基于动态选择机制的低信噪比单声道语音增强算法*

台文鑫，王钇翔，李 森，蓝 天，刘 峤
（电子科技大学 信息与软件工程学院，成都 610054）

摘 要：为了提升模型在复杂场景下的信息处理能力，提出了一种基于注意力的动态选择机制，根据当前信息选择性地分配权重，有效融合形变卷积和普通卷积的特征输出，自适应地在卷积形变和标准卷积之间做权衡，从而提高其表示能力。此外，通过借鉴渐进学习，在不增加额外参数的前提下，通过循环迭代的方式进一步增强了模型的学习能力。在 TIMIT 公开语料库上使用 7 种来自 NoiseX92 的不同噪声，在多种信噪比环境下进行实验，结果表明无论信噪比高低，噪声是否在训练数据集中出现，所提出的算法在可懂度和语音质量等客观评价指标上均优于近期其他的深度学习算法。

关键词：语音增强；低信噪比；动态选择机制；形变卷积；渐进学习
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2021)09-007-2604-05
doi:10.19734/j.issn.1001-3695.2020.12.0549

Monaural speech enhancement algorithm based on deformable convolution

Tai Wenxin, Wang Yixiang, Li Sen, Lan Tian, Liu Qiao
(*School of Information & Software Engineering, University of Electronic Science & Technology of China, Chengdu* 610054, *China*)

Abstract: In order to improve the information processing ability of the model in complex scenes, this paper proposed a dynamic selection mechanism based on attention, which selectively allocated weights according to the current information, effectively fused the feature outputs of deformation convolution and ordinary convolution, and adaptively balanced deformable convolution and standard convolution, so as to improve its representation ability. In addition, the learning ability of the model is further enhanced by means of iteration without additional parameters. It used seven different kinds of noises from Noise-X92 in TIMIT corpus, and carried out experiments in various SNR environments. The results show that the proposed algorithm outperforms other recent deep learning algorithms in terms of intelligibility and speech quality, regardless of SNR and whether noise appears in the training data set.

Key words: speech enhancement; low SNR; dynamic selection; deformable convolution; progressive learning

0 引言

现实复杂场景下存在大量环境噪声，这将会严重降低语音信号的质量。因此，降低背景噪声、提高带噪语音的质量和清晰度成为了语音处理应用中亟待解决的关键问题。语音增强的主要目标是从含噪语音中提取原始纯净语音信号，通过抑制或分离噪声来提升语音的可懂度和感知质量。其经常被用作其他语音处理任务的预处理过程，例如声纹识别[1]、语音识别[2]和助听设备[3]。相比于多麦克风阵列的语音增强任务，单声道的语音增强任务难度更高。近年来，语音增强在语音处理领域得到了广泛的研究[4]。传统的语音增强方法大多基于信号处理，包括谱减法[5]、维纳滤波[6,7]和基于统计模型的方法[8]。这些方法主要是利用无监督的语音信号分析算法，通过分解语音信号来提取干净语音和噪声的特征，继而将噪声从混合语音信号中分离出来。然而，由于过分依赖人为假设，上述方法仅适用于噪声平稳或缓慢变化的场景[9]。

深度学习的兴起以及在声学领域的成功应用，为解决复杂环境下的语音增强提供了建模思路。与经典方法相比，基于深度学习的方法显著提高了模型在非平稳噪声场景下的性能。Wang 等人[10]首先将深度神经网络(deep neural network, DNN)应用于语音分离，然后训练二值分类器预测理想二值掩码(ideal binary mask, IBM)，以达到去除背景噪声的目的。Wang 等人[11]后来证明理想比率掩码(ideal ratio mask, IRM)能够产生比 IBM 更好的语音质量。Xu 等人[12]利用深度回归方法学习噪声语音的对数功率谱到干净语音的对数功率谱的映射函数。在复杂环境下，上述方法的有效性得到了证实。然而，基于 DNN 的模型需要大量的参数，这导致了训练的瓶颈。

随着卷积神经网络(convolutional neural network, CNN)的兴起，许多研究者试图将其应用于语音处理领域。与 DNN 相比，共享卷积核的 CNN 体系结构具有更快的计算速度和更低的空间复杂度。受图像去噪领域的基于编解码结构的 CNN 模型架构的启发[13]，一些研究者尝试将这种编码解码的方式也应用于语音增强任务。Part 等人[14]认为传统的下采样会导致信息丢失，使得解码器更难重建干净语音信号。因此，他们提出了一种冗余卷积编译码网络结构，并验证了其有效性。此外，一些工作融合了卷积神经网络和递归神经网络，例如文献[15,16]，相比于单一卷积网络取得了更好的性能。Tan 等人[17]在传统的编码解码结构中间加入了瓶颈层，基于门控线性单元设计了一种残差门控机制，以更好地筛选过滤有效信息。

收稿日期：2020-12-25；**修回日期**：2021-02-24 **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：台文鑫(1997-)，男，甘肃兰州人，硕士研究生，主要研究方向为语音识别、语音增强、推荐系统等(wxtai@std.uestc.edu.cn)；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、情感分析等；李森，男，安徽淮北人，硕士研究生，主要研究方向为语音识别、语音增强等；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；刘峤，男，四川成都人，教授，博导，博士，主要研究方向为知识图谱、自然语言处理、深度学习.

由于潜在环境的复杂性,对于信噪比较低的场景,建模难度会进一步提高[18]。在低信噪比条件下,噪声分量高,干净语音被噪声所淹没,从中恢复出原始信号的难度大幅增加。为了解决低信噪比条件下的语音增强任务,总的来说有三种方式。方式一是从模型构建角度入手,增强模型的信息处理能力,从根本上提升降噪性能。方式二是通过外部记忆机制,引入额外的知识例如环境表征等辅助决策。方式三是通过信息融合,例如融合时域和时频域等多角度信息以提升模型表现。本工作基于方式一,通过更加科学的模型架构设计来提升模型处理信息的能力。为了增强模型的性能,一些工作受视觉皮层中同一区域(例如 V1 区)神经元的局部感受野大小不同[19],使神经元能够在同一阶段收集多尺度的空间信息[20]此生物学机制的启发,试图设计相应的架构以模仿生物学行为。InceptionNet[21]从空间维度入手,结合多尺度信息,聚合多种不同感受野的特征来获取额外的信息增益。SENet[22]指出,InceptionNet 等很大一部分工作基本上默认对输入特征图的所有通道进行融合,不同通道之间默认具有相同的重要权重。其更加关注通道之间的潜在联系,并提出了一种压缩—激发架构的注意力机制,在 ImageNet 等分类任务上收益显著。Lan 等人[23]受 SENet 设计的启发,将该方案引入到语音增强任务中,提出了基于通道注意力的语音增强模型 RCNA,该模型在以单帧为输入的语音增强任务上优于其他模型。Li 等人[24]认为注意力机制有助于模型更好地进行信息融合,其设计了一个自网络,用来自适应的产生注意力权重分布,在解码阶段融合对应位置的编码特征,以获得更好的频谱估计。Li 等人[25]指出 Inception 线性融合这种方式的表征能力有限,通过结合 Inception 的多尺度和 SENet 的压缩—激发机理,提出了一种基于核选择的注意力机制,在分类等图像领域任务上相比于 SENet 有了更进一步的性能提升。除此之外,Dai[26]从另固定参数的卷积核设计限制了传统卷积的提取能力。其提出了一种形变卷积,根据输入动态地学习每个位置的偏移量,实现卷积核的自适应变形,该架构增强了模型的信息处理能力,并成功应用于目标检测等图像领域的任务。文献[27]通过实验证明,尽管形变卷积对于特征的空间支持比规则的卷积网络更接近于区域对象结构,但在某些情况下由于卷积核过度的可调节性质,可能产生负增益现象,导致局部特征受到不相关区域的影响,反而弱化了模型的信息筛选能力。

标准的傅里叶分析仅适用于全局平稳信号,而语音信号具有非平稳特点,因此研究人员往往采用基于短时傅里叶变换对语音信号进行分析。受短时傅里叶变换分帧、加窗、重叠等机制的影响,最终得到的时频图存在随时间分布变换的频谱纹理,纹理特征在形状大小和几何方向上的不确定性增加了对模型处理信息能力的要求。受上述工作的启发,本文提出了一种动态选择机制的信息融合方案,有效融合形变卷积和普通卷积的输出特征,根据不同的刺激动态调节当前感受野,在自适应性和稳定性之间做了一种动态的权衡,以更好地应对低信噪比等复杂噪声场景。为了进一步提升模型性能,本文借鉴了时域模型 RTNet[28]的渐进学习机制,并将其迁移到时频域中,利用共享参数的循环机制,在不增加模型复杂度的情况下通过多次迭代,进一步提升了模型表现。本文首先利用短时快速傅里叶变换,将时域波形转换到时频域并获得二维时频空间表示,然后通过一个编码解码卷积神经网络,在幅度谱上进行噪声消除。在解码阶段,本文利用基于注意力的动态选择机制,在通道维度自适应地融合可变形卷积和传统卷积。最后通过短时快速傅里叶逆变换,利用增强后的幅度谱和带噪相位恢复信号。

本文创新点如下:

a)首次将形变卷积引入语音增强领域,通过可变形卷积提升了模型信息的信息处理能力,使其能够拟合局部声纹纹理并更好地应对低信噪比复杂环境。

b)提出了一种基于注意力的动态选择机制,其有效融合传统卷积与形变卷积的输出特征,一方面弱化了部分情况下由于形变卷积的自适应性导致的负增益,另一方面通过信息融合增强了模型信息处理能力,使其能够更好地应对复杂场景。

c)在公开数据集 TIMIT 的实验结果表明,本文所提出的模型在 STOI、PESQ 等指标上的表现更好,在低信噪比环境时仍能保持一定的降噪性能。

1　研究问题描述

在时域,复杂场景下的带噪语音通常可以表示为

$$x(n)=s(n)+d(n) \tag{1}$$

其中:x、s 和 d 分别代表带噪语音、干净语音和噪声波形,n 代表时间帧的索引。值得注意的是,通常情况下语音的时间帧总数不是固定的,因为不同的语音片段往往具有不同的持续时间。给定一个长度为 N 的实值向量 $\boldsymbol{X}$,可通过短时傅里叶变换将其转换为时频域。因此,可以将其重写为

$$X_{t,f}=S_{t,f}+D_{t,f} \tag{2}$$

其中:$X_{t,f}$、$S_{t,f}$和 $D_{t,f}$分别代表带噪语音、干净语音和噪声在时间帧 t 和频点 f 时的值。在直角坐标系中复数谱 $X_{t,f}$又可以写为实部和虚部的组合:

$$X_{t,f}=X^r_{t,f}+X^i_{t,f} \tag{3}$$

近期的工作[21]提出,直接在实数谱和虚数谱上面利用均方误差等损失函数进行计算,其效果往往低于基于幅度谱的方式。因此,本文通过组合实数谱和虚数谱来获得带噪幅度谱:

$$X^m_{t,f}=\sqrt{(X^r_{t,f})^2+(X^i_{t,f})^2} \tag{4}$$

通过神经网络模型,获取增强后的幅度谱,并利用带噪相位和快速傅里叶逆变换还原到时域空间,最终得到降噪后的语音波形。

2　总体模型架构

2.1　形变卷积

标准卷积单元在固定位置对输入特征图进行采样,并通过计算样本的加权和生成对应位置的输出。由于预先固定了采样形状和卷积核大小,标准卷积缺乏根据输入特征自适应调整的能力。近期,文献[19]提出的形变卷积可以利用一种添加偏移量的方式,巧妙地解决了标准卷积核固定形状大小的问题,其整体架构如图 1 所示。本文认为,语音信号在转换到时频域空间时,受窗函数和帧移等短时快速傅里叶变换操作手段的影响,语音时频图中形成了密集的频谱纹理。借助于形变卷积对图像领域几何变换所表现出来的强适应性,本文将形变卷积迁移到语音领域,使用附加卷积层估计每个网格位置的偏移量,然后计算偏移后位置的加权值以重建输出特征。偏移量的引入使网络对目标的几何变形具有一定的适应性,提高了语音增强模型处理复杂场景的能力,特别是在信噪比较低或噪声类型不可见的情况下。假设定义一个大小为 3×3,步长为 1 的卷积核,应有如下 9 个位置的采样点:

$$R=\{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\} \tag{5}$$

常规标准卷积的做法是对固定矩形位置进行采样,继而根据对应位置加权求和得到当前的输出值:

$$y(p_0)=\sum_{p_n\in\mathbb{R}}w(p_n)\cdot x(p_0+p_n) \tag{6}$$

其中:w 代表对应位置的卷积核权重,而 x 代表相应的位置。对于形变卷积来说,其在每个位置通过一个额外的卷积操作获得响应位置的偏移量,使得采样的总体形状不再局限于标准大小 3×3 矩形:

$$y(p_0)=\sum_{p_n\in\mathbb{R}}w(p_n)\cdot x(p_0+p_n+\Delta p_n) \tag{7}$$

通过引入偏移量,模型在信息处理过程中具备了一定的自适应性,这种可调整性质增强了模型应对复杂场景的能力,例如当信噪比较低或者噪声种类未曾见过时。值得注意的是,由于偏移量是由卷积产生的,不一定是个整数,所以在具体计算过程中需要使用双线性插值的方法求解。

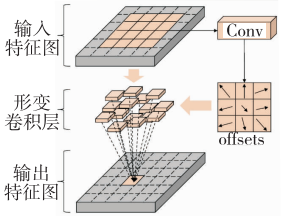

图 1 形变卷积

Fig. 1 Deformable convolution

2.2 基于注意力的动态选择机制

形变卷积过度的自适应性在某些情况下往往会造成负增益的现象。受 SKNet[23] 的启发,本文提出了一种基于注意力的动态选择机制,通过全局平均池化编码全局信息,进而构建一种门控机制,利用通道之间的软注意力机制,动态的控制流入下一个卷积层中不同分支的信息流,具体操作如图 2 所示。

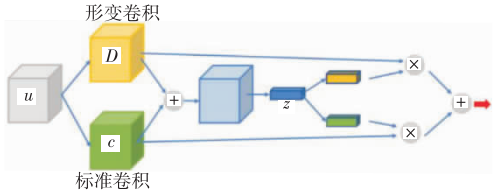

图 2 基于通道注意力的动态选择机制

Fig. 2 Dynamic selection based on channel attention

首先本文融合形变卷积与标准卷积的信息,并通过全局平均池化编码得到包含全局信息的压缩向量:

$$s_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} C_c(i,j) + D_c(i,j) \tag{8}$$

紧接着,对已获取的密集向量 s_c 进行一次空间变换,为精确和自适应的选择相应的分支提供指导:

$$z = W \cdot s + b \tag{9}$$

在向量 z 的基础上,再通过相应的线性变换获得不同支路的信息流控制向量,在通道维度使用软注意力机制自适应地控制不同支路地信息流量:

$$a_c = \frac{e^{A_c z}}{e^{A_c z} + e^{B_c z}}, b_c = \frac{e^{B_c z}}{e^{A_c z} + e^{B_c z}} \tag{10}$$

其中:$\boldsymbol{A}$、$\boldsymbol{B}$ 分别对应了不同支路的变换矩阵,偏执向量 $\boldsymbol{b}$ 在此处被省略。最后,根据不同支路的注意力权重得到最终的输出特征:

$$O_c = a_c \cdot C_c + b_c \cdot D_c \tag{11}$$

2.3 渐进学习

当模型逐渐加深,伴随而来的是参数量的急剧增加以及梯度消失的问题,这些问题都在一定程度上损害了深度神经网络的性能。

本文借鉴了时域语音增强模型 RTNet 的渐进学习架构来缓解上述问题,该架构在之前的基础上进一步增强了模型的信息处理能力。该架构利用了一个存储机制来整合神经网络在各个阶段抽象得到的信息,并在每一个阶段对整合的信息进行迭代。渐进学习网络架构主要由两个部分组成,即一个 2D 的卷积模块和一个卷积-RNN 模块。第一个部分的作用是将输入特征抽象为一个隐向量表示,接着第二部分主要负责利用卷积-RNN 模块来更新当前阶段模型的状态。假设 2D 卷积以及卷积-RNN 模块在阶段的输出分别为 和 ,渐进学习架构为

$$\hat{h}^l = f_{conv}(|X|, |\tilde{S}^{l-1}|) \tag{12}$$

$$h^l = f_{conv\_rnn}(\hat{h}^l, h^{l-1}) \tag{13}$$

其中:f_{conv} 以及 f_{conv_rnn} 分别表示 2D 卷积以及卷积 RNN 模块的函数,h^{l-1} 是上一阶段的状态。其中,本文使用 Conv-GRU 作为一个卷积 RNN 单元,计算公式如下所示。

$$z^l = \sigma(\boldsymbol{W}_z^l \otimes \hat{h}^l + \boldsymbol{U}_z^l \otimes h^{l-1}) \tag{14}$$

$$r^l = \sigma(\boldsymbol{W}_r^l \otimes \hat{h}^l + \boldsymbol{U}_r^l \otimes h^{l-1}) \tag{15}$$

$$n^l = \tanh(\boldsymbol{W}_n^l \otimes \hat{h}^l + \boldsymbol{U}_n^l \otimes (r^l \odot h^{l-1})) \tag{16}$$

$$h^l = (1 - z^l) \odot \hat{h}^l + z^l \odot n^l \tag{17}$$

其中:$\boldsymbol{W}$ 和 $\boldsymbol{U}$ 分别表示元件内的权重矩阵,$\sigma(\cdot)$ 和 $\tanh(\cdot)$ 分别表示 sigmoid 以及 tanh 激活函数,$\otimes$ 表示卷积操作,$\odot$ 表示点乘操作。

2.4 总体架构

基于 U-Net 的编码器—解码器体系结构在近期的语音增强领域被广泛应用。本文模型架构包含四层编码器与四层解码器,在此基础上提出了一种基于形变卷积的渐进学习网络,如图 3 所示。其中,编码器的每一层都使用大小为 11 × 11 的标准卷积,解码器的每一层都使用基于通道注意力的动态选择模块。在模型的中间部位,使用了一个堆叠线性门控单元的模块来进一步提取信息。本文参考文献[21]中基于自注意力机制的门控单元的设置,使用扩张卷积替代每个分支的线性变换。为了降低模型的空间复杂度,首先使用卷积核大小为(1 × 1)的二维卷积来降低输入维数,然后,利用基于自注意力机制的门控机制筛选和过滤噪声信息:

$$y = (x * W_1 + b_1) \odot \sigma(x * W_2 + b_2) \tag{18}$$

此外,每个门控单元还增加了残差连接,以防止由于堆叠门控单元造成的梯度消失问题。

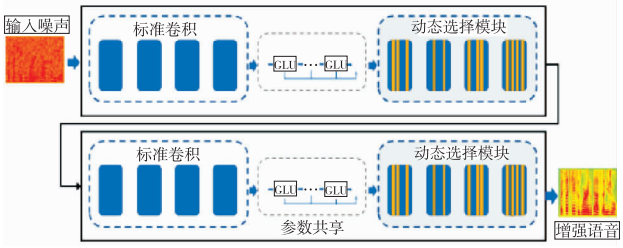

图 3 模型总体架构

Fig. 3 The overall architecture

3 实验设置描述

3.1 数据集描述

本文使用 TIMIT 开源数据集构建实验环境,并将所有的语音信号统一下采样到 8 KHz。TIMIT 包含 6 300 条语句,其来自 8 个主要方言地区的 630 名说话人,每人说 10 句话。从训练集中去除方言句子,并使用剩下的 3 696 个句子进行训练。使用包含 192 条语句的官方数据构造实验所需的验证集和测试集。干净语音与四种类型的噪声(babble、factory1、destroyers ops 和 destroyer engine)以及三种不同的信噪比(-5、0 和 5 dB)随机结合构建训练数据集和验证集。在测试集中,每一条语音都分别与七种噪声(包括三种不可见噪声:buccaneer1,factory2 和 white)以各种不同的信噪比(-10,-5,0,5)混合共计 3 584 条测试语音。测试集中加入的三种不可见噪声和 -10 dB信噪比主要用来验证所提出模型的泛化性。所有使用的 7 种噪声均来源于 NoiseX92 数据集。

3.2 实验环境与实验设置

实验选择了短时目标可懂度(STOI)和感知语音质量评估方法(PESQ)作为评估准则。其中,STOI 是短时客观可懂度,

用于衡量语音可懂度,得分在 0 ~ 1,分值越高表明语音质量越好。PESQ 是 ITU-T(国际电信联盟电信标准化部)推荐的语音质量评估指标,是一种用来评价语音的主观试听效果的客观计算方法,能够很好地近似平均意见得分。PESQ 的值在 - 0.5 和 4.5 之间,值越高表示语音越清晰。在数据预处理期间,本文使用汉明窗将信号分割为一组帧,窗口大小为 20 ms,相邻时间帧重叠 50%,因此步长为 10 ms。之后,利用 256 个点的 STFT 得到频率轴为 129 维的二维幅度谱作为模型的输入。

在训练阶段,为了保证实验环境的一致性,所有比较模型(包括本文提出的模型)均使用 Adam 优化器迭代 60 轮次,学习率设置为 0.002,并选择 MAE 损失作为损失函数。在每一个小批量样本数据中,都预先进行了零填充,以保证所有数据与当前批量中最长的样本具有相同的时间帧数。此外,所有实验结果取了 3 次实验的平均值,以降低实验结果的偶然性。

4　实验结果及分析

4.1　性能评估

为验证本文模型的有效性,与近期的几个经典算 法进行比较,分别为 RCED[12]、GRN[15]和 RCNA[21]和 DARCN[22]。表 1 为不同信噪比条件下的降噪性能。

表 1　不同信噪比下的测试结果

Tab. 1　The result under different SNRS

指标	模型	噪声(可见 + 不可见)			
		- 10	- 5	0	5
STOI/%	(noisy)	46.32	56.38	67.65	78.41
	(RCED)	55.42	69.64	80.27	87.71
	(GRN)	55.92	70.5	80.83	87.05
	(RCNA)	54.7	69.32	79.8	86.38
	(DARCN)	59.6	72.48	81.12	86.7
	(本文算法)	**59.82**	**73.21**	**82.63**	**88.93**
PESQ	(noisy)	1.27	1.5	1.78	2.1
	(RCED)	1.71	2.12	2.48	2.79
	(GRN)	1.69	2.1	2.45	2.72
	(RCNA)	1.69	2.11	2.44	2.7
	(DARCN)	1.81	2.18	2.47	2.69
	(本文算法)	**1.83**	**2.24**	**2.61**	**2.91**

首先,从数值上可以直观地观察到,本文模型在各种 信噪比条件下均优于几种经典模型。GRN 中独特的门控机制使其具备过滤无效信息的能力,因此其表现优于完全使用卷积架构的 RCED。相比于 RCED,RCNA 在单帧上虽然有优异的性能表现,但当输入为整条语音信号的时频图时,基于局部的注意力抑制方式对于全局的信息把控能力较弱,反而导致了负增益现象。与同样使用了注意力机制的 DARCN 相比,本文模型具备自适应形变卷积核的能力,在低信噪比等复杂场景条件下具有更优异的性能。其中,在 - 10 dB 的条件下,基于注意力机制的 DARCN 和本文模型远优于其他模型,这证明注意力机制所提供的动态调整能力契合低信噪比等复杂场景的降噪需求。在低信噪比环境下,噪声和干净语音严重混叠,传统的卷积神经网络在复杂场景下的特征提取的能力有限,而基于形变卷积和普通卷积的自适应性动态选择机制可以通过刺激相应的调整感受野,在自适应性和稳定性之间权衡,因而能更好地提取干净语音信息;此外,渐进学习在不增加模型计算负载的前提下,通过多阶段迭代去噪的方式给予了模型渐进学习的能力,进一步增强了模型在低信噪比环境的去噪表现。

为了更直观地比较模型的语音增强效果,采用时频图可视化的方式展示了 - 10 dB 下的降噪结果,如图 4 所示。总体上,各种深度学习算法都在一定程度上对语音进行了有效的降噪处理。RCED 由于缺乏额外的过滤机制,其对于局部噪声细节的筛选过滤能力有限,在去除大量噪声的同时也去除了部分语音成分信息;GRN 相比于其他模型,高频分量存在过多的噪声残留;DARCN 作为基于注意力机制的最新模型,相比于其他架构在低信噪比环境下有更好的表现;本文模型由于利用形变卷积对局部信息处理的自适应性,有效拟合时频图中所显现的纹理信息,并配合基于注意力机制的动态选择机制,通过压缩后编码向量动态分配各支路权重,自适应的分配形变卷积和标准卷积的权重比例,抑制了形变卷积可能出现的负增益对语音增强所造成的影响。其对于局部噪声处理的能力明显优于其他几种算法,增强后的语音时频图也更接近于干净语谱图。

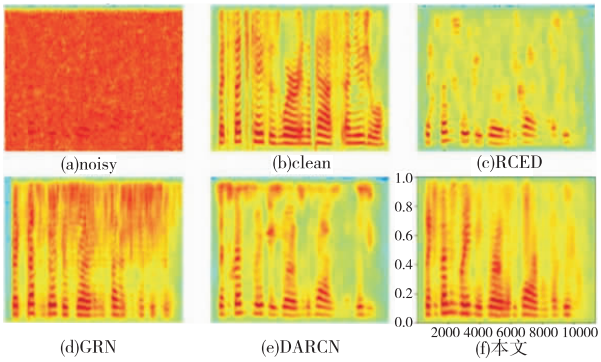

图 4　- 10 dB 条件下语音增强效果可视化

Fig. 4　Visualization under the - 10 dB condition

4.2　不同噪声环境下的消融实验分析

除此之外,为进一步探究模型各组件在语音增强过程的作用,本文在不同的信噪比条件下(- 10 dB, - 5 dB,0 dB,5 dB)进行了一系列消融实验,并在表 2 中展示了不同噪声环境下的模型性能。

从中能够得到以下观察结果:a)本文模型在绝大多数情况下都是性能最优的模型,这证明了该架构的整体优异性;b)形变卷积和动态选择机制所产生的贡献互相补充和增益,两者的结合促使了所有指标的进一步改进;c)相比于可见噪声(babble,destroyerEngine,destroyerOps,factory1),基于形变卷积的模型在面对未知噪声环境时(bucaneer1,factory2,white)性能增幅明显,基于形变卷积的解码器可以根 据输入动态的调整感受大小和形状,增强了模型应对复杂噪声场景的鲁棒性;d)动态选择机制的引入提升了模型整体的特征学习能力,在各种噪声环境下相比于无动态选择机制的框架在 STOI 指标上性能提升明显,但是在 PESQ 指标上无明显优势。

表 2　不同噪声环境下的消融实验结果

Tab. 2　Ablation study under different type of noises

指标	噪声	模型简述			
		无形变卷积	无动态选择	标准卷积	本文算法
STOI/%	(babble)	71.4	71.61	71.38	**72.16**
	(destroyerEngine)	81.09	81.32	80.88	**81.35**
	(destroyerOps)	78.77	79.03	78.81	**79.35**
	(factory1)	74.51	75.02	74.47	**75.25**
	(buccaneer1)	68.69	69.21	68.33	**69.49**
	(factory2)	81.75	82.7	81.39	**82.74**
	(white)	71.94	72.81	71.62	**73.01**
PESQ	(babble)	2.24	2.28	2.23	**2.28**
	(destroyerEngine)	2.56	**2.59**	2.58	**2.59**
	(destroyerOps)	2.53	**2.56**	2.53	**2.56**
	(factory1)	2.37	**2.40**	2.37	**2.40**
	(buccaneer1)	2.02	2.03	2.01	**2.10**
	(factory2)	2.67	**2.69**	2.64	**2.69**
	(white)	2.27	**2.28**	2.26	**2.28**

4.3　模型参数量分析

考虑到模型参数对于实验结果的影响,本文进一步分析了各个模型的参数总量,并同时比对了其性能,如图 5 所示。其

中从左到右分别为(RCED, RCNA, GRN, DARCN 和本文算法)。从图中可以看到,RCED 由于纯标准卷积架构,参数量最小,但其性能也有限。RCNA 模型尽管在单帧任务上表现良好,但当输入改为完整时频图时,基于局部的注意力机制性能反而有所下降。本文对 GRN 原有的网络架构进行了一定的调整,以在本文的任务上取得最优解,仍需要 1.06(百万)的参数。DARCN 由于构建了一个与主网络并行的注意力机制子网络,其参数量最大。本文模型在保持较小参数量的同时取得了最好的性能表现,这一定程度上可以体现本文模型的优异性。

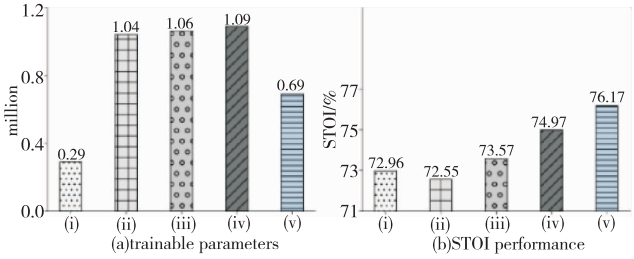

图 5 模型参数量及性能分析

Fig. 5 Model parameters and performance

5 结束语

为了提升复杂环境下神经网络模型的降噪性能,通过分析语音时频图构造方式发现语音纹理特征,继而将图像识别任务的可变形卷积迁移至语音领域,并提出了一种基于注意力机制的动态选择机制,以自适应地调整形变卷积和标准卷积的权值,有效缓解形变卷积的自适应形变所导致的负增益现象,在获取更强的信息处理能力的同时仍保留了模型的稳定性,增强了模型应对复杂噪声场景的去噪能力。除此之外,进一步利用渐进学习,在不增加模型复杂度的情况下变相增加模型深度,使得模型具备更强的学习能力。实验证明本文模型在 STOI、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,将尝试将本文模型架构应用于复数谱,以更好地适应复杂噪声环境。

参考文献:

[1] Shi Yanpei, Huang Qiang, Hain T. Robust speaker recognition using speech enhancement and attention model[EB/OL]. 2020. https://arxiv.org/abs/2001.05031.

[2] Zhang Xueliang, Wang Zhongqiu, Wang Deliang. A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2017: 276-280.

[3] 蓝天, 彭川, 李森, 等. 单声道语音降噪与去混响研究综述[J]. 计算机研究与发展, 2020, 57(5): 928-953. (Lan Tian, Peng Chuan, Li Sen, *et al*. A review of monaural speech noise reduction and dereverberation[J]. Computer Research and Development, 2020, 57(5): 928-953.)

[4] Reddy C, Shankar N, Bhat G, *et al*. An individualized super-Gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device[J]. IEEE Signal Processing Letters, 2017, 24(11): 1601-1605.

[5] Boll S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Trans on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113-120.

[6] Hu Xiaohu, Wang Shiwei, Zheng Chengshi, *et al*. A cepstrum-based preprocessing and postprocessing for speech enhancement in adverse environments[J]. Applied Acoustics, 2013, 74(12): 1458-1462.

[7] Jensen S, Hansen P, Hansen S, *et al*. Reduction of broad-band noise in speech by truncated QSVD[J]. IEEE Trans on Speech and Audio Processing, 1995, 3(6): 439-448.

[8] Loizou P. Speech enhancement: theory and practice[M]. Florida: CRC Press, 2013.

[9] 韩伟, 张雄伟, 周星宇, 等. 联合优化深度神经网络和约束维纳滤波的单通道语音增强方法[J]. 计算机应用研究, 2017, 34(3): 706-709, 713. (Han Wei, Zhang Xiongwei, Zhou Xingyu, *et al*. Single channel speech enhancement method based on combined optimization of deep neural network and constrained Wiener filter[J]. Application Research of Computers, 2017, 34(3): 706-709, 713.)

[10] Wang Yuxuan, Wang Deliang. Towards scaling up classification-based speech separation[J]. IEEE Trans on Audio, Speech, and Language Processing, 2013, 21(7): 1381-1390.

[11] Wang Yuxuan, Narayanan A, Wang Deliang. On training targets for supervised speech separation[J]. IEEE/ACM Trans on audio, Speech, and Language Processing, 2014, 22(12): 1849-1858.

[12] Xu Yong, Du Jun, Dai Lirong, *et al*. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Trans on Audio, Speech, and Language Processing, 2015, 23(1): 7-19.

[13] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C]//Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[14] Park S, Lee J. A fully convolutional neural network for speech enhancement[EB/OL]. 2016. https://arxiv.org/abs/1609.07132.

[15] Tan Ke, Wang Deliang. A convolutional recurrent neural network for real-time speech enhancement[C]//Proc of InterSpeech. Piscataway, NJ: IEEE Press, 2018: 3229-3233.

[16] Zhao Han, Zarar S, Tashev I, *et al*. Convolutional-recurrent neural networks for speech enhancement[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2018: 2401-2405.

[17] Tan Ke, Chen Jitong, Wang Deliang. Gated residual networks with dilated convolutions for monaural speech enhancement[J]. IEEE/ACM Trans on Audio, Speech, and Language Processing, 2018, 27(1): 189-198.

[18] Cheng Shuai, Zhang Haijian, Hua Guang. Speech enhancement in low SNR environments by designing a time-frequency binary mask[C]//Proc of the 23rd IEEE International Conference on Digital Signal Processing. Piscataway, NJ: IEEE Press, 2018: 1-5.

[19] Hubel D, Wiesel T. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of Physiology, 1962, 160(1): 106-154.

[20] Li Xiang, Wang Wenhai, Hu Xiaolin, *et al*. Selective kernel networks[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 510-519.

[21] Szegedy C. Going deeper with convolutions[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 1-9.

[22] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 7132-7141.

[23] Lan Tian, Lyu Yilan, Hui Guoqiang, *et al*. Redundant convolutional network with attention mechanism for monaural speech enhancement[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2020: 6654-6658.

[24] Li Andong, Zheng Chengshi, Fan Cunhang, *et al*. A recursive network with dynamic attention for monaural speech enhancement[EB/OL]. 2020. https://arxiv.org/abs/2003.12973.

[25] Li Xiang, Wang Wenhai, Hu Xiaolin, *et al*. Selective kernel networks[C]//Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 510-519.

[26] Dai Jifeng. Deformable convolutional networks[C]// Proc of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 764-773.

[27] Zhu Xizhou, Hu Han, Lin S, *et al*. Deformable convnets v2: more deformable, better results[C]//Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 9308-9316.

[28] Li Andong, Zheng Chengshi, Cheng Lijuan, *et al*. A time-domain monaural speech enhancement with recursive learning[EB/OL]. 2020. https://arxiv.org/abs/2003.09815.