基于改进STAM的语音端点检测算法

吴荣波[1],周斌[1*],胡波[2]

(1 中南民族大学 a.计算机科学学院;b.国家民委信息物理融合智能计算重点实验室,武汉 430074;2 武汉东信同邦信息技术有限公司,武汉 430074)

摘 要 在低信噪比的背景下,由于背景噪声干扰信号特征,存在语言端点检测误判和漏判的风险.现有的解决方法存在易受干扰、精度有限、鲁棒性差等问题.针对上述问题,对STAM进行优化,提出了一种改进的语音端点检测算法Inception-ResNet STAM(IR-STAM).该算法通过改用音频指纹(AFP)特征来取代传统的Log-Mel特征,实现了对音频信号更深层次的特征提取;对频率注意力模块的卷积方式进行改进,采用深度可分离卷积,有效降低了模型的参数量;加入Inception-ResNet模块,进一步增强了模型对不同尺度特征的捕捉和分析能力.实验结果表明:在TIMIT测试集上,IR-STAM相较于STAM,模型的参数量降低150 k,并且在不同信噪比环境下F1分数均提高了0.5以上.

关键词 低信噪比;Inception-ResNet模块;音频指纹特征;语音端点检测

中图分类号 O625.67;O643.3 **文献标志码** A **文章编号** 1672-4321(2025)03-0384-09

doi:10.20056/j.cnki.ZNMDZK.20250312

Voice activity detection algorithm based on improved STAM

WU Rongbo[1], ZHOU Bin[1*], HU Bo[2]

(1 South-Central Minzu University, a. College of Computer Science; b. Key Laboratory of Information-Physics-Fusion-based Intelligent Computing of the National Ethnic Affairs Commission of the People's Republic of China, Wuhan 430074, China; 2 Wuhan Dongxin Tongbang Information Technology Co., Ltd., Wuhan 430074, China)

Abstract In low Signal-to-Noise Ratio (SNR) scenarios, voice activity detection is impeded by background noise that disrupts signal characteristics, leading to the risks of false and missed detections. Existing solutions are prone to interference, have limited accuracy, and lack robustness. To tackle these challenges, an enhanced version of the voice activity detection Model (STAM) has been developed, named the Inception-ResNet STAM (IR-STAM). The algorithm facilitates more profound feature extraction from audio signals by substituting traditional Log-Mel features with Audio Fingerprint (AFP) features. The convolution method within the frequency attention module is enhanced through the use of depthwise separable convolution, significantly reducing the model's parameter count. Furthermore, the integration of an Inception-ResNet module bolsters the model's capacity to detect and analyze features across various scales. The experimental results show that on the TIMIT test set, IR-STAM has reduced the model's parameter count by 150 k compared to STAM and has achieved an increase of more than 0.5 in the F1 score across various Signal-to-Noise Ratio conditions.

Keywords low signal to noise ratio; Inception-ResNet; audio fingerprinting features; voice activity detection

随着互联网技术的不断发展,语音处理相关的技术也在不断的发展成熟,语音端点检测(Voice Activity Detection,VAD)是一种检测音频中是否存在语音的信号处理技术,它将帧序列划分为语音和非语音.在清晰无干扰的语音环境中,这项技术能够轻松实现.然而,在低信噪比环境中,特别是在非平稳、不匹配的噪声条件下,VAD变得具有挑战性.

早期的研究通过分析音频信号的各种属性,如能

收稿日期 2024-08-24

* **通信作者** 周斌(1971-),男,教授,博士,研究方向:大数据处理,E-mail:binzhou@mail.scuec.edu.cn

基金项目 中南民族大学中央高校基本科研业务费专项资金资助(CZY23006);湖北省技术创新专项基金资助项目(2019ADC071)

量、频率分布和时域特征[1-2]，来识别语音的存在.例如测量信号的零交叉率(Zero Crossing Rate，ZCR)或梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients，MFCCs)，这些都是捕捉语音信号特性的有效手段.

然而，这些传统方法在处理复杂环境或高噪声条件下的音频信号时，往往表现出一定的局限性.首先，它们通常需要精心设计的特征选择和阈值设置[3]，这在不同应用场景中可能需要大量的调整和优化；其次，这些方法[4]在面对多种噪声类型和非平稳的噪声时，鲁棒性较差，容易受到干扰，导致检测精度下降；最后，传统算法可能难以适应音频信号中的快速变化和动态特性[5]，限制了它们在实时系统中的应用.深度学习[6-7]在处理这些问题时表现出来较好的效果，其主要分为两种方法：端到端方法和特征工程方法.

端到端方法主要包括CLDNN[8]、AV-VAD[9]、AM-VAD[10]等，这类方法主要是将时域信号直接带入深度学习网络，减少了手动特征提取的需求，简化了从数据预处理到模型训练的整个流程.上述方法将所有步骤都集成在一个统一的框架中，不依赖于特定任务的特征，因此速度较快，但这些端到端方法在低信噪比环境下，检测精度依然受限.

特征工程方法依据使用的特征类型和组合方式可以被分为单特征方法和多种特征融合方法.单特征方法依赖于单一类型的特征来实现语音端点检测.这些特征通常是从原始音频信号中直接提取的，不需要复杂的预处理步骤，它主要包括bDNN[11]、ACAM[12]、SA[13]、STAM[14]等.这类方法由于只处理单一特征，导致模型无法充分捕捉语音信号中的复杂信息.在低信噪比环境下，单特征方法可能无法有效区分噪声和语音，导致检测精度下降，此外低信噪比环境会影响特征的稳定性和可靠性.

多种特征融合方法从语音信号中提取多种类型的特征来进行语音端点检测.例如，AM-cIRM[15]采用了一种基于注意力模型的新型深度神经网络架构.该方法将cIRM特征与Log-Mel特征结合，通过特征融合来实现噪声抑制和语音增强.尽管AM-cIRM在VAD任务中表现出色，但其引入的大量参数和复杂的模型结构，使得其在实际应用中需要消耗更多的计算资源.PC-ARN[16]将Log-Mel特征与相位相关特征[17]结合，虽然在一定程度上在提高模型精度的同时降低了模型的参数量，但其模型精度不足和泛化能力差的缺点依然存在.

针对上述问题，本文提出了基于STAM改进的语音端点检测算法.首先，引入了AFP[18]特征，使模型能够更准确地捕捉音频信号的细微特征；然后，将频率注意力模块内的传统卷积层优化为深度可分离卷积，这一改进在保持特征提取质量的同时减少了模型的参数量；最后，在主干网络中融合了Inception-ResNet模块[19]，此优化不仅增强了网络的多尺度特征学习能力，还通过残差连接提升了深层网络的泛化性.实验表明：在TIMIT测试集上，本文模型相较于STAM，在不同信噪比环境下F1分数均提高了0.5以上.

1　IR-STAM模型的工作原理

STAM模型包括四个模块：频率注意力、管道网络、时间注意力和后处理网络.

(1)频率注意力模块：使模型能够更加关注频谱中的关键频率信息，特别是那些包含有用语音成分的频率信息.这有助于提高模型的抗噪性.该模块由多个块组成，每个块包含一对门控卷积层.在每个块之后，沿频率轴应用一个额外的一维最大池化层.

(2)管道网络：包含两个隐藏维度为N_d的全连接网络(Fully Connected Network，FCN)，其输出表示为$G \in \mathbb{R}^{N_d \times L}$，其中$L$是上下文维度.

(3)时间注意力模块：STAM采用了多头自注意力机制，允许模型同时关注不同位置的信息.

(4)后处理网络：后处理网络包括两个全连接层，随后是一个Sigmoid激活函数来进行预测，最后得到每一帧是否为语音的概率值.

STAM算法通过融合频谱和时间注意力模块，展现了在低信噪比环境中良好的泛化性能.然而，它仍然面临着一些挑战，包括精度提升的局限性和模型参数数量的庞大.为了在提升精确度的同时优化参数效率，本文在STAM算法的基础上进行了深入改进，提出了IR-STAM算法.IR-STAM算法主要由特征提取、模型预测、分类决策三个部分构成.在特征提取阶段，首先采用音频指纹(Audio Fingerprinting Features，AFP)特征替换了传统的Log-Mel特征，从而丰富了输入特征所包含的信息量.接着，在模型预测阶段，使用IR-STAM模型.IR-STAM是在STAM模型的基础上，通过加入Inception-ResNet模块来增强模型获取多尺度特征的能力.同时，将原有的频率注意力模块中的传统卷积层替换为深度可分离卷积，这样做在不牺牲模型性能的前提下，有效减少了模型的参数数量.最终，在结果分类阶段，对模型的预测结果进行分类处理.

IR-STAM算法流程如图1所示.

1.1 特征提取

在本节中,简要介绍用于提取本工作中的声学特征的预处理步骤.输入的噪声$x[n]$被建模为:

$$x[n]=s[n]+w[n], \tag{1}$$

其中,$s[n]$表示干净的语音信号,$w[n]$表示加性背景噪声,$n\in Z$是离散时间索引,处理通过在频率域中对$x[n]$应用短时傅里叶变换(STFT)来实现,其计算公式如下:

$$X(t,f)=\sum_{n=0}^{N-1}x\left[n+tL_{\text{hop}}\right]h[n]e^{-j2\pi f\frac{n}{N}}, \tag{2}$$

其中t为帧索引,L_{hop}为帧移,即短时傅里叶变换窗口在时间上的间隔,$f\in\{0,1,2,\cdots,N/2\}$为频率索引,N表示窗函数大小,即STFT的窗口长度,$h[n]$是窗函数.

计算STFT输出$X(t,f)$的功率,即其模的平方$|X(t,f)|^2$.这个步骤将复数的STFT输出转换为能量或功率的度量,反映了信号在各个频率成分上的强度.然后,为了使频率分辨率适应人耳的特性,通过一系列按照Mel尺度设计的滤波器来对功率谱$|X(t,f)|^2$进行处理.将对数函数应用于每个滤波器的输出,对于每一时刻t和每个滤波器b,滤波后的功率谱值可表示为:

$$\text{FB}(t,b)=20\log_{10}\{\sum_{f=l_b}^{h_b}u_b(f)\left|X(t,f)\right|^2\}. \tag{3}$$

设定$b\in\{0,1,\cdots,B-1\}$为滤波器索引,B是滤波器组中的滤波器数量,$u_b(f)$是第b个子带的频谱整形滤波器,l_b和h_b分别是$u_b(f)$的下限和上限频率.当前第t帧的对数Mel滤波器组特征向量表示为:

$$\text{FB}_t=[\text{FB}(t,0),\cdots,\text{FB}(t,b),\cdots,\text{FB}(t,B-1)]. \tag{4}$$

离散余弦变换(DCT)被应用于对数Mel滤波器组特征,以获得Mel频率倒谱系数(MFCC):

$$\text{MFCC}(t,b)=\frac{1}{20}\sqrt{\frac{2}{B}}\sum_{b=0}^{B-1}\text{FB}(t,b)\cos\left(\frac{p\pi}{B}(b-0.5)\right), \tag{5}$$

第t帧的MFCC特征向量为:

$$\text{MFCC}_t=[\text{MFCC}(t,0),\cdots,\text{MFCC}(t,B-1)]. \tag{6}$$

频谱子带质心(SSC)常用于测量子带频谱的中心频率.为了计算第b个整形滤波器的SSC时,使用加权平均值,其计算公式如下:

$$\text{SSC}(t,b)=\frac{\sum_{f=l_b}^{h_b}fu_b'(f)\left|X(t,f)\right|^2}{\sum_{f=l_b}^{h_b}u_b'(f)\left|X(t,f)\right|^2}, \tag{7}$$

其中$u_b'(f)$是子带滤波器.为了简单起见,本工作中计算MFCC和SSC特征时使用相同的滤波器$u_b'(f)$.为了高效训练,使用归一化SSC(NSSC),其取值范围为[-1, 1],计算公式为:

$$\text{NSSC}(t,b)=\frac{\text{SSC}(t,b)-(h_b-l_b)}{h_b-l_b}. \tag{8}$$

类似地,第t帧的NSSC特征向量定义为:

$$\text{NSSC}_t=[\text{NSSC}(t,0),\cdots,\text{NSSC}(t,B-1)]. \tag{9}$$

音频指纹组合(AFPC)是MFCC和NSSC的组合,当作为生成对抗网络(GAN)的输入用于语音增强时[12]表现出优越的性能.在本文的研究中,使用类似的特征组合:

$$\begin{aligned}\text{AFPC}_t=[&\text{MFCC}_t,\Delta\text{MFCC}_t,\Delta^2\text{MFCC},\text{NSSC}_t,\\&\Delta\text{NSSC}_t,\Delta^2\text{NSSC}_t],\end{aligned} \tag{10}$$

其中Δ和Δ^2分别表示一阶差分和二阶差分操作.

1.2 IR-STAM模型预测

鉴于STAM模型[14]在精度上的局限性以及其

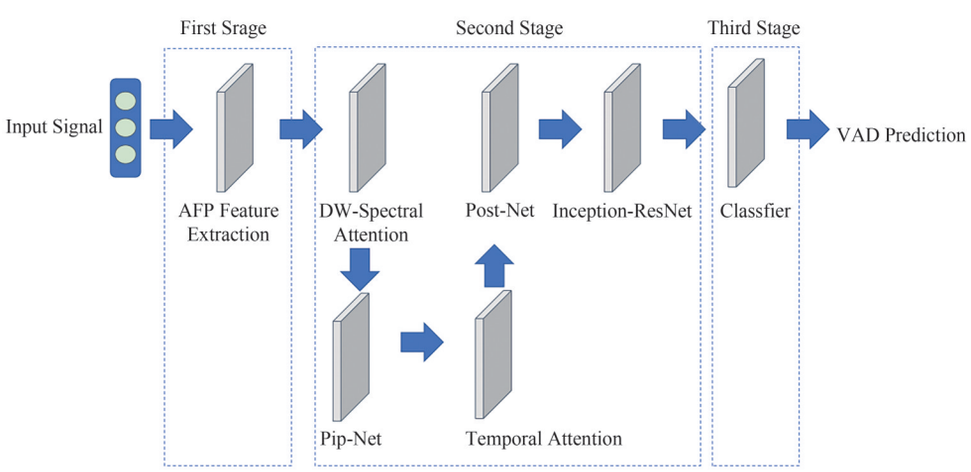


图1 IR-STAM算法流程图

Fig. 1 IR-STAM algorithm flowchart

较大的参数量，本文采用基于STAM改进的模型IR-STAM来进行模型预测，其具体改进措施如下：将STAM的频率注意力模块替换为DW-Spectral Attention模块，在后处理网络之后加入Inception-ResNet模块.

1.2.1　DW-Spectral Attention模块

DW-Spectral Attention模块的设计思想是在STAM原有频率注意力模块的基础之上，用深度可分离卷积层替代传统的卷积层.频率注意力模块是语音端点检测的核心部分，因此很大程度上决定了网络的大小.为了能够显著减少模型的参数量和计算量，本模型在频率注意力模块中将传统的卷积层替换为深度可分离卷积.

深度可分离卷积是一种高效的卷积方法，它将标准卷积过程分解为深度卷积(Depthwise Convolution)和逐点卷积(Pointwise Convolution)两个阶段.对于一个标准的卷积过程，如果输入特征的维度为$N \times H \times W \times C$(其中$N$是批次大小，$H$和$W$分别是输入特征图的高度和宽度，$C$是通道数)，并且有$K$个$3 \times 3$的卷积核，设置pad = 1，stride = 1，那么标准卷积输出为$N \times H \times W \times K$.对于深度可分离卷积，在深度卷积阶段将输入的$N \times H \times W \times C$分成$C$组，然后对每组数据应用$3 \times 3$的卷积核，这样可以提取每个通道的空间特征；在逐点卷积阶段对输入的$N \times H \times W \times C$做$K$个$1 \times 1$卷积，提取特征图中每个点的特征.深度可分离卷积与传统卷积的参数量之比和计算量之比分别为：

$$\frac{3^2 \times C + K \times C}{3^2 \times C \times K} = \frac{1}{9} + \frac{1}{K}, \tag{11}$$

$$\frac{3^2 \times C \times 1^2 + K \times C \times 1^2}{3^2 \times C \times K \times 1^2} = \frac{1}{9} + \frac{1}{K}. \tag{12}$$

这种方法显著减少了模型的参数数量和计算复杂度，同时保持了特征提取的能力.为了防止过拟合，在深度卷积和逐点卷积之间加入批归一化层，替换后的深度可分离卷积结构如图2所示.

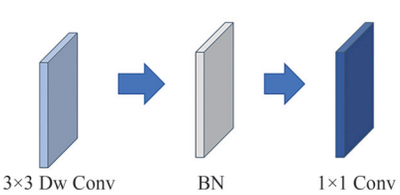


图2　深度可分离卷积层示意图

Fig. 2　Depthwise separable convolution illustration

通过这种改进，VAD系统能够在保持高精度的同时，降低模型的计算压力.这不仅加快了处理速度，还减少了模型对硬件资源的需求，使得VAD系统更适合部署在边缘设备或资源受限的环境中.此外，深度可分离卷积的引入，增强了模型对不同频率特征的适应性，提升了模型在复杂声学条件下的鲁棒性.

1.2.2　Inception-ResNet模块

在主干网络中融入Inception网络结构是一种提升模型特征提取能力的有效策略.Inception模块通过并行卷积，能够在不同尺度上捕获图像特征，这使得网络能够同时关注局部细节和全局轮廓.本文加入的Inception网络模块如图3所示.首先对后处理网络模块产生的输入特征$G \in \mathbb{R}^{N_d \times L}$进行重塑得到$G \in \mathbb{R}^{4 \times \frac{N_d}{4} \times L}$，以便输入特征与Inception网络匹配，然后再将重塑后的语音特征送入Inception网络中.Inception网络各层参数如表1.

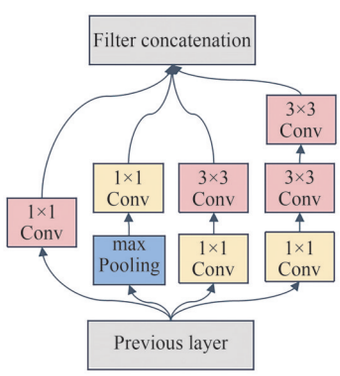


图3　Inception网络结构

Fig. 3　Inception network architecture

表1　Inception各层参数

Tab. 1　The parameters of each layer in Inception

分支	网络层	卷积核(窗口大小)	输入维度	输出维度
branch1	1×1 Conv	1	4×7×32	1×7×32
branch2	max pooling	3	4×7×32	4×7×32
	1×1 Conv	1	4×7×32	1×7×32
branch3	1×1 Conv	1	4×7×32	1×7×32
	1×1 Conv	3	1×7×32	1×7×32
branch4	1×1 Conv	1	4×7×32	1×7×32
	1×1 Conv	5	1×7×32	1×7×32

将Inception网络集成到主干网络中，可以充分利用其多尺度特征融合的优势，增强模型对输入数据的表征能力.例如，在图像分类任务中，Inception模块可以帮助模型更好地识别不同大小的物体；在语音端点检测中，则能更准确地识别语音信号.此外，Inception网络的引入还有助于提高模型对噪声和遮挡等不利因素的鲁棒性，使得模型在复杂环境中也能保持较高的性能.为了提高模型训练的稳定性，将Inception网络与残差网络(ResNet)结合使用，其结构如图4所示.通过这种方式可以提高深层网

络的训练稳定性和表征能力,增强其特征融合和泛化能力,简化模型的网络设计,并加速模型的训练过程.

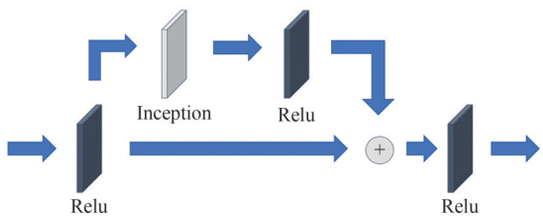


图4 Inception-ResNet网络结构

Fig. 4 Inception-ResNet network architecture

1.3 损失函数

将时间注意力模块产生的输出Y_{att}与Inception-ResNet模块产生的输出Y_{inc}相加以强化特征差异:

$$Y_{\text{branch}} = Y_{\text{att}} + Y_{\text{inc}} \in \mathbb{R}^{N_d \times L}, \tag{13}$$

利用STAM的Post-Net进行VAD预测.与STAM[14]相似,总损失可以表述为

$$L = L_{\text{post}} + L_{\text{pipe}} + \lambda L_{\text{att}}, \tag{14}$$

其中L_{post}和L_{pipe},L_{att}都是交叉熵损失.

1.4 分类决策

对于模型产生的输出,其计算方式如下:

$$\hat{y}_t = \frac{1}{L}\sum_{l \in T} y_{t+l} \in \mathbb{R}^L, \tag{15}$$

其中,y_{t+l}代表第$t+l$帧的软预测结果,T表示临近帧集合,而L为集合T中元素的数量,为了确保模型的可比较性和复现性,集合T与文献[7]中保持一致.最终的决策标签$\bar{y}_t$通过将预测标签$\hat{y}_t$与一个正阈值θ_{VAD}进行比较来确定:

$$\bar{y}_t = \begin{cases} 1, & \text{if } \hat{y}_t \geqslant \theta_{\text{VAD}} \\ 0, & \text{otherwise} \end{cases}, \tag{16}$$

1表示语音存在,而0表示语音不存在.通过这种方法,VAD模块能够为每一帧提供准确的语音存在性预测.

2 实验结果与分析

2.1 实验数据集

使用TIMIT[20]语料库进行训练,其具体描述如表2所示.在实验中,训练数据集中95%的语音话语用于训练,5%用于模型验证.为了防止TIMIT数据集中静音片段占比过小,在每个句子的开头和结尾各添加了1秒的静音.训练和验证集通过添加NOISEX-92语料库[21]中的八种类型的噪声进行增强(包括背景交谈声、F16飞机声、驱逐舰声、M109坦克声、沃尔沃汽车声、白噪声以及两种工厂噪声),信噪比设置为-10、-5、0、5和10 dB.在测试阶段,使用TIMIT测试数据集,添加的所有八种未见过的噪声类型来自AURORA噪声数据集[22],用于干扰干净的语音信号,信噪比设置为-10、-5、0和5 dB.

表2 TIMIT语料库描述

Tab. 2 Description of TIMIT corpus

属性	描述
语种	英语
说话者	630名,包括来自8个主要地区的男性和女性,年龄在20到50岁之间.
录音时长	每个说话者大约15分钟.
录音环境	录音在安静的环境中进行,使用高质量的麦克风.
数据集大小	约4 GB

2.2 实验评价指标

选择F1分数和检测成本函数(Detection Cost Function,DCF)作为评价指标.F1分数是二值分类问题的常用评价指标,定义为:

$$\text{F1} = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})}, \tag{17}$$

其中TP为正确预测为正类(语音)的样本数,FP为错误预测为正类的样本数.FN为误预测为负类的样本数.

DCF的目的是通过结合假阳性率和假阴性率来综合衡量模型的错误表现,其计算公式如下:

$$\text{DCF} = (1-\beta)P_{\text{FN}} + \beta P_{\text{FP}}, \tag{18}$$

其中P_{FN}是FN占总样本数的比率,P_{FP}是FP占总样本数的比率.根据文献[16],β被设置为0.25,以便更严重地惩罚缺失的语音帧.

2.3 实验环境

本实验基于Linux搭配Pytorch深度学习框架实现,具体环境如表3所示.

表3 实验环境配置单

Tab. 3 Experimental environment configuration sheet

硬件设备	参数
CPU	12th Gen Intel(R)Core(TM)
GPU	NVIDIA GeForce RTX 4070TI
操作系统	Ubuntu 20.04
加速环境	CUDA 11.4.56
深度学习框架	Pytorch

2.4 训练设置

对于每个来自训练和测试数据集的话语,采样率为16 kHz,随后,通过应用32 ms的汉宁窗,并以16 ms的步长进行窗位移,以实现对语音信号的精确帧分割.用于频谱分析的STFT大小设置为$N = 512$,

本文对MFCC，ΔMFCC，Δ²MFCC和NSSC，ΔNSSC各计算16个系数，得到80个特征．阈值θ_{VAD}设置为0.5，训练批次大小为4096，其它设置与STAM的设置相同．

损失函数曲线如图5所示，从图中可以看到损失函数随着epoch次数的增加总体呈下降趋势，最终在在第10个epoch处，训练集和验证集的损失趋于稳定．

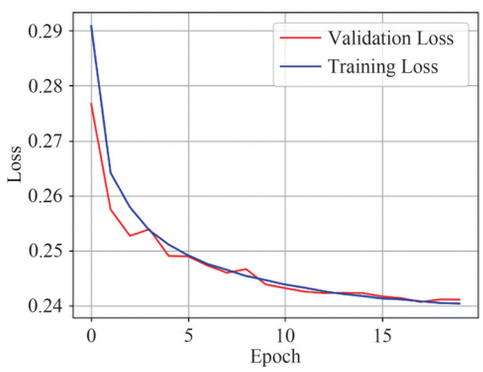


图5　损失函数曲线图

Fig. 5　Loss function graph

2.5　消融实验

进行消融实验以评估不同模块在模型中的作用和效果，在TIMIT测试集上的平均F1分数（以百分比表示）、平均DCF（以百分比表示）和参数量如表4-6所示．其中Afpc指将Log-Mel特征替换为AFP特征，DW指将频率注意力模块中的卷积层替换为深度可分离卷积，Inc指在主干网络中加入Inception-ResNet模块．

表4　F1分数对比

Tab. 4　Comparison of F1-Score

Method	STAM	Afpc+STAM	Afpc+DW+STAM	Afpc+DW+Inc+STAM
F1	97.6	98.4	98.3	98.5

表5　DCF对比

Tab. 5　Comparison of DCF

Method	STAM	Afpc+STAM	Afpc+DW+STAM	Afpc+DW+Inc+STAM
DCF	1.2	0.67	0.69	0.64

表6　参数对比

Tab. 6　Comparison of parmeters

Method	STAM	Afpc+STAM	Afpc+DW+STAM	Afpc+DW+Inc+STAM
parmeters/K	559	559	409	414

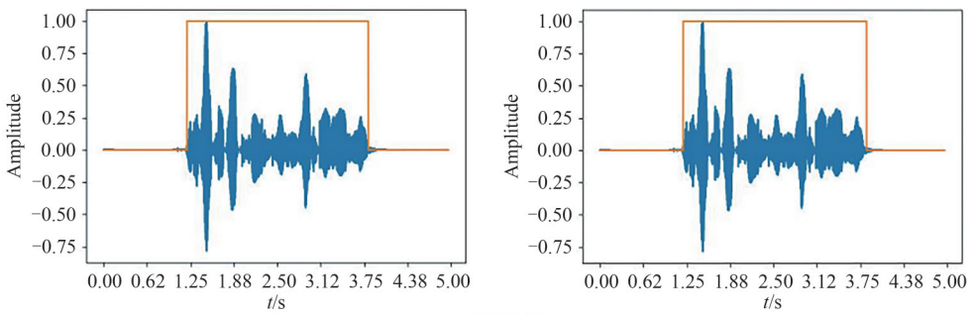
从表4中可以看出加入的Inception-ResNet模块与AFP特征对语音端点检测的准确率有不同程度的提升．将Log-Mel特征替换为AFP特征时，观察到F1分数增加了0.8%；相对于Log-Mel特征，AFP特征提供了更丰富的音频信息，使模型能够更准确地捕捉到音频中的关键特征，从而提高了算法的精度；另一方面，AFP特征通常对音频的噪声和变化具有较高的鲁棒性，即使在音频质量不佳或低信噪比的情况下也能够保持较高的识别准确率．在主干网络中加入Inception-ResNet模块，模型的F1分数增加了0.2%，这是由于Inception-ResNet模块能够学习不同尺度的特征，有助于捕捉图像中不同大小的对象，提高模型对不同特征的表征能力．此外，残差连接允许在网络中前几层学习到的特征直接或间接地被后面的层所利用，这有助于特征的重用和信息的整合．

从表5中可以看出，AFP特征相比于Log-Mel特征在降低检测成本函数DCF方面表现出色．这一结果说明AFP特征能够有效提高语音活动检测的准确性，减少了漏检和误检的发生．这是由于AFP特征通过更好地模拟人耳的听觉感知机制，捕捉语音信号中的关键信息，从而在复杂的声音环境中实现了更为精确的语音识别．

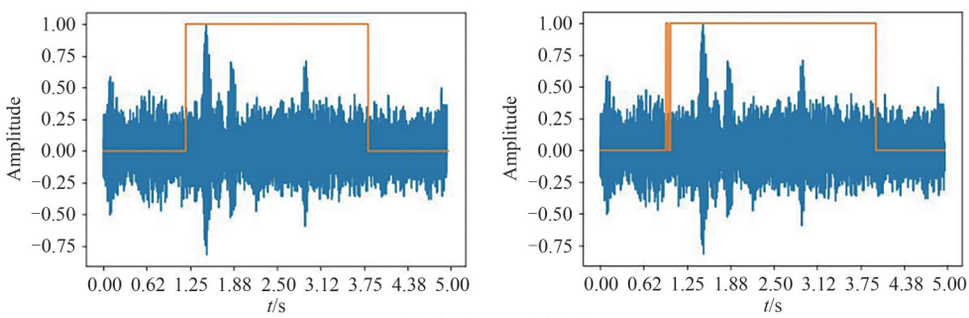
从表6可以看出将频率注意力模块中的卷积层替换为深度可分离卷积，在保持F1分数和DCF性能基本不变的前提下，成功减少了150K的参数量．这种改进归功于深度可分离卷积仅涉及输入输出通道间的线性变换，而非全空间卷积，从而在不损失特征提取能力的基础上，显著降低了模型的参数量和计算成本，提升了模型的效率和实用性，使其更适合于资源受限的环境．

为了进一步分析模型相对于STAM在测试数据集上的性能，本文对STAM与本文提出的模型的语音端点检测效果进行对比，其结果如图6所示．

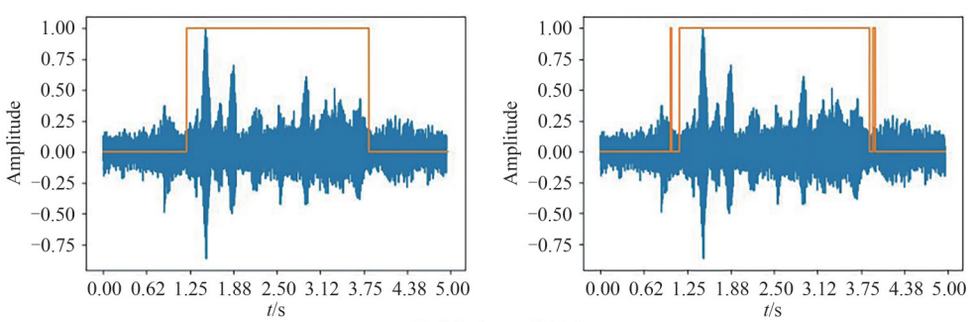
图6清晰地揭示了IR-STAM算法在低信噪比环境中的卓越性能．与STAM算法相比，在一系列信噪比条件下，IR-STAM均展现出更高的识别准确性．特别地，在信噪比极端降低至−5 dB的挑战性环境下，IR-STAM依然能够稳定地识别出语音信号，显示出非凡的噪声抵抗能力．这一显著成果凸显了特征提取模块和Inception-ResNet模块的有效性．一方面AFP特征为算法提供了更丰富的语音信号信息；另一方面，Inception-ResNet模块的引入通过多尺度信息处理和残差连接的设计，提高了特征提取能力．



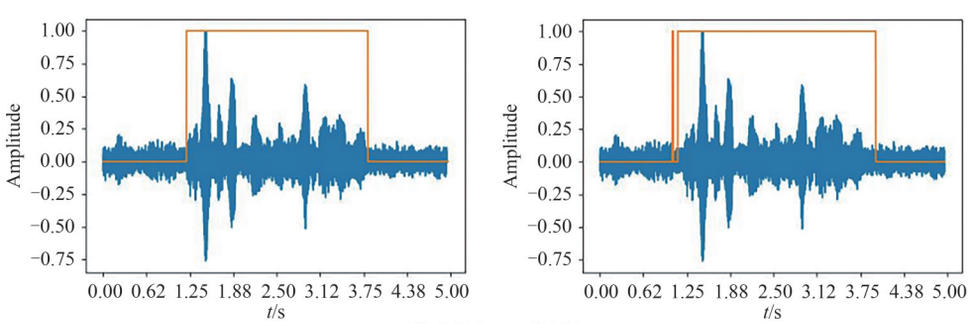
(a) 干净语音



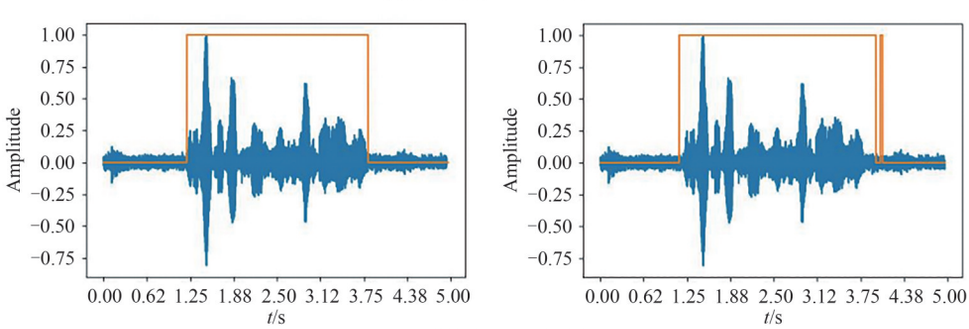
(b) 信噪比为−5 dB的语音



(c) 信噪比为0 dB的语音



(d) 信噪比为5 dB的语音



(e) 信噪比为10 dB的语音

图6 STAM(右)与本模型(左)语音端点检测效果对比

Fig. 6 Comparison of voice activity detection effects between STAM (right) and the presented model (left)

2.6　与主流模型的性能对比

为了展示所提模型的有效性，同时对比了rVAD[2]、DCU-10[23]、ACAM[12]等经典语音端点检测模型．在测试数据集上的结果如表7所示．

表7　在TIMIT测试集上的F1分数和DCF对比

Tab. 7　Comparison of F1-Score and DCF on TIMIT test set SNR

SNR	Metric	rVAD	DCU-10	ACAM	STAM	AM-cIRM	IR-STAM
−5 dB	F1	79.5	86.4	85.9	97.7	98.0	98.6
	DCF	8.3	7.8	6.2	1.5	1.0	0.6
0 dB	F1	86.0	89.8	90.7	98.0	98.5	99.0
	DCF	5.8	5.7	3.7	1.3	0.7	0.4
5 dB	F1	92.4	92.3	95.4	98.3	98.9	99.0
	DCF	3.9	4.0	2.6	1.2	0.5	0.5
10 dB	F1	94.0	94.2	96.0	98.4	98.9	99.1
	DCF	3.4	2.8	2.3	1.1	0.5	0.4

表7给出了F1分数和DCF的比较结果（均以百分比表示）在不同信噪比上的平均值．显然，所有基于注意力的方法（ACAM、STAM 、AM-cIRM 和 IR-STAM）都比非基于注意力的方法（rVAD 和DCU-10）取得了更好的结果．与 ACAM 相比，STAM 通过利用频率和时间注意力大大提高了性能．AM-cIRM 通过cIRM 特征同时利用了幅值和相位信息，与STAM 相比，该模型的性能得到了进一步提高．所提出的IR-STAM 模型利用 AFP 特征和 Inception-ResNet 模块，相对于 AM-cIRM 在不同信噪比背景下 F1 分数均有所提升和 DCF 均有所下降，其中在−5 dB 噪声背景下 F1 分数和 DCF 分别提升了 0.6 和降低了 0.4．这说明 AFP 特征和 Inception-ResNet 模块在提高检测精度和降低误检率方面的有效性．

表8显示了不同模型的大小和处理10秒话语的平均运行时间，结果验证了该算法的有效性，相较于AM-cIRM，参数量大幅减少，执行时间缩短29 ms．这说明深度可分离卷积对于模型轻量化是一个行之有效的方案．

表8　参数和平均运行时间对比

Tab. 8　Comparison of parameters and average running time

模型	rVAD	DCU-10	ACAM	STAM	AM-cIRM	IR-STAM
参数	NA	2808 K	957 K	559 K	3613 K	409 K
运行时间(ms)	86	251	1263	132	269	240

3　结语

本文针对STAM存在参数量大，精度低等问题，提出了IR-STAM语音端点检测算法．该算法首先对原始信号提取AFP特征，然后对提取到的特征进行模型预测，最后对预测的结果进行分类．实验结果表明：在TIMIT测试集上，本文提出的改进模型相较于STAM在低信噪比语音的情况下具有更高的F1分数和更低的DCF，其中在−5 dB噪声背景下提升最为显著，F1分数和DCF分别提升了0.9和降低了0.9，并且参数量降低了150 K．

参　考　文　献

[1] SOHN J, KIM N S, SUNG W. A statistical model-based voice activity detection[J]. IEEE Signal Processing Letters, 1999, 6(1): 1-3.

[2] TAN Z H, SARKAR A K, DEHAK N. rVAD: An unsupervised segment-based robust voice activity detection method[J]. Computer Speech & Language, 2020, 59: 1-21.

[3] 肖思，龚杰，李宝清．低信噪比环境下的多通道语音端点检测算法[J]．中国科学院大学学报，2023，40(5)：687-693.

[4] 张洪德，韩鑫怡，柳林，等．基于谱减与自适应子带对数能熵积的端点检测[J]．兵器装备工程学报，2022，43(2)：267-273.

[5] 刘艳辉．改进型多特征语音端点检测方法[J]．河南工程学院学报(自然科学版)，2022，34(4)：69-73，78.

[6] SUN T, LEI T, ZHANG X, et al. A lightweight hybrid multi-channel speech extraction system with directional voice activity detection [C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul:IEEE, 2024: 1486-1490.

[7] YANG Q, LIU Q, LI N, et al. SVAD: A robust, low-power, and light-weight voice activity detection with spiking neural networks [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul:IEEE, 2024: 221-225.

[8] ZAZO R, SAINATH T N, SIMKO G, et al. Feature learning with raw-waveform CLDNNs for voice activity detection [C]//Interspeech 2016. San Francisco: ISCA, 2016: 3668-3672.

[9] ARIAV I, COHEN I. An end-to-end multimodal voice activity detection using WaveNet encoder and residual networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(2): 265-274.

[10] LARSEN C M, KOCH P, TAN Z H. Adversarial multi-task deep learning for noise-robust voice activity detection with low algorithmic delay[EB/OL]. 2022: 2207.01691. https://arxiv.org/abs/2207.01691v1

[11] ZHANG X L, WANG D. Boosting contextual information for deep neural network based voice activity detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(2): 252-264.

[12] KIM J, HAHN M. Voice activity detection using an adaptive context attention model [J]. IEEE Signal Processing Letters, 2018, 25(8): 1181-1185.

[13] JO Y R, MOON Y K, CHO W I, et al. Self-attentive VAD: Context-aware detection of voice from noise[C]// 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto:IEEE, 2021: 6808-6812.

[14] LEE Y, MIN J, HAN D K, et al. Spectro-temporal attention-based voice activity detection[J]. IEEE Signal Processing Letters, 2019, 27: 131-135.

[15] ZHAO Y, ATTABI Y, CHAMPAGNE B, et al. Complex IRM-aware training for voice activity detection using attention model[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore:IEEE, 2022: 3698-3702.

[16] TANG M, HUANG H, ZHANG W, et al. Phase continuity-aware self-attentive recurrent network with adaptive feature selection for robust VAD[C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul:IEEE, 2024: 11506-11510.

[17] KIM D, HAN H, SHIN H K, et al. Phase continuity: Learning derivatives of phase spectrum for speech enhancement[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore:IEEE, 2022: 6942-6946.

[18] FARAJI F, ATTABI Y, CHAMPAGNE B, et al. On the use of audio fingerprinting features for speech enhancement with generative adversarial network [C]//2020 IEEE Workshop on Signal Processing Systems (SiPS), Coimbra: IEEE, 2020: 1-6.

[19] 张瑞博, 李凌均. 基于注意力机制与Inception-ResNet的轴承故障诊断方法[J]. 电子测量技术, 2023, 46(21): 107-113.

[20] GAROFOLO J S, LAMEL L, FISHER W, , et al. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1 [J]. NASA STI/Recon technical report n, 1993, 93: 27403.

[21] 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.

[22] PEARCE D, HIRSCH H G. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions [C]//6th International Conference on Spoken Language Processing (ICSLP 2000), Paris: ISCA, 2000: 181-188.

[23] LIU F, WANG L. UNet-based model for crack detection integrating visual explanations [J]. Construction and Building Materials, 2022, 322: 126265.

(责编&校对 雷建云)