计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：基于区域自适应多尺度卷积的单声道语音增强算法
作者：王钇翔，吕忆蓝，台文鑫，孙建强，蓝天
DOI：10.19734/j.issn.1001-3695.2021.03.0131
收稿日期：2021-03-05
网络首发日期：2021-08-30
引用格式：王钇翔，吕忆蓝，台文鑫，孙建强，蓝天．基于区域自适应多尺度卷积的单声道语音增强算法[J/OL]．计算机应用研究．
https://doi.org/10.19734/j.issn.1001-3695.2021.03.0131

网络首发时间：2021-08-30 17:33:10
网络首发地址：https://kns.cnki.net/kcms/detail/51.1196.TP.20210830.1244.026.html

基于区域自适应多尺度卷积的单声道语音增强算法 *

王钇翔，吕忆蓝，台文鑫，孙建强，蓝　天

(电子科技大学 信息与软件工程学院，成都 610054)

摘　要：近年来，基于卷积神经网络的方法在语音增强领域得到了广泛的应用。卷积神经网络的感受野大小与卷积核的尺寸相关，传统的卷积采用了固定大小的卷积核，这限制了网络模型的特征感知能力。此外，卷积神经网络使用参数共享机制，对空间区域中所有的样本点采用了相同的特征提取方式，然而，带噪频谱图噪声信号与干净语音信号的分布存在差异，特别是在复杂噪声环境下，这使得传统卷积方式难以实现高质量的语音信号特征提取和过滤。为了解决上述问题，提出了多尺度区域自适应卷积模块，利用多尺度信息提升模型的特征感知能力；同时，根据对应采样点的特征值自适应的分配区域卷积权重，实现区域自适应卷积，提升模型过滤噪声的能力。在 TIMIT 公开数据集上的实验表明，提出的算法在语音质量和可懂度的评价指标上取得了更优的实验结果。

关键词：语音增强；卷积神经网络；多尺度卷积；区域自适应

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2021.03.0131

Region-aware multi-scale based monaural speech enhancement algorithm

Wang Yixiang, Lyu Yilan, Tai Wenxin, Sun Jianqiang, Lan Tian

(*School of Information & Software Engineering*, *University of Electronic Science & Technology of China*, *Chengdu* 610054, *China*)

Abstract: In recent years, methods based on convolutional neural networks have been widely used in the field of speech enhancement. The size of the receptive field of the convolutional neural network is related to the size of the convolution kernel. And the traditional convolution uses a fixed-size convolution kernel, which limits the feature perception ability of the network model. In addition, due to the parameter sharing mechanism of the convolutional neural network, the same feature extraction method is used for all pixels in the spatial region. However, there are differences in the distribution of noise signals and clean speech signals in the noisy spectrogram. Especially in the complex noise environment, the general convolution method is difficult to achieve high-quality speech signal feature extraction and choosing. In order to solve the problems above, a multi-scale region adaptive convolution module is proposed, which uses multi-scale information to improve the feature perception ability of the model and the area adaptive convolution automatically allocates the area convolution to improve the denoising ability of the model. The experiments on the TIMIT public datasets show that the proposed algorithm has achieved satisfactory results in the metrics of speech quality and intelligibility.

Key words: speech enhancement; convolutional network; multi-scale convolution; region-aware

0　引言

现实声学环境中，由于存在大量未知且复杂的噪声，语音信号极易被干扰，这导致了语音信号的质量及可懂度的严重下降。在自动语音识别[1]、说话人识别[2]、助听设备[3]等下游任务中，语音信号被干扰将对性能造成较大影响，所以往往将语音增强作为这些任务的上游任务。语音增强任务是将受噪声干扰的语音信号作为输入，还原干净语音信号，实现语音信号的质量和可懂度的提升，因此常被用于语音处理任务的预处理阶段。本文主要研究单声道语音增强算法，探究如何在单源噪声干扰环境下进一步提升语音的质量和可懂度。

近年来，随着深度学习算法在各个领域的快速发展，基于深度学习的语音增强算法在该领域得到了更广泛的关注。由于深度神经网络(deep neural network)具有出色的建模复杂非线性函数的能力，即使在高度不稳定的噪声环境，基于深度神经网络的语音增强方法仍能实现稳定的性能[4]。早期基于深度神经网络的方法一般利用了大量的全连接层。这种方法可以实现较好的去噪效果[5,6]，但同时这也为模型带来了巨大的参数量，极大增大了计算机的负载。

与基于深度神经网络的方法相比，基于卷积神经网络(convolutional neural network，CNN)由于其共享卷积核参数的机制大大减少了网络的参数量，且同样具有较强的特征抽象能力，因此开始受到更多的关注[7~9]。Park 等人[10]首次将卷积神经网络引入语音增强领域，以编解码器体系作为模型架构，提出冗余卷积编解码网络体系结构，解决了卷积神经网络在特征提取过程中的信息丢失问题。Fu 等人[11]使用基于卷积神经网络的模型来直接获取干净语音的复数频谱，缓解了由于使用带噪语音的相位信息而造成的误差。Tan 等人[12]在卷积神经网络结构的基础上，引入了循环神经网络(Recurrent Neural Network，RNN)以建模时序信息。

然而，一般卷积神经网络具有的特征表达能力极大依赖于网络的深度以及卷积核的大小[13]，因此网络提炼声学特征

收稿日期：2021-03-05；**修回日期：**2021-06-29　**基金项目：**国家自然科学基金项目(U19B2028，61772117)、科技委创新特区项目(19-163-21-TS-001-042-01)、提升政府治理能力大数据应用技术国家工程实验室重点项目(10-2018039)、中央高校基本科研业务费项目(ZYGX2019J077)

作者简介：王钇翔(1996-)，男，浙江义乌人，硕士研究生，主要研究方向为语音增强，情感分析，推荐系统等(yxwang@std.uestc.edu.cn)；吕忆蓝(1997-)，女，河南焦作人，硕士研究生，主要研究方向为语音增强；台文鑫，男，甘肃兰州人，硕士研究生，主要研究方向为语音增强，推荐系统等；孙建强，男，湖北十堰人，硕士研究生，主要研究方向为语音增强，推荐系统等；蓝天，男，四川宜宾人，副教授，博士，主要研究方向为语音增强，语音识别，医学图像处理等.

的能力被计算机负载能力所限制。在视觉感官神经内，同一神经元可以在同一时间表现不同大小的局部感受野[14]，这表明来自多尺度的特征信息可以实现更好的进行特征表达。Szegedy 等人[15]提出了 InceptionNet，通过并联多个不同卷积核大小的卷积网络，获得对输入特征的多尺度表达，这极大提升了卷积神经网络的特征表达能力。Xian 等人[16]使用不同尺寸卷积核来捕获信号内局部以及上下文信息之间的依赖，其网络结构采用了特征上下采样的特征提取方式，这造成了一定的信息损失。受相关工作的启发，本文将多尺度卷积结构与冗余编解码器网络体系结构结合，以提升模型整体的信息感知能力。但是这仍然没有解决卷积神经网络所存在的由共享参数机制引起的问题。

卷积神经网络由于其卷积核具有共享参数的机制，经过卷积的特征图中的每一个元素的信息都被平等地考虑了，使得模型难以区分重要信息和冗余信息，这限制了卷积神经网络获取有效信息的能力。文献[17,18]提出了一种替代一般卷积的方法，为特征图中的每一个元素分配一个独立的卷积核，以使模型获得较一般卷积更为有效的空间特征提取能力。然而这样的方法存在这两个问题：第一，虽然这样的卷积方式不会增加计算负载，但是带来了巨大的参数量。第二，由于语音信号的帧长不确定，导致这样的方式很难应用于语音处理相关领域。同时，这样的卷积方式与一般的卷积一样，在训练过程中确定了卷积核参数，导致模型对各个样本的特定特征不够敏感，这一定程度上影响了模型处理复杂噪声环境的能力。Chen 等人[19]提出利用输入自适应的将特征图划分为不同区域，并对各个区域分别应用特定卷积核，得到区域特定感知结果。

为了进一步提高模型处理声学特征的能力，本文提出了一种基于区域自适应多尺度卷积的语音增强框架。以编码器解码器网络结构作为模型的基础框架，在编解码阶段首先使用多尺度卷积代替一般的卷积层以多个视角的形式提取更充分的声学特征。在多尺度卷积网络的基础上，本文利用了区域卷积，设计了一种区域自适应卷积块，根据输入动态的调整针对不同区域的卷积，进一步提升了模型对于特定区域的去噪能力。[20]提出了在模型输出增强语音的阶段，模型需要具有足够大小的感受野以包含相关重要信息，因此本文引用了多个线性门控单元(Gated Linear Unit，GLU)，并利用空洞卷积代替一般卷积来增大模型的感受野进一步提升了增强语音的质量和可懂度。

1　研究问题描述

一般来说，一条带噪语音通常可以表示为如下：

$$y(t) = x(t) + n(t) \tag{1}$$

其中，t表示时间帧序列，y、x以及n分别表示对应时间帧的带噪语音信号、干净语音信号以及噪声信号。因此，语音增强任务可以理解为是从带噪语音信号y中消除噪声n，最终得到干净语音x的过程。

基于神经网络的语音增强方法可以被大致分为两类，一类是基于时频掩码的方法，另一类是基于特征映射的方法。基于时频掩码的方法是学习带噪语音信号y中干净语音信号x以及噪声信号n之间的关系；基于特征映射的方法是学习带噪语音信号y与干净语音信号x之间的直接映射，即：

$$y(t) \rightarrow x(t) \tag{2}$$

本文主要研究基于特征映射的方法，首先通过短时傅里叶变换(short-time Fourier transform, STFT)将时域的语音信号转换为时频域，获得对应的语音频谱图。因此可将式(1)写为

$$\mathbf{Y}_{t,f} = \mathbf{X}_{t,f} + \mathbf{N}_{t,f}, \tag{3}$$

其中$Y_{t,f}$，$X_{t,f}$和$N_{t,f}$分别代表带噪语音，干净语音和噪声频谱图中时间帧序列t和频点f上的值。通过神经网络模型，获取带噪语音频谱图到增强语音频谱图的映射，最终利用带噪语音相位进行快速傅里叶逆变换(Inverse short-time Fourier transform, iSTFT)将增强语音频谱图还原回时域空间，最终得到了降噪的增强语音信号。

2　总体模型架构

2.1　编码解码框架(encoder-decoder)

一些基于卷积神经网络的方法被应用于语音增强领域，一般的卷积神经网络的模型由输入层、卷积层、池化层、反卷积层、全连接层和输出层组成，通过卷积层与池化层的级联挖掘特征信息，卷积网络中的权重共享性质可以减少训练参数的数量。考虑到池化层和上采样层会引起信息丢失，不利于增强语音细节信息的恢复，Park 等人[11]提出冗余卷积编码解码网络(redundant convolutional encoder-decoder，R-CED)。

本文将冗余卷积编码解码网络结构作为模型的体系结构。其中编码器中每个构造模块的输出特征图个数逐渐增加，这样编码器可以看做为每个时频点生成一个高维向量表示。这个向量的每一维都代表了一种内在特征，向量维度增加的过程可以看做编码器不断地从上层输出信息中尽可能多地挖掘信息。而这些信息中，可能只有一部分对增强语音语谱图的生成是有用的，剩余的都可看做冗余信息，应该被丢弃。因此在解码器中，每个构造模块中的卷积层的输出语谱图数量逐渐减少，这个过程可以看做是对编码器生成特征的筛选过滤。编码器通过不断地过滤无用信息，保留重要信息。

图 1 展示了模型的总体架构，编码器解码器中的构造模块在 2.2 章节中进行了详细的介绍，2.3 章节描述了编码器解码器之间的瓶颈层的内部细节。

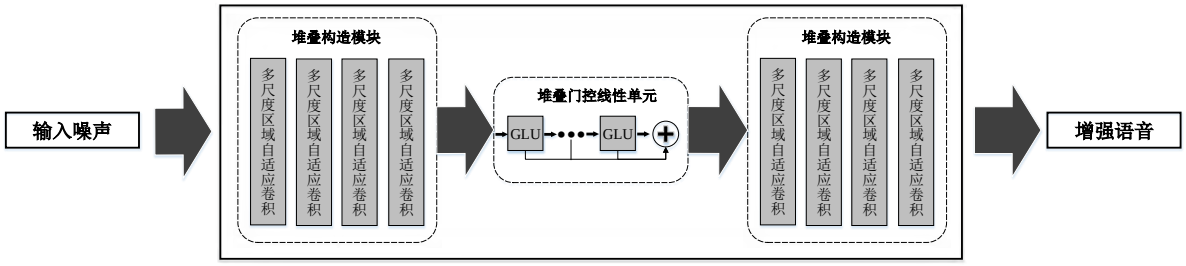

图 1　模型总体架构图

Fig. 1　Overall architecture

2.2　多尺度区域自适应卷积模块

在标准卷积中，卷积核的大小在模型设计阶段被提前设置好，这意味着模型只能以固定大小感受野进行特征抽取，这极大限制了模型的特征感知能力。同时，由于卷积神经网络具有的参数共享性质，在特征过滤的过程中它平等地考虑了同一频谱图中的所有采样点，然而由于不同的采样点存在不同噪声与干净语音比率，对所有样本点采取相同的特征提炼方式将极大限制模型的泛化性。

为了解决上述的问题，本文提出了一种多尺度区域自适应卷积模块，该模块结构如图 2 所示。

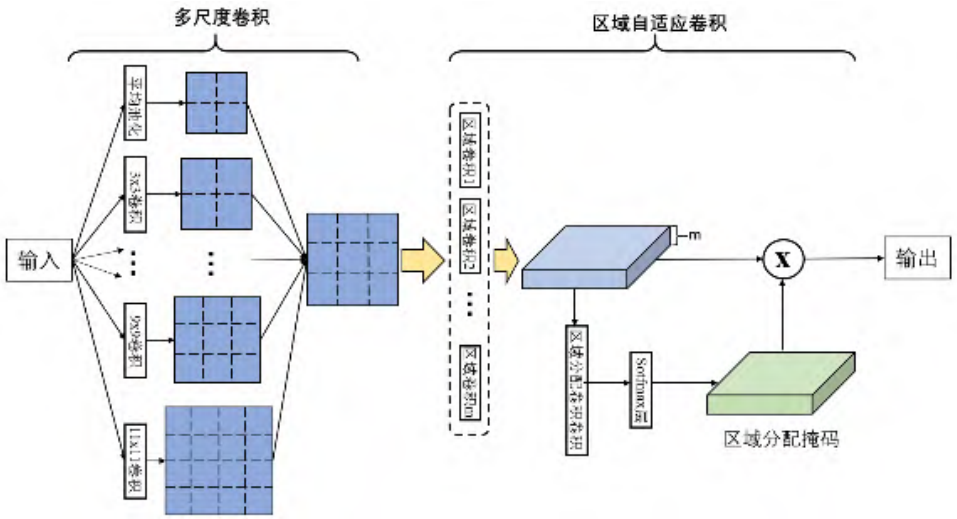

图 2　区域自适应多尺度卷积模块

Fig. 2　Region adaptive multi-scale convolution module

设模块的输入为$x \in \mathbb{R}^{B\times T\times F}$，其中 B、T 和 F 分别表示输入的批量、帧数和频率。首先使用卷积核大小分别为 3、5、7、9 和 11 的卷积层以及一个平均池化卷积对输入进行多尺度卷积来获取不同尺度的信息，输出结果分别记作x_{dl_1}、x_{dl_2}、x_{dl_3}、x_{dl_4}、x_{dl_5}以及x_{avg}。将以上特征图沿着通道维度进行拼接，得到x_{ms}：

$$x_{ms} = [x_avg; x_{dl_1}; x_{dl_2}; x_{dl_3}; x_{dl_4}; x_{dl_5}] \tag{4}$$

其中，；表示拼接操作。

为了保持通道的一致性，将拼接得到的特征图x_{ms}输入到一个 1×1 卷积中进行特征融合，最终得到多尺度卷积结果z。同时，为了防止多尺度卷积过程中梯度消失的问题，在输入以及输出之间添加了残差连接。

接着将得到的特征图z输入到区域自适应卷积模块中，根据输入隐式地将特征图中相似的采样点分配 m 个区域，并对特定区域的采样点，以解决一般卷积方法由于参数共享存在的限制。

将z分别输入到m个相同卷积核大小的卷积$conv_i$ $(i = 1, 2, ..., m)$对其进行处理得到区域卷积结果M_i，并对结果进行拼接得到$M \in \mathbb{R}^{B\times(C\times m)\times T\times F}$。

将M输入到区域分配模块内获得区域分配掩码，具体公式如下：

$$Mask = softmax(conv(W(M))) \tag{5}$$

其中$W \in \mathbb{R}^{(C\times m)\times C}$。

将得到的区域分配掩码用来为每个采样点进行区域卷积的权值分配。在这个过程中，根据当前特征图信息来对不同区域卷积层获得的信息分配相应的权重，实现区域自适应卷积。

2.3　扩张线性门控单元层

在瓶颈层阶段，参考[21]使用线性门控单元来对信息流进行控制。同时，为了增大模型感受野同时避免过大的参数量，使用扩张卷积[22]代替了线性门控单元内的一般卷积。

线性门控单元的结构如图 3 所示。设输入为$x \in \mathbb{R}^{H\times W\times C}$，首先使用 1×1 卷积来改变其通道数。然后分别使用两个扩张卷积层对其进行卷积操作，将得到的结果点乘，实现基于自注意力机制的门控机制筛选和过滤噪声的功能。

$$\mathbf{y} = (\mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1) \odot \sigma(\mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2) \tag{6}$$

使用扩张卷积增大了感受野，使模型获得更全局的信息，但是其在卷积过程中容易遗漏一些局部信息，这对于密集预测的任务十分不利[23]。因此在多个线性门控单元之间，使用跳连接(skip connection)方式来对这些信息进行融合，这样模型能够弥补损失的信息，有利于提升模型的泛化性。

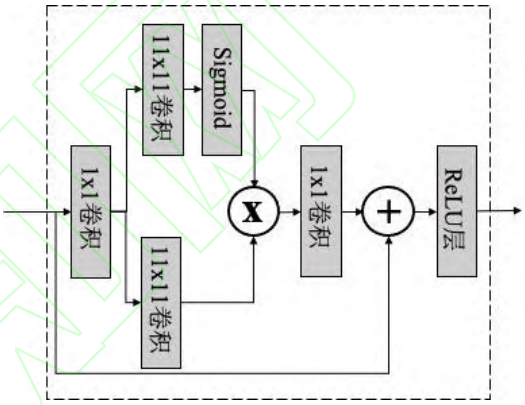

图 3　扩张线性门控单元

Fig. 3　Dialated linear gating unit

3　实验

3.1　实验设置

实验选取 TIMIT 公开语音数据集作为干净语音数据集，选取 Noisex92 中的噪声作为实验中的噪声数据集。在训练过程中，首先选取 TIMIT 训练集中的非方言语句，共 3696 条作为干净语音片段；选取 Noisex92 中的 babble、factory1、destroyerops 和 destroyerengine 噪声片段作为训练噪声，并随取选取一个噪声以随机信噪比(-5，0，5)进行混合。在测试过程中，选取 TIMIT 核心测试中的非方言语句，共 192 条干净语音片段；除了训练集中出现的噪声，还另外选取 Noisex92 中的非平稳噪声 buccaneer1 和平稳噪声 white 进行混合；除了训练集中出现的信噪比，还另外选取了-10 和 10dB 来测试模型的泛化性，与训练阶段不同的是，为了避免实验的偶然性，对于测试阶段的每一条干净语音，将其分别与每一种噪声以不同的信噪比进行混合，共得到了 192*5*6 条带噪语音信号。

将所有的语音片段降采样到 8kHz，并使用窗口大小与帧长分别为 32ms 和 16ms 的汉明窗，对语音信号进行短时傅里叶变换(Short-Time Fourier Transform，STFT)，实现将语音信号从时域转换到时频域。使用均方误差(Mean Square Error，MSE)作为损失函数，并通过 Adam 优化器来优化模型参数，学习率设置为 0.0002。

为了直观体现模型性能，实验选取了短时客观可懂度(Short-Time Objective Intelligibility，STOI)和语音质量感知评估(Perceptual Evaluation of Speech Quality，PESQ)作为评价指标。其中，STOI 的取值范围为[0,1]，得分越高说明语音的可

懂度越好；PESQ 的取值范围为[-0.5,4.5]，得分越高说明语音质量越好。

3.2 模型对比

实验过程中，本文将如下模型进行对比，来验证所添加机制的作用：

(a) Net：基于编码器解码器的结构，其中瓶颈层部分采用了扩张先行门控单元层。

(b) Mnet：在 Net 的基础上，将编码器解码器结构中的一般卷积层替换为多尺度卷积模块。

(c) RMNet：文中所提出的模型，是在 Net 的基础上，将编码器解码器结构中的所有一般卷积层替换为多尺度区域自适应卷积模块。

此外，为了证明所提出模型的有效性和先进性，选取如下近期模型用于对比：

(a) RCED[10]：该模型主体由四个冗余编码器和四个冗余解码器组成，并使用 skip connection 操作连接编码器和解码器对应的层。

(b) RTNet[24]：该模型基于渐进学习，在时域实现。本文在时频域进行复现，并使用 2D 卷积来代替 1D 卷积，使之适用于时频域。

(c) RCNA[25]：该模型包含九个 ACB 模块和一个卷积层。每个 ACB 模块包含一个卷积层、一个批归一化层(Batch Normalization，BN)和一个 relu 激活层和注意力机制层。

3.3 实验结果

表 1 和 2 分别给出了在可见噪声和不可见噪声条件下，经过各模型增强后的语音的 STOI 和 PESQ 指标，其中每个值表示该模型在所有信噪比条件下的平均值。

表 1　在可见噪声条件下各个模型的平均 STOI 和 PESQ 得分

Tab. 1　Average STOI and PESQ scores of each model under seen noise conditions

Metric	Model	Babble	Destroyerengine	Factory1	Destroyer-ops
STOI (%)	Noisy	65.51	67.98	65.14	68.82
	Net	74.84	83.60	77.80	81.64
	MNet	76.90	84.88	79.11	83.12
	RMNet	**77.46**	**85.35**	**79.56**	**83.50**
	RCED	72.90	82.22	76.04	80.04
	RTNet	75.46	84.00	78.27	81.94
	RCNA	74.41	82.60	76.27	80.20
PESQ	Noisy	1.89	1.84	1.81	1.91
	Net	2.41	2.70	2.52	2.67
	MNet	2.46	2.78	**2.57**	2.74
	RMNet	**2.49**	**2.79**	**2.57**	**2.76**
	RCED	2.23	2.57	2.40	2.53
	RTNet	2.24	2.53	2.35	2.50
	RCNA	2.33	2.56	2.42	2.54

表 2　在不可见噪声条件下各模型的平均 STOI 和 PESQ 得分

Tab. 2　Average STOI and PESQ scores of each model under unseen noise conditions

Model	STOI		PESQ	
	Buccaneer1	White	Buccaneer1	White
Noisy	63.10	71.14	1.66	1.74
Net	73.83	80.42	2.24	2.61
MNet	74.36	81.18	2.25(2.28)	**2.64**
RMNet	**74.81**	**81.52**	**2.26**	**2.64**
RCED	71.78	78.91	2.12	2.48
RTNet	73.62	79.98	2.13	2.48
RCNA	73.04	80.27	2.18	2.57

首先将本文提出的 RMNet 方法与近期的实验成果对比，从表 1 中可以直观的看到，在不同可见噪声条件下，RMNet 方法相比于其他方法均实现了明显的性能提升，其中 STOI 得分最高提升了 2.00，PESQ 最高提升了 0.23。实验结果证明本文提出的模型具有更强的特征抽象能力以及去噪能力。这是因为传统卷积神经网络受限于单一形式卷积核，难以拟合复杂的声学环境，限制了模型的特征表达以及过滤能力。同时，从表 2 中可以发现，即使是在不可见噪声的条件下，RMNet 方法与近期的方法相比在各项指标上仍能取得明显的性能提升，其中 STOI 得分最少提升了 1.19，PESQ 得分最少提升了 0.08。这样的实验结果显示了本文提出的模型 RMNet 具有更强的泛化性能，本文认为这是因为提出的区域自适应卷积模块可以根据输入自适应的改变区域卷积权重，这极大增加了模型处理不同类型噪声的能力，同时多尺度卷积提升了模型的特征感知能力。从表中可以看出，所提出方法的 STOI 和 PESQ 指标在所有噪声条件下，均明显优于 GRN、RTNet 和 RCNA 方法，证明了模型的先进性。

为了验证模型在复杂噪声环境下的性能，表 3 中展示了信噪比-10dB 条件下模型在不同噪声下的实验性能，以更好地验证模型的泛化性和有效性。从表中可以看出，在 STOI 指标上，MNet 方法相比于 Net 方法实现了显著提升，这得益于多尺度卷积带来的特征表达能力的提升，增强了模型处理复杂噪声环境的能力。此外，在-10dB 条件下，RMNet 方法相比于 MNet 方法，在所有可见噪声条件下都实现了 STOI 指标上的提升，且在不可见噪声条件下可以观察到同样的增长趋势，这说明该模型能够更好的处理低信噪比条件下的带噪语音，这是由于区域自适应卷积模块通过区域卷积实现频谱图上的区域特定卷积，并且能够根据输入自适应调整区域卷积的赋权，这提升了模型处理噪声的能力以及泛化性。在 PESQ 指标下，与 Net 方法相比，MNet 方法和 RMNet 方法虽然有提升，但是不算明显。

表 3　10dB 下各噪声环境下的消融实验结果

Tab. 3　Results of ablation experiments under various noise environments at -10db

Metric	Noise	Model		
		Net	MNet	RMNet
STOI	Babble	46.88	50.10	**50.77**
	Destroyerengine	65.79	67.92	**68.80**
	Factory1	55.76	57.48	**58.13**
	Destroyerops	63.78	66.19	**66.79**
	Buccaneer1	51.05	51.12	**52.16**
	White	63.65	63.70	**64.38**
PESQ	Babble	1.54	1.55	**1.57**
	Destroyerengine	2.03	2.07	**2.08**
	Factory1	1.79	**1.80**	**1.80**
	Destroyerops	2.00	**2.03**	**2.03**
	Buccaneer1	1.51	**1.53**	1.52
	White	2.61	2.63	**2.64**

4　结束语

本文提出了基于时频域 RMNet 模型来实现单声道语音增强。为了解决一般卷积神经网络存在的由于感受野大小固定导致的特征表达受限的问题以及由于参数共享机制影响了模型去除噪声能力的问题，提出使用区域自适应多尺度卷积模块，在提升了模型整体性能的同时提升了模型的泛化能力，增强了模型处理复杂噪声的能力。实验证明所提出的模型在 STOI 和 PESQ 等常见指标上显著优于近期所提出的语音增强算法。在未来研究中，将尝试对频谱图中的每一个像素进行特定处理，进一步提升模型处理带噪语音的能力。

参考文献：

[1] Donahue C, Li B, Prabhavalkar R. Exploring speech enhancement with generative adversarial networks for robust speech recognition [C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 2018: 5024-5028. \

[2] Michelsanti D, Tan Z H. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification [J]. arXiv preprint arXiv: 1709. 01703, 2017.

[3] Guo H, Pu X, Chen J, *et al.* A highly sensitive, self-powered triboelectric auditory sensor for social robotics and hearing aids [J]. Science robotics, 2018, 3 (20) .

[4] Wang D L, Chen J. Supervised speech separation based on deep learning: An overview [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26 (10): 1702-1726.

[5] 张明亮，陈雨. 基于全卷积神经网络的语音增强算法 [J]. 计算机应用研究，2020, 1. (Zhang Mingliang, Chen Yu. Speech Enhancement Algorithm Based on Fully Convolutional Neural Network [J]. Application Research of Computers, 2020, 1.)

[6] Wang Y, Wang D L. Towards scaling up classification-based speech separation [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21 (7): 1381-1390.

[7] 袁文浩，孙文珠，夏斌，等. 利用深度卷积神经网络提高未知噪声下的语音增强性能 [J]. 自动化学报，2018, 44 (4): 751-759. (Yuan Wenhao, Sun Wenzhu, Xia Bin, *et al.* Using deep convolutional neural network to improve speech enhancement performance under unknown noise [J]. IEEE/CAA Journal of Automatica Sinica (JAS) , 2018, 44 (4): 751-759.)

[8] Li X, Li Y, Li M, *et al.* A Convolutional Neural Network with Non-Local Module for Speech Enhancement [C]// INTERSPEECH. 2019: 1796-1800.

[9] Pandey A, Wang D L. Dense CNN with Self-Attention for Time-Domain Speech Enhancement [J]. arXiv preprint arXiv: 2009. 01941, 2020.

[10] Park S R, Lee J. A fully convolutional neural network for speech enhancement [J]. arXiv preprint arXiv: 1609. 07132, 2016.

[11] Fu S W, Hu T, Tsao Y, *et al.* Complex spectrogram enhancement by convolutional neural network with multi-metrics learning [C]// 2017 IEEE 27th international workshop on machine learning for signal processing (MLSP) . IEEE, 2017: 1-6.

[12] Tan K, Wang D L. A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement [C]// Interspeech. 2018: 3229-3233.

[13] Wang P, Chen P, Yuan Y, *et al.* Understanding convolution for semantic segmentation [C]// 2018 IEEE winter conference on applications of computer vision (WACV) . IEEE, 2018: 1451-1460.

[14] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex [J]. The Journal of physiology, 1962, [0 (1): 106-154.

[15] Szegedy C, Liu W, Jia Y, *et al.* Going deeper with convolutions [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[16] Xian Y, Sun Y, Wang W, *et al.* A Multi-Scale Feature Recalibration Network for End-to-End Single Channel Speech Enhancement [J]. IEEE Journal of Selected Topics in Signal Processing, 2020.

[17] Gregor K, LeCun Y. Emergence of complex-like cells in a temporal product network with local receptive fields [J]. arXiv preprint arXiv: 1006. 0448, 2010.

[18] Sun Y, Wang X, Tang X. Deep learning face representation by joint identification-verification [J]. arXiv preprint arXiv: 1406. 4773, 2014.

[19] Chen J, Wang X, Guo Z, *et al.* Dynamic region-aware convolution [J]. arXiv preprint arXiv: 2003. 12243, 2020.

[20] Luo W, Li Y, Urtasun R, *et al.* Understanding the effective receptive field in deep convolutional neural networks [J]. arXiv preprint arXiv: 1701. 04128, 2017.

[21] Tan K, Chen J, Wang D L. Gated residual networks with dilated convolutions for monaural speech enhancement [J]. IEEE/ACM transactions on audio, speech, and language processing, 2018, 27 (1): 189-198.

[22] Chen L C, Papandreou G, Schroff F, *et al.* Rethinking atrous convolution for semantic image segmentation [J]. arXiv preprint arXiv: 1706. 05587, 2017.

[23] Wang P, Chen P, Yuan Y, *et al.* Understanding convolution for semantic segmentation [C]// 2018 IEEE winter conference on applications of computer vision (WACV) . IEEE, 2018: 1451-1460.

[24] Li A, Zheng C, Cheng L, *et al.* A Time-domain Monaural Speech Enhancement with Recursive Learning [J]. arXiv e-prints, 2020: arXiv: 2003. 09815.

[25] Lan T, Lyu Y, Hui G, *et al.* Redundant Convolutional Network With Attention Mechanism For Monaural Speech Enhancement [C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 2020: 6654-6658.