解，再使用OC-Softmax损失函数[19]调整优化网络模型的参数。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

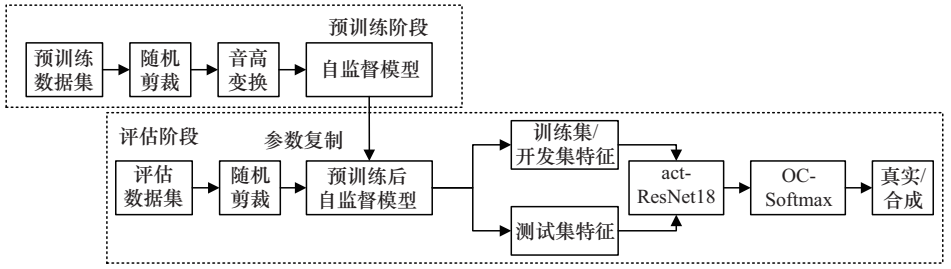

图3　合成语音检测系统SSCL-actResNet18流程

个语音样本的嵌入向量。最后，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。

3 实验结果和性能评估

3.1 数据集

为了训练检测系统的前端模型，使用的语料库是ASVspoof 2015挑战赛[20]和ASVspoof 2017挑战赛[21]中的逻辑访问（logical access，LA）场景数据集，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，预训练数据集信息见表1。

表1 预训练数据集信息

语料库	子集	真实语音/条	伪造语音/条	总语音/条
ASVspoof 2015 LA	—	16 651	—	16 651
ASVspoof 2017 LA	—	3 511	—	3 511

为了训练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，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合成语音检测的性能评估，评估数据集信息见表2。

表2 评估数据集信息

语料库	子集	真实语音/条	伪造语音/条	总语音/条
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

由于ASVspoof 2021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，合成语音主要采用神经声学模型和深度学习方法生成。而ASVspoof 2021 LA测试集在ASVspoof 2019 LA测试集的基础上进行了扩展，进一步考虑了电话编码和传输过程中对ASV系统的影响。重要拓展是纳入了信道转换环节，以模拟语音数据经由不同电话系统传输的场景，例如基于网络协议的语音传输（VoIP）以及公共交换电话网络（PSTN）等实际应用环境。同时，无论是真实语音数据还是伪造语音数据，都考虑了7种不同的编解码器作为传输流程的一部分。该设计旨在检验并增强ASV系统在面对复杂攻击和真实电话网络条件下的稳健性与泛化能力。

实验使用NoiseX-92噪声数据集[23]，该数据集是由英国感知技术研究院测量的噪声数据集，按照不同信噪比与评估数据集的语音进行混合得到带噪评估数据集，以模拟SSD的噪声环境，将带噪评估数据集的测试集特征向量输入actResNet18以验证SSCL-actResNet18的噪声鲁棒性。

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actResNet18后端模型实验参数设置，使用Adam优化模型参数，其中，参数$\beta_1=0.9$、$\beta_2=0.999$，用于更新ResNet18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$\rho=20$，$m_0=0.9$，$m_1=0.2$，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）来评价伪装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）的计算表达式为[24]：

$$P_{\mathrm{FAR}}(\lambda)=\frac{\text{得分大于阈值}\lambda\text{的伪造语音数量}}{\text{伪造语音总数}} \tag{16}$$

$$P_{\mathrm{FRR}}(\lambda)=\frac{\text{得分小于或等于阈值}\lambda\text{的伪造语音数量}}{\text{真实语音总数}} \tag{17}$$

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

首先实验使用纯净语音即评估数据集训练actResNet18后端模型，验证SSCL-actResNet18提高合成语音检测系统性能的能力以及在面对复杂攻击和真实电话网络条件下的稳健性与泛化能力。

在评估SSCL-actResNet18检测性能之前，实验评估了不同激活函数代替跳跃连接后对系统性能的影响。将ReLU激活函数、max-pooling激活函数和本文提出的激活函数分别代替跳跃连接得到后端模型ReLUResNet18、maxResNet18和actResNet18。将评估数据集中训练集和开发集的特征向量输入不同激活函数代替跳跃连接后的后端模型进行训练，并计算评估数据集中开发集的EER值，不同激活函数的性能对比见表3，其中ResNet18表示采用跳跃连接的网络。实验结果表明，ReLU激活函数和max-pooling激活函数代替跳跃连接都会降低检测系统的EER值。其中ReLU激活函数代替跳跃连接只限制输入特征中的负值而不会增强显著特征，并且非显著特征也可能是噪声特征。而max-pooling激活函数代替跳跃连接会增强显著性特征，该显著性特征为反伪造特征。因此ReLUResNet18和maxResNet18分别通过抑制噪声和增强反伪造特征提高了检测系统的检测性能，相比于ReLUResNet而言，maxResNet检测性能更好。同时可以看到，本文提出的激活函数对于检测系统性能的提升更加明显，该激活函数结合了ReLU激活函数和max-pooling激活函数两者的优势，不仅可以抑制噪声特征而且可以增强反伪造特征，因此可以更好地提升系统的检测性能。

表3　不同激活函数的性能对比

激活函数	EER
ResNet18	2.31%
ReLUResNet18	2.27%
maxResNet18	2.11%
actResNet18	**1.87%**

实验将评估数据集的训练集和开发集的特征向量输入不同激活函数代替跳跃连接后的后端模型进行训练，训练结束后将测试集的特征向量输入SSCL-actResNet18系统得到EER值，并将SSCL-actResNet18与主流的基于深度学习的合成伪造语音检测方法进行了对比分析，与基于深度学习的合成语音伪造检测方法的EER对比见表4。由表4可知，SSCL-actResNet18系统的性能不仅优于现有方法，也在前期工作SSCL-ResNet18的基础上将EER降低了7.38%。这是因为SSCL-actResNet18系统不仅考虑了合成语音存在韵律泄漏的情况，还通过设计激活函数来代替残差块的跳跃连接，从而限制噪声对残差块输出的影响，提高合成语音检测的准确性。

表4　与基于深度学习的合成语音伪造检测方法的EER对比

检测系统	EER
ECAPA-TDNN[25]	5.46%
RawNet2+RawBoost[26]	5.31%
mGMM-MobileNet[27]	4.10%
GMM-LCNN[28]	3.62%
SSCL-ResNet18[18]	**3.25%**
SSCL-actResNet18	**3.01%**

为了进一步验证SSCL-actResNet18系统的泛化能力，与ASVspoof 2021挑战赛中的基线系统进行了对比分析。ASVspoof 2021 LA测试集数据

中无论是真实语音数据还是伪造语音数据都考虑了7种不同的编解码器作为信道传输的一部分，在传输过程中不可避免会出现噪声干扰问题，非常考验检测系统在面对复杂攻击和真实电话网络条件下的鲁棒性与泛化能力。与ASVspoof 2021基线系统的EER对比见表5。由表5可得，SSCL-actResNet18的性能均优于ASVspoof 2021挑战赛中的基线系统，表明SSCL-actResNet18有良好的泛化能力和噪声鲁棒性。

表5　与ASVspoof 2021基线系统的EER对比

检测系统	EER
CQCC-GMM[29]	15.62%
LFCC-GMM[29]	19.30%
LFCC-LCNN[29]	9.26%
RawNet2[30]	9.50%
SSCL-ResNet18[18]	**5.64%**
SSCL-actResNet18	**3.01%**

为进一步验证SSCL-actResNet18系统中激活函数的噪声鲁棒性，实验选取NoiseX-92数据库中的噪声，分别按照−5 dB、0 dB、5 dB和10 dB的信噪比生成带噪语音以模拟SSD的噪声环境，用来评估SSCL-actResNet18系统的噪声鲁棒性。实验在带噪评估数据集下将SSCL-actResNet18系统与SSCL-ResNet18系统进行对比分析，噪声鲁棒性性能对比见表6。从实验结果来看，SSCL-actResNet18由于采用了激活函数去代替跳跃连接，在各种信噪比的情况下性能都优于SSCL-ResNet18，这表明SSCL-actResNet18不仅提升了SSD检测系统的性能，而且具有很好的噪声鲁棒性。

表6　噪声鲁棒性性能对比

检测系统	EER				
	−5 dB	0 dB	5 dB	10 dB	平均
SSCL-ResNet18	19.6%	15.37%	13.61%	12.82%	**15.35%**
SSCL-actResNet18	12.86%	10.16%	6.52%	5.90%	**8.86%**

4　结束语

本文提出了一个新的激活函数并用于深度神经网络中残差块的跳跃连接，结合基于自监督对比学习的前端模型构建了具有噪声鲁棒性的SSD系统SSCL-actResNet18。通过分析不同激活函数对残差块跳跃连接的影响后，将输入特征划分为非显著特征、显著特征和无法判断特征，并据此设计了一个新的激活函数来实现跳跃连接，再通过方差增长的算法来寻找激活函数的最优参数。实验结果表明，SSCL-actResNet18系统相比传统检测方法在EER性能上具有明显的优势，同时与SSCL-ResNet18相比，由于采用了激活函数代替跳跃连接，SSCL-actResNet18具有非常好的噪声鲁棒性，进一步提升了系统的泛化能力。

参考文献：

[1] 付毅冲. 零样本个性化语音合成的研究[D]. 北京: 北京邮电大学, 2025.
Fu Y C. Research on zero-shot personalized speech synthesis[D]. Beijing: Beijing University of Posts and Telecommunications, 2025.

[2] 乔喆. 人工智能生成内容技术在内容安全治理领域的风险和对策[J]. 电信科学, 2023, 39(10): 136-146.
Qiao Z. Risks and countermeasures of artificial intelligence generated content technology in content security governance[J]. Telecommunications Science, 2023, 39(10): 136-146.

[3] Huang W, Gu Y M, Wang Z M, et al. Generalizable audio deepfake detection via latent space refinement and augmentation[C]// Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2025: 1-5.

[4] Isolde W, Dagmar B, Vincent H, 等. 欧洲法庭科学联盟说话人鉴定方法的实践指南[J]. 中国语音学报, 2024(1): 93-101.
Isolde W, Dagmar B, Vincent H, et al. Practical guide to the European network of forensic speaker typing methods[J]. Chinese Journal of Phonetics, 2024(1): 93-101.

[5] 许裕雄, 李斌, 谭舜泉, 等. 语音深度伪造及其检测技术研究进展[J]. 中国图象图形学报, 2024, 29(8): 2236-2268.

Xu Y X, Li B, Tan S Q, et al. Research progress on speech deepfake and its detection techniques[J]. Journal of Image and Graphics, 2024, 29(8): 2236-2268.

[6] Mo Y C, Wang S L. Multi-task learning improves synthetic speech detection[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6392-6396.

[7] 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.

[8] Mutica I, Mihalache S, Burileanu D. Synthetic speech detection using deep neural networks[C]//Proceedings of the 2024 47th International Conference on Telecommunications and Signal Processing (TSP). Piscataway: IEEE Press, 2024: 53-57.

[9] Li C T, Yang F R, Yang J. The role of long-term dependency in synthetic speech detection[J]. IEEE Signal Processing Letters, 2022, 29: 1142-1146.

[10] Liu C, Xu X L, Xiao F. ASSD: an AI-synthesized speech detection scheme using whisper feature and types classification[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025, 33: 542-556.

[11] Bhukya R K, Raj A. Automatic speaker verification spoof detection and countermeasures using Gaussian mixture model[C]//Proceedings of the 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). Piscataway: IEEE Press, 2022: 1-6.

[12] Rahmeni R, Ben A A, Ben A Y. Speech spoofing detection using SVM and ELM technique with acoustic features[C]//Proceedings of the 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). Piscataway: IEEE Press, 2020: 1-4.

[13] Yu H, Tan Z H, Ma Z Y, et al. Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(10): 4633-4644.

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

[15] Alzantot M, Wang Z Q, Srivastava M B. Deep residual neural networks for audio spoofing detection[C]//Proceedings of the Interspeech 2019. Farmington Hills: Cengage Learning, 2019: 1078-1082.

[16] Gao S H, Cheng M M, Zhao K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662.

[17] Parasu P, Epps J, Sriskandaraja K, et al. Investigating light-ResNet architecture for spoofing detection under mismatched conditions[C]//Proceedings of the Interspeech 2020. Farmington Hills: Cengage Learning, 2020: 1111-1115.

[18] 杨曼, 简志华, 梁承涵. 采用自监督对比学习的合成伪造语音检测方法[J]. 电信科学, 2024, 40(11): 40-49.
Yang M, Jian Z H, Liang C H. A method of synthetic spoofing speech detection using self-supervised contrastive learning[J]. Telecommunications Science, 2024, 40(11): 40-49.

[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] Wu Z Z, Kinnunen T, Evans N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge[C]//Proceedings of the Interspeech 2015. Farmington Hills: Cengage Learning, 2015: 2037-2041.

[21] Kinnunen T, Sahidullah M, DELGADO H, et al. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection[C]//Proceedings of the Interspeech 2017. Farmington Hills: Cengage Learning, 2017: 2-6.

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

[23] Varga A, Steeneken H J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993, 12(3): 247-251.

[24] Wang L B, Yoshida Y, Kawakami Y, et al. Relative phase information for detecting human speech and spoofed speech[C]//Proceedings of the Interspeech 2015. Farmington Hills: Cengage Learning, 2015: 2092-2096.

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

[26] 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 the 2022 IEEE In-

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6382-6386.

[27] 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, 2023.
Wen Y. Research on synthetic and converted speech detection based on multi-branch convolutional neural network[D]. Nanchang: Jiangxi Normal University, 2023.

[28] Das R K. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021[C]//Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. 2021: 29-36.

[29] Liu X C, Wang X, Sahidullah M, et al. ASVspoof 2021: towards spoofed and deepfake speech detection in the wild[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2507-2522.

[30] Tak H, Patino J, Todisco M, et al. Evans and A. Larcher, "End-to-End anti-spoofing with RawNet2[C]//Proceedings of the 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。