计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220129
60(7): 1639−1648, 2023

基于跨维度协同注意力机制的单通道语音增强方法

康宏博[1] 冯雨佳[1] 台文鑫[1] 蓝 天[1] 吴祖峰[1] 刘 峤[2]
[1](电子科技大学信息与软件工程学院 成都 610054)
[2](电子科技大学计算机科学与工程学院 成都 611731)
(hbkang@std.uestc.edu.cn)

Monaural Speech Enhancement Based on Cross-Dimensional Collaborative Attention Mechanism

Kang Hongbo[1], Feng Yujia[1], Tai Wenxin[1], Lan Tian[1], Wu Zufeng[1], and Liu Qiao[2]
[1](*School of Information and Software Engineering*, *University of Electronic Science and Technology of China*, *Chengdu* 610054)
[2](*School of Computer Science and Engineering*, *University of Electronic Science and Technology of China*, *Chengdu* 611731)

Abstract Monaural speech enhancement aims to recover clean speech from complex noise scenes, thus improving the quality of the noise-corrupted voice signals. This problem has been studied for decades. In recent years, convolutional encoder-decoder neural networks have been widely used in speech enhancement tasks. The convolutional models reflect strong correlations of speech in time and can extract important voiceprint features. However, two challenges still remain. Firstly, skip connection mechanisms widely used in recent state-of-the-art methods introduce noise components in the transmission of feature information, which degrades the denoising performance inevitably; Secondly, widely used standard fix-shaped convolution kernels are inefficient of dealing with various voiceprints due to their limitation of receptive field. Taking into consideration the above concerns, we propose a novel end-to-end encoder-decoder-based network CADNet that incorporates the cross-dimensional collaborative attention mechanism and deformable convolution modules. In specific, we insert cross-dimensional collaborative attention blocks into skip connections to further facilitate the ability of voice information control. In addition, we introduce a deformable convolution layer after each standard convolution layer in order to better match the natural characteristics of voiceprints. Experiments conducted on the TIMIT open corpus verify the effectiveness of the proposed architecture in terms of objective intelligibility and quality metrics.

Key words speech enhancement; self-attention; cross-dimensional collaborative attention; deformable convolution; skip connection

摘 要 近年来,卷积神经网络在语音增强任务中得到了广泛的应用.然而,目前广泛使用的跳跃连接机制在特征信息传输时会引入噪声成分,从而不可避免地降低了去噪性能;除此之外,普遍使用的固定形状的卷积核在处理各种声纹信息时效率低下,基于上述考虑,提出了一种跨维度协同注意力机制和形变卷积模块的端到端编-解码器网络 CADNet. 具体来说,在跳跃连接中引入跨维度协同注意力模块,进一步提高

收稿日期: 2022-01-29; **修回日期**: 2022-08-08
基金项目: 国家自然科学基金项目(U19B2028); 国家科技重大专项 (2021YFC3330403); 中国电子科技集团 54 所开放课题 (201148); 攀钢集团有限公司开放课题 (211129)
This work was supported by the National Natural Science Foundation of China (U19B2028), the National Science and Technology Major Project (2021YFC3330403), the Open Subject of the 54th Research Institute of China Electronics Technology Group Corporation (201148), and the Open Subject of the Pangang Group Company Limited (211129).
通信作者: 蓝天(lantian1029@uestc.edu.cn)

信息控制能力.并且在每个标准卷积层之后引入形变卷积层,从而更好地匹配声纹的自然特征.在TIMIT公开数据集上进行的实验验证了所提出的方法在语音质量和可懂度的评价指标方面的有效性.

关键词 语音增强;自注意力;跨维度协同注意力;形变卷积;跳跃连接

中图法分类号 TP391

日常生活中,人们经常会使用到移动电话和微信聊天等,在这些语音通信中,环境噪声和其他干扰不可避免地影响了通话质量.因此,降低背景噪声,提高语音的质量和清晰度一直都是语音处理应用中的一个关键问题.语音增强的目的就是消除噪声和干扰,最大可能地提高语音的听觉质量和可懂度.目前很多的语音增强算法仅仅改善了语音质量,在可懂度方面存在很大不足,即在低信噪比环境下,噪声得到了控制,但引入了较大的语音失真.因此,如何保证在语音不失真的前提下,有效抑制噪声和干扰,是语音增强领域一个主要挑战.

语音增强方法主要有传统方法和深度学习方法.传统方法如谱减法[1]、维纳滤波[2]和最小均方误差(minimum mean square error, MMSE)估计[3]在语音处理领域得到了广泛地研究,其主要采用无监督数字信号分析方法,通过对语音信号进行分解,确定干净语音和噪声的特征,实现语音和噪声的分离.然而,这些方法都是基于稳定噪声的假设,在处理非平稳噪声时性能会大大降低,为了解决这些局限性,基于深度学习的方法被提出,深度学习方法主要有基于时频掩蔽的语音增强方法和基于特征映射的语音增强方法.基于时频掩蔽的语音增强方法主要将噪声和干净语音相互关系的时频掩蔽作为学习目标,该方法需要假设纯净语音和噪声具有一定的独立性;而基于特征映射的语音增强方法主要学习干净语音特征和带噪语音特征之间的复杂关系,从而找到两者间的映射,网络的输入与输出通常是同种类型的声学特征,且在实现过程中几乎不会对语音和噪声信号做任何假设.研究人员在此基础上提出了许多基于神经网络的模型[4-6],如深度神经网络(deep neural network, DNN)[7-9]、循环神经网络(recurrent neural network, RNN)[10-12]、卷积神经网络(convolutional neural network, CNN)[13-17]和其他一些变体.Xu等人[8]首先使用了一种基于DNN的回归方法,学习噪声语音到干净语音的对数功率谱的映射函数.该方法取得了令人满意的结果,从而证明了基于深度学习方法的有效性.然而,DNN由几个完全连通的层组成,这些层对语音信号[18]的时域结构建模困难.此外,参数的数量随着层数和节点数的增加而迅速增加,从而增加了计算量.

近年来CNN被不断应用于语音处理领域,使得其在减少参数量的同时,可以显著捕捉语音信号中的隐含信息.在一定范围内,CNN可以保持语音特征在频域内的小幅度移动,从而应对说话人和环境的变化.为了提高去噪性能和重构语音,很多方法会利用跳跃连接帮助模型恢复频谱.然而,目前广泛使用的跳跃连接机制在特征信息传输时会引入噪声成分,不可避免地降低了去噪性能;除此之外,CNN中普遍使用的固定形状的卷积核在处理各种声纹信息时效率低下.基于上述考虑,本文提出了一种基于跨维度协同注意力机制(cross-dimensional collaborative attention mechanism)和形变卷积(deformable convolution)的端到端编-解码器网络CADNet.具体地:1)在跳跃连接中引入卷积注意模块,跳跃连接自适应分配注意权值来抑制噪声成分.不同于单一方面的注意力机制设计,我们从通道和空间2个维度及其相互依赖等方面对其进行了细化,进一步提高信息控制能力,从而有效解决噪声引入的问题.2)在编码器和解码器之间,将多个自注意力模块串联起来对信息进行处理,为了防止注意力操作堆叠造成的信息遗忘,在每个自注意力模块(self-attention block, SAB)中添加残差连接,将每个原始输入直接传递到下一层.3)在每个标准卷积层之后引入可形变卷积层,以更好地匹配声纹的自然特征,增强信息的解析能力.

本文的主要贡献包括3个方面:

1)提出了一种基于通道和空间的跨维度协同注意力机制方法,通过在跳跃连接中协同学习注意力机制来抑制噪声成分,进一步提高信息控制能力,且该方法只涉及很少的参数;

2)在解码器的每个标准卷积层后引入形变卷积层,对解析结果进行重新校准和校正,以获得更好的特征处理能力;

3)在TIMIT公开数据集上对多个基准模型进行了充分实验,并与本文所提出的模型进行了比较分析,实验验证了所提出的算法在语音质量和可懂度的评价指标方面的有效性.

1 相关工作

1.1 基于 CNN 的语音增强

随着 CNN 被引入到语音处理中，Fu 等人[14]提出了一种信噪比感知的 CNN 来估计话语的信噪比，然后自适应增强，从而提高泛化能力. Hou 等人[16]使用了音频和视觉信息来进行语音增强. Bhat 等人[17]提出了一个多目标学习 CNN，并将其作为应用程序在智能手机上实现. 由于 CNN 在特征提取方面更为有效，所以近年来的研究多采用基于卷积编码器-解码器（convolutional encoder-decoder）的网络进行语音增强. 此外，Park 等人[19]去掉了 CNN 中的全连接层，将全卷积网络（fully convolutional network，FCN）引入语音增强领域. 近年来，基于 FCN 的工作层出不穷. Tan 等人[20]提出了卷积循环网络（convolutional recurrent network，CRN），它在 FCN 的编码器和解码器之间插入了 2 个长短期记忆网络（long-short-memory，LSTM）层，可以有效地捕获局部和序列属性. Grzywalski 等人[21]将门控循环单元（gate recurrent unit，GRU）层添加到 FCN 的每个构建块中. 文献 [14, 16, 17, 19, 20, 21] 所述的模型利用神经网络的时间建模能力，提高了自身的信息表示能力. FCN 中的最大池化层用于提取特定区域中最活跃的部分，但会导致细节信息的丢失. 因此，FCN 在一些只需要获得整体特征的语音识别领域中就可以取得良好的效果. 但在语音增强领域中，细节信息是恢复干净语音的关键，如果没有细节信息，将会大大影响语音增强的效果. 为了解决这个问题，文献 [19] 中还提出了冗余卷积编码器-解码器（redundant convolutional encoder-decoder，RCED）网络，该网络丢弃了 FCN 中的最大池化层和相应的上采样层，保持特征图的大小，从而达到保留细节信息的目的，提高了语音增强的性能. 为了进一步提高去噪性能，Lan 等人[22]在每个卷积层之后引入注意力块来捕获复杂的依赖关系. 此外，考虑到深度神经网络通常缺乏细粒度的信息，使得语音重构更加困难，文献 [22] 所述的方法还利用了跳跃连接帮助模型恢复频谱.

1.2 基于注意力机制的神经网络

注意力机制最初被用于机器翻译工作[23]，现在已经成为了神经网络领域中的一个重要概念. 在人工智能领域，注意力机制已经作为神经网络结构的重要组成部分，并在自然语言处理、语音和计算机等领域得到了广泛地应用. 注意力机制来源于人类视觉机制，通常人类的视觉系统倾向于关注图像中重要的信息和忽略掉不相关的信息[24]. 而注意力的基本思想是：在预测输出时允许模型动态地关注有利于执行任务输入的部分，并将这种相关性概念结合起来. 注意力机制不仅能够告诉模型应该注意什么，同时也能增强特定区域的表征能力. 传统的注意力方法是将网络的特征图的权重传递到下一层，而 Hu 等人[25]提出了一种注意力机制，其学习了每一个卷积模块的通道注意力，提升了 CNN 的性能. 其核心思想在于建模通道之间的相互依赖关系，通过网络的全局损失函数自适应地重新矫正通道之间的权重.

在此基础上，本文首先在通道维度使用了注意力机制，依赖于特征图的全局信息来确定每个通道的重要性. 但由于卷积操作是将跨通道信息和空间信息融合在一起来提取信息特征，因此从通道级全局视角进行细化、确定每个通道的重要性之后，我们还利用特征的空间关系生成空间注意图，以区分不同权重的内空间关系特征，把 2 个分支的信息融合在一起，从而形成了一种跨维度的协同注意力机制.

2 基于跨维度协同注意力机制和形变卷积的神经网络

本节介绍基于跨维度协同注意力机制和形变卷积的单通道语音增强算法及应用.

2.1 问题描述

通常情况下，一条带噪语音可以表示为

$$\boldsymbol{Y}(t)=\boldsymbol{X}(t)+\boldsymbol{N}(t), \tag{1}$$

其中t表示时间帧的索引，$\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{N}$表示对应时间帧的含噪语音信号、纯净语音信号及噪声波形. 从含噪语音信号$\boldsymbol{Y}$中消除噪声$\boldsymbol{N}$，得到纯净语音$\boldsymbol{X}$的过程就是语音增强的任务. 一般情况下，不同语音往往具有不同的时间长度，因为语音的时间帧的总数不是固定的. 语音的时域信号通过分帧、加窗以及短时傅里叶变换得到短时傅里叶变换幅度谱. 给定一个长度为L的实值向量$\boldsymbol{Y}$，可通过短时傅里叶变换（short-time Fourier transform，STFT）将其转换为时频域，即

$$\boldsymbol{Y}\rightarrow\boldsymbol{Y}_{t,f}, \tag{2}$$

式（1）可以重写为

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

其中$\boldsymbol{Y}_{t,f}$, $\boldsymbol{X}_{t,f}$, $\boldsymbol{N}_{t,f}$分别代表含噪语音、纯净语音和噪声在时间帧t和频点f时的值. 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.

2.2 网络结构

如图 1 所示, CADNet 的整体框架图一共包含了 4 个模块, 分别为: 编码层(encoder)、自注意力模块(self-attention blocks, SAB)、解码层(decoder)和跨维度协同注意力模块(cross-dimensional collaborative attention blocks).

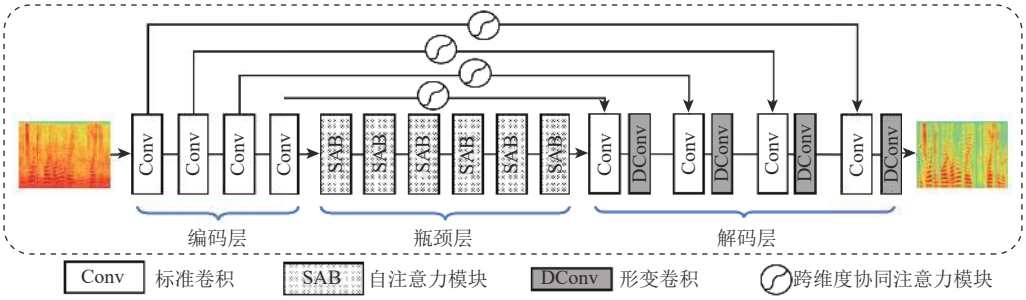


Fig. 1 The overall architecture of CADNet

图 1 CADNet 的总体架构

1) 编码层

在特征处理和语音重构的后续步骤中, 充分的特征表示将会起到很重要的作用. 考虑到特定的帧长和时间轴与频率轴之间的分辨率关系, 我们决定采用一个更大的卷积核来对特征进行提取, 从而在层数较少的情况下能够获得更大的感受野[26]. 首先, 利用短时傅立叶变换将语音波形转换为频谱域, 并以幅度谱作为输入. 然后, 使用由 4 个大小为 11×11 的卷积核组成的卷积层进行特征提取, 并在每个卷积层之后使用 PReLU 激活函数. 除此之外, 我们将输出特征通道数分别设置为 4, 8, 16, 32, 并将步长和膨胀率值均设置为 1, 以避免信息丢失.

2) 自注意力模块

受到文献 [27] 的启发, 将 6 个 SAB 串联起来对信息进行处理.每个 SAB 的第 1 步是使用 1 层卷积(卷积核大小为 11×11)进行线性变换.然后, 将特征输入到 2 个并行的卷积层中(卷积核大小均为 11×11), 其中一个卷积层生成重要权值来控制另一个卷积层的信息流. 使用注意力机制的本意是更好地获取全局信息, 但堆叠的注意力层也会使模型丧失了捕捉局部特征的能力, 使得特征图包含的图像信息会逐层减少, 造成一定的信息遗忘.为了防止注意力操作堆叠造成的信息遗忘, 还在每个 SAB 中添加了残差连接[28], 将每个原始的输入直接传递到下一层. 在具体的实现过程中, 将每个 SAB 的输入和输出通道的数量都固定为 32 个, 并在每个卷积层之后应用 PReLU 激活函数.

3) 解码层

标准卷积使用固定的卷积核在输入的特征图上采样时, 将卷积核与特征图中对应位置的数值逐个相乘, 最后求出加权和, 就得到该位置的卷积结果, 不断移动卷积核, 就可算出各个位置的卷积结果. 在同一层标准卷积中, 所有的激活单元的感受野是一样的. 但由于不同位置可能对应着不同尺度或形变的特征信息, 因此在提取特征信息的过程中对尺度或者感受野大小进行自适应地调整是必须的. 受到文献 [29] 的启发, 我们在解码层的卷积模块中引入了形变卷积, 即在每个标准的卷积层之后添加了形变卷积层来重新校正解析结果. 形变卷积基于一个平行网络学习偏移量, 对卷积核中每个采样点的位置都增加了一个偏移变量, 可以实现在当前位置附近随意采样而不局限于之前的规则格点.

设 $R=\{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$表示以 p_0 为中心的 3×3 邻域区域, 则位置 p_0处的偏移增广特征映射为

$$h(p_0)=\sum_{p_n\in R} w(p_n)\cdot u(p_0+p_n+\Delta p_n). \tag{4}$$

我们可以看到, 形变卷积可以通过操纵偏移 p_n 来改变感受野的大小, 从而更好地匹配声纹的自然特征. 在具体实现中, 我们在每个核大小为 11×11 的标准卷积之后应用核大小为 3×3 的形变卷积. 解码器的 8 层输入、输出通道数目分别为(32, 32, 16, 16, 8, 8, 4, 4)和(16, 16, 8, 8, 4, 4, 1, 1). 形变卷积的结构如图 2 所示.

4) 跨维度协同注意力模块

为了区分不同类型的信息, 抑制不相关的噪声部分, 我们在跳跃连接操作中插入一个跨维度协同注意力模块[24], 如图 3 所示. 由于卷积是将跨通道信息和空间信息融合在一起来提取信息特征的, 因此我们沿着这 2 个主要维度设计了跨维度协同注意力

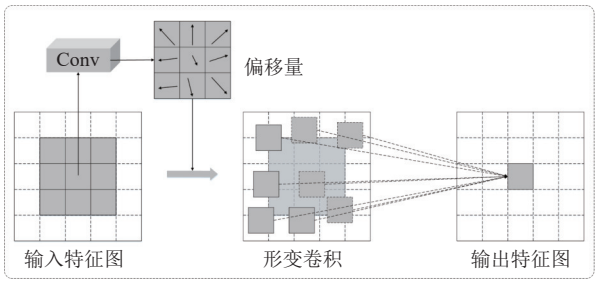


Fig. 2 Details of deformable convolution

图 2 形变卷积细节

机制. 首先, 我们采用一种基于通道的注意力机制, 该机制依赖于特征图的全局信息来确定每个通道的重要性.

设一个卷积层的输出为 $\boldsymbol{X}\in\mathbb{R}^{W\times H\times C}$, 其中 W, H, C 分别表示宽度、高度和通道维度. 我们用 $\boldsymbol{X}_{\rm e}$ 和 $\boldsymbol{X}_{\rm d}$ 分别表示编码器和解码器的一个卷积块的输出. 首先, 我们通过元素求和的方式融合 $\boldsymbol{X}_{\rm e}$ 和 $\boldsymbol{X}_{\rm d}$, 即

$$\boldsymbol{X}=\boldsymbol{X}_{\rm e}+\boldsymbol{X}_{\rm d}. \tag{5}$$

其次, 采用全局平均池化方法, 沿通道维度对信息进行挤压, 即聚合特征

$$\boldsymbol{y}=g(\boldsymbol{X})=\frac{1}{W\times H}\sum_{i=1,\ j=1}^{W,\ H}X_{ij}. \tag{6}$$

由于注意力机制的有效性, Lan 等人[22] 使用了 SE-Net, SE-Net 学习了每一个卷积模块的通道注意力从而提升了网络的性能, 但是 SE-Net 中的 2 个全连接层之间的维度缩减不利于学习通道注意力的权重, 且文献 [30] 表明降维会对学习通道注意力产生副作用, 效率较低, 没有必要捕获所有通道之间的依赖关系. 适当的跨通道交互对于保持性能和显著降低模型复杂度很有价值. 因此我们提出了一种获取局部跨通道交互的方法, 旨在保证效率和有效性. 假设经过全局平均池化后的聚合特征为 $\boldsymbol{y}\in\mathbb{R}^C$ 且没有降维, 其中 C 为特征维度, 即过滤器的数量. $\sigma(\cdot)$ 表示 sigmoid 函数, 则可以通过式(7)来学习通道注意力 $\boldsymbol{\omega}$:

$$\boldsymbol{\omega}=\sigma(\boldsymbol{W}_y\boldsymbol{y}), \tag{7}$$

其中 $\boldsymbol{W}_y$ 是一个 $C\times C$ 的矩阵, C 表示通道维度.

我们使用一个带状矩阵 $\boldsymbol{W}_k$ 来学习通道注意力:

$$\boldsymbol{W}_k=\begin{pmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & 0 & \cdots & 0\\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & 0 & \cdots & 0\\ \vdots & \vdots & & \vdots & \ddots & \vdots & & \vdots\\ 0 & 0 & \cdots & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{pmatrix}, \tag{8}$$

其中 $\boldsymbol{W}_k$ 包含 $k\times C$ 个参数.

我们可以仅考虑 y_i 和它相邻的 k 个通道之间的相互关系来计算 y_i 的权值, 让所有的通道共享相同的学习参数, 即

$$\omega_i=\sigma\left(\sum_{i=1}^{k}\omega^j y_i^j\right),\ y_i^j\in\Omega_i^k, \tag{9}$$

其中 Ω 表示 y_i 的 k 个相邻通道的集合, 并且该种方法可以使用卷积核大小为 k 的快速一维卷积来实现:

$$\boldsymbol{\omega}=\sigma(\mathrm{C1D}_k(\boldsymbol{y})). \tag{10}$$

其中C1D指一维卷积, 且该一维卷积仅包含 k 个参数就可以获取局部跨通道交互的信息.

由于该模块旨在适当地捕获跨通道信息交互, 因此需要确定交互的覆盖范围(即一维卷积核的大小 k), 我们可以通过手动调整具有不同通道数的卷积块来确定跨通道交互信息的范围, 但是手动调优会消耗大量计算机资源, 组卷积已经成功运用于改进 CNN 架构[31-33], 其中高维通道拥有较长的固定组卷积, 而低维通道拥有较短的固定组卷积. 因此, 我们可以得到通道交互的覆盖范围(即一维卷积的核大小 k)与通道维度 C 成正比. 也就是说, k 和 C 之间存在映射 ϕ:

$$C=\phi(k). \tag{11}$$

最简单的映射是一个线性函数, 即

$$\phi(k)=\gamma\times k-b, \tag{12}$$

其中, γ 和 b 是线性函数的参数. 然而, 以线性为特征的关系功能太有限了.另一方面, 通道维度 C 通常设置为 2 的幂因子, 我们可以通过此关系将线性函数(式(12))扩展到非线性函数, 即

$$C=\phi(k)=2^{\gamma\times k-b}. \tag{13}$$

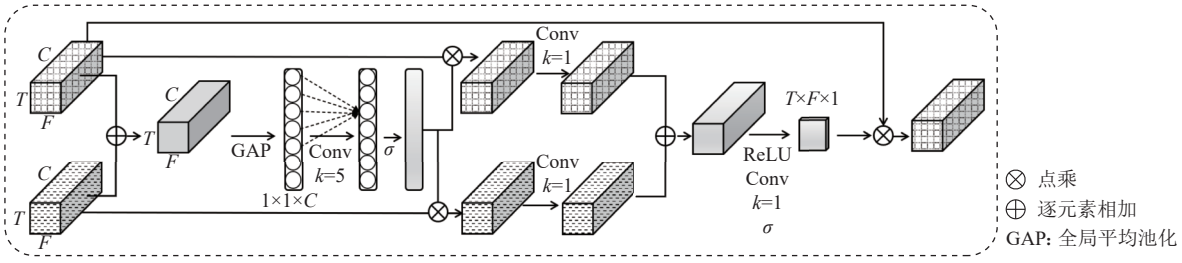


Fig. 3 Cross-dimensional collaborative attention mechanism blocks

图 3 跨维度协同注意力机制模块

那么，给定通道维度 C，内核大小 k 可以是自适应地确定：

$$k=\psi(C)=\left|\frac{\mathrm{lb}(C)}{\gamma}+\frac{b}{\gamma}\right|_{\mathrm{odd}}, \tag{14}$$

其中$|t|_{\mathrm{odd}}$表示最接近 t 的奇数.

本文中，我们把所有实验中的 y 和 b 分别设置为 2 和 1. 显然，通过映射函数$\psi(\cdot)$，高维通道具有更远距离的相互作用. 在自适应地确定了内核大小 k 之后，执行一维卷积，然后，使用 sigmoid 激活函数来标准化每个权值，通过乘法求出权值来细化$\boldsymbol{X}_{\mathrm{e}}$和$\boldsymbol{X}_{\mathrm{d}}$. 从通道级全局视角进行细化后，利用特征的空间关系生成空间注意图，以区分不同权重的内空间关系特征. 首先融合 2 个分支的信息，然后应用一系列的操作（ReLU 激活函数，1×1的卷积操作和 sigmoid 激活函数）来获得空间注意图. 特别地，将卷积层的输出通道维度 C 设置为 1，以通过通道尺寸缩小注意图（参见图 3）.

3 实　　验

本节详细介绍实验的数据集及其实验配置，以及基准模型和不同模型的实验结果对比分析.

3.1 数据集及其实验配置

本文在 TIMIT 语料库[34]上进行了实验，同时选取 NOISEX92[35]中的噪声以及其他 105 种噪声作为实验中的噪声数据集，并从 TIMIT 语料库的训练数据集中分别选取 2 000 和 100 条干净语音进行训练和验证，并选择 TIMIT 核心测试中的 192 条语音进行测试.

训练过程中我们采用了 105 种噪声，其中包括 100 种非言语噪音[29]和自助餐厅、餐厅、公园、办公室和会议这 5 种不同环境的生活噪音[36]. 在−5~10 dB 的信噪比（signal noise rate, SNR）范围内以 1 dB 为间隔混合训练集的共 2 100 条干净语音分别创建了训练和验证数据集. 测试过程中，我们在 105 种噪声的基础上还增加了 NOISEX92 中的 5 种噪音（babble, f16, factory2, m109, white），并在 4 种信噪比（−5 dB, 0 dB, 5 dB, 10 dB）条件下混合 192 条干净测试语音创建测试数据集.

在混合语音的过程中，我们首先把所有的噪声拼接成一个长向量；然后随机选择一个切割点，在一定的信噪比条件下将其与待混合的干净语音进行混合.如表 1 所示，混合语言分别生成了 36 h 的训练数据、1.5 h 的验证数据和 1 h 的测试数据.

我们将实验中所使用到的所有语音片段均采样到 16 kHz. 在数据预处理过程中，使用窗口大小为 20 ms 的汉明窗对语音信号进行短时傅里叶变换，将语音信号分割成一组帧，相邻帧重叠 50%. 每帧对应一个 161 维的特征向量. 我们将 epoch 大小设置为 30，并将学习率固定为 0.000 2，使用平均绝对误差（mean absolute error, MAE）作为损失函数，并通过 Adam 优化器来优化模型参数. 我们通过短时客观可解性（STOI）[37]、感知评价语音质量（PESQ）[38]和比例不变信号失真比（SI-SDR）[39]来评估不同模型的性能.对于这 3 个评价指标，我们分别计算了每一个基准模型在不同信噪比下的评价指标的平均值，避免了实验结果的偶然性.本文实验所用到的实验平台是 Ubuntu LTS 18.04，其带有 i7-9700 和 RTX 2060.

Table 1 Dataset of Our Experiment

表 1 实验数据集

语音种类	数据集分类	数量	时长/h
干净语音	训练	2 000	1.2
	验证	100	0.06
	测试	192	0.12
噪声	可见	105	0.17
	不可见	5	0.4
混合语音	训练	20 000	36
	验证	2 000	1.5
	测试	768	1

3.2 对比方法

我们在近 3 年较为流行的几个时频域网络中选取了 4 个模型（CRN[20], GRN[40], AUNet[41], DARCN[42]），与本文所提出的方法 CADNet 进行了对比，验证了 CADNet 方法的有效性.

1）CRN 将卷积编码器-解码器网络和长短期记忆网络结合到 CRN 体系结构中，并引入了跳跃连接操作，从而形成了一个适用于实时处理的单通道语音增强系统；

2）GRN 将语音增强视为序列到序列的映射，结合了扩展卷积和门控机制进行序列建模，在扩大感受野的同时使用简单的聚合操作来保留重要特征，从而实现语音增强；

3）AUNet 在编码器和解码器之间采用了一种空间上的注意机制，用于学习聚焦不同形状和大小的目标结构，以尽可能地消除噪声，从而提升语音增强的准确率；

4）DARCN 设计了一个并行注意子网络来控制信息流，同时引入动态注意力机制和递归学习，通过

在多个阶段重用网络来动态减少可训练参数的数量，增强单通道语音.

3.3 实验结果

表 2 分别显示了不同模型的性能，其中 Noisy 表示带噪语言的语音质量. 我们可以观察到：在不同的信噪比条件下，CRN 的 3 项指标结果都优于带噪语音 Noisy，但在低信噪比水平上（−5dB）性能会严重下降；采用了空间注意力机制的 AUNet 获得了比 CRN 更好的结果；DARCN 利用注意力机制动态控制信息流，即使在低信噪比的情况下也能取得良好的效果，而在低信噪比的情况下，跳跃连接引入的噪声更加明显，从而降低了去噪性能，这与之前的分析一致；没有采用跳跃连接的 GRN 即使性能指标不是最佳，但也获得了相对较好的结果，这可能是由于其简单的聚合操作，以一种粗略的方式提供所有流程信息. 与目前最先进的模型相比，本文所提出的 CADNet 性能提升显著，即在 STOI, PESQ, SI-SDR 指标方面，CADNet 的性能分别提高了 0.70%, 3.83%, 2.61%.

Table 2 Indicators Comparison of Different Models Under Different SNR Conditions

表 2 不同信噪比条件下各个模型的指标比较

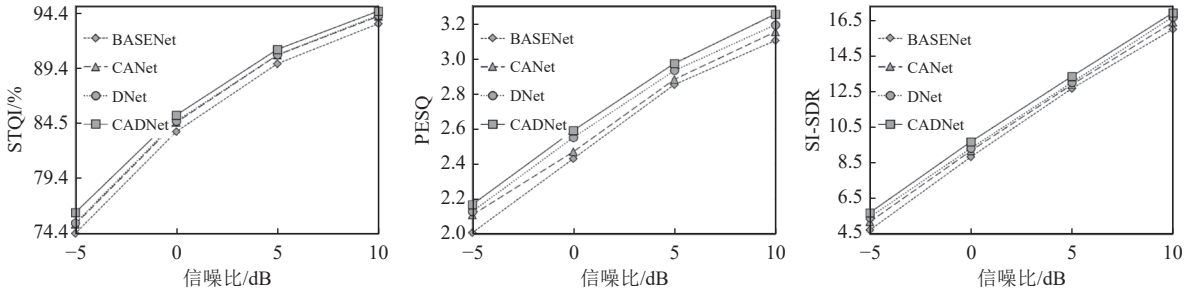
信噪比/dB	指标	Noisy	模型				
			CRN	GRN	AUNet	DARCN	CADNet
−5	STOI/%	60.82	68.23	74.13	72.29	75.35	**76.35**
0		71.89	80.03	83.71	82.34	84.59	**85.14**
5		81.83	88.08	90.11	89.02	90.58	**91.02**
10		89.09	92.59	93.76	92.92	94.05	**94.46**
	平均/%	75.91	82.22	85.42	84.13	86.14	**86.74**
−5	PESQ	1.43	1.78	2.04	1.97	2.09	**2.17**
0		1.76	2.22	2.49	2.41	2.53	**2.59**
5		2.15	2.64	2.89	2.80	2.92	**2.97**
10		2.47	2.94	3.19	3.10	3.19	**3.25**
	平均	1.95	2.39	2.65	2.57	2.68	**2.75**
−5	SI-SDR	-5.00	3.39	4.92	4.62	5.34	**5.72**
0		0.00	7.75	8.99	8.76	9.26	**9.67**
5		4.99	11.78	12.78	12.51	12.93	**13.33**
10		10.00	15.43	16.34	15.94	16.42	**16.88**
	平均	2.49	9.59	10.75	10.45	10.98	**11.40**

注：黑体数字表示每种实验设置下的最佳结果.

为了进一步研究模型中单个模块对性能的影响，我们进行了消融实验，分别构建了去掉形变卷积模块的 CANet 网络、去掉协同注意力模块的 DNet 网络以及同时去掉这 2 个模块的基本网络. BASENet 分别比较了这 3 个模型和 CADNet 在不同信噪比下（−5dB, 0dB, 5dB, 10dB）的 STOI, PESQ, SI-SDR 值，实验结果如图 4 所示. 跨维度协同注意力模块和形变卷积模块均能够较好地提升模型 BASENet 的性能指标，从而改善实验结果. 更为重要的是，两者的贡献是相辅相成的，因此消融实验结果也表明了跨维度协同注意力和形变卷积这 2 种机制的重要性.

通道注意力模块中包含一个参数 k，即卷积核大小，在第 2 节我们提到卷积核大小 k 和通道维度具有某种关系并且给出了推导，因此，我们做了参数敏感性实验来证明我们的结论. 通过设置 k=3, 5, 7, 9，我们分别对模型进行了训练，然后在混合信噪比语音测试集上进行了测试. 结果如表 3 所示，可以观察到当 k=5 时，模型可以获得最优的结果.

为了证明 CADNet 的效率，我们还对模型复杂度进行了分析. 表 4 显示了不同模型的训练参数量和测试每个样本（即每条语音）的测试时长. 可以看到，与其他模型相比，CADNet 的参数量相对较少. 这是由于在模型设计过程中使用了小通道搭配较大的卷积核，从而降低了模型的参数复杂度，达到令人满意的性能. 因此，CADNet 由于内存消耗少且推理时间快，具有广泛的应用性.



(a) 各模型STOI值折线图 (b) 各模型PESQ值折线图 (c) 各模型SI-SDR值折线图

Fig. 4 Experimental results of ablation studies under different SNRs

图 4 不同信噪比下的消融实验结果

Table 3 Experimental results of Channel-Based Attention Modules with Various *k* values

表 3 不同 *k* 值对应的通道注意力模块的实验结果

k	PESQ	STOI/%	SI-SDR
3	2.74	86.61	11.36
5	**2.75**	**86.73**	**11.40**
7	2.73	86.51	11.17
9	2.75	86.60	11.30

注：黑体数字表示最优值.

Table 4 Number of Training Parameters and Inference Time (for Each Sample) of Different Models

表 4 不同模型的训练参数量和测试时间（每个样本）

模型	训练参数量	测试时间/s
CRN	1 719 万	0.15
GRN	313 万	**0.11**
DARCN	123 万	0.17
CADNet	**72 万**	0.17

注：黑体数字表示最优值.

4 结 论

本文首先剖析了近年来基于 CNN 的语音增强模型，并提出了其所面临的挑战. 然后，提出了一种新的 CADNet 方法，该方法包含 2 个具体的设计：1）在编码器和解码器之间引入了一个跨维度协同注意力模块，首先依赖于特征图的全局信息来确定每个通道的重要性，从通道级全局视角进行细化后；再利用特征的空间生成空间注意力图，以区分不同权重的内空间关系特征；最后将 2 个维度的信息进行融合，从而更好地控制信息，抑制不相关的噪声部分. 2）在解码器部分引入形变卷积模块，对尺度或者感受野大小进行自适应地调整，提取不同位置对应的不同尺度或形变的特征信息，更好地匹配声纹的自然特征，从而增强了信息解析能力. 为了进一步验证模型的有效性和效率，我们在 TIMIT 语料库上进行了一系列实验. 实验结果表明，所提出的模型能够在降低参数量的同时，性能始终优于现有的先进模型. 由于低级的粒度特征和高级语义表示之间的直接连接会削弱自注意力模块的重要性，我们未来的工作将会重点研究如何在上述条件下提高自注力模块的重要性.

作者贡献声明：康宏博和冯雨佳对本文具有相同的贡献. 其中康宏博提出了算法思路和实验方案并撰写论文；冯雨佳负责完成实验并撰写论文. 台文鑫提出论文修改意见；蓝天、吴祖峰和刘峤提出了指导意见.

参 考 文 献

[1] Boll S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1979, 27(2): 113−120

[2] 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

[3] Gerkmann T, Hendriks R C. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay[J]. IEEE Transactions on Audio Speech & Language Processing, 2012, 20(4): 1383−1393

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

[5] Sun Lei, Du Jun, Dai Lirong, et al. Multiple-target deep learning for LSTM-RNN based speech enhancement [C] //Proc of the 5th Hands-free Speech Communications and Microphone Arrays. Piscataway, NJ: IEEE, 2017: 136 − 140

[6] Pandey A, Wang Deliang. A new framework for CNN based speech enhancement in the time domain[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179−1188

[7] Kumar A, Florencio D. Speech enhancement in multiple-noise conditions using deep neural networks[J]. arXiv: preprint, arXiv: 1605. 02427, 2016

[8] Xu Yong, Du Jun, Huang Zhen, et al. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement[J]. arXiv: preprint, arXiv: 1703. 07172, 2017

[9] Kang T G, Shin J W, Kim N S. DNN-based monaural speech enhancement with temporal and spectral variations equalization[J]. Digital Signal Processing, 2017, 74: 102−110

[10] Weninger F, Erdogan H, Watanabe S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR [C] //Proc of the 12th Int Conf on Latent Variable Analysis and Signal Separation. Berlin: Springer, 2015: 91−99

[11] Shan Dongjing, Zhang Xiongwei, Zhang Chao, et al. A novel encoder-decoder model via NS-LSTM used for bone-conducted speech enhancement[J]. IEEE Access, 2018, 6: 62638−62644

[12] Gao Tian, Du Jun, Dai Lirong, et al. Densely connected progressive learning for LSTM-based speech enhancement [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5054−5058

[13] Healy E W, Yoho S E, Wang Yuxuan, et al. An algorithm to improve speech recognition in noise for hearing-impaired listeners[J]. The Journal of the Acoustical Society of America, 2013, 134(4): 3029−3038

[14] Fu Szu-Wei, Tsao Y, Lu Xugang. SNR-aware convolutional neural network modeling for speech enhancement [C] //Proc of the 17th Interspeech 2016. Grenoble, France: ISCA, 2016: 3768−3772

[15] Kounovsky T, Malek J. Single channel speech enhancement using convolutional neural network [C] //Proc of the 13th IEEE Int Workshop of Electronics, Control, Measurement, Signals and Their Application to Mechatronics(ECMSM). Piscataway, NJ: IEEE, 2017: 1−5

[16] Hou Jen-Cheng, Wang Syu-Siang, Lai Ying-Hui, et al. Audio-visual speech enhancement using multimodal deep convolutional neural networks[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, 2(2): 117−128

[17] Bhat G S, Shankar N, Reddy C, et al. A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone[J]. IEEE Access, 2019, 7: 78421−78433

[18] Fu Szu-Wei, Tsao Yu, Lu Xugang, et al. Raw waveform-based speech enhancement by fully convolutional networks [C] //Proc of the 8th Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. Piscataway, NJ: IEEE, 2017: 6−12

[19] Park S R, Lee J W. A fully convolutional neural network for speech enhancement [C] //Proc of the 18th Interspeech 2017. Grenoble, France: ISCA, 2017: 1993−1997

[20] Tan Ke, Wang Deliang. A convolutional recurrent neural network for real-time speech enhancement [C] //Proc of the 19th Interspeech 2018. Grenoble, France: ISCA, 2018: 3229−3233

[21] Grzywalski T, Drgas S. Application of recurrent U-net architecture to speech enhancement [C] //Proc of the 22nd Signal Processing: Algorithms, Architectures, Arrangements, and Applications. Piscataway, NJ: IEEE, 2018: 82 − 87

[22] Lan Tian, Lyu Yilan, Ye Wenzheng, et al. Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement[J]. IEEE Access, 2020, 8: 78979−78991

[23] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint, arXiv: 1409. 0473, 2014

[24] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [C] //Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2048−2057

[25] Hu Jie, Li Shen, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011−2023

[26] Ding Xiaohan, Zhang Xiangyu, Zhou Yizhuang, et al. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs[J]. arXiv preprint, arXiv: 2203. 06717, 2022

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Proc of the 31st Int Conf on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2017: 6000−6010

[28] Wang Fei, Jiang Mengqing, Qian Chen, et al. Residual attention network for image classification [C] //Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6450−6458

[29] Dai Jifeng, Qi Haozhi, Xiong Yuwen, et al. Deformable convolutional networks [C] //Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 764−773

[30] Wang Qilong, Wu Banggu, Li Pengfei, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C/OL] // Proc of the 38th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020[2022-06-16]. https://ieeexplore.ieee.org/document/9156697

[31] Ioannou Y, Robertson D, Cipolla R, et al. Deep roots: Improving CNN efficiency with hierarchical filter groups [C] //Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5977−5986

[32] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks [C] //Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5987−5995

[33] Zhang Ting, Qi Guojun, Xiao Bin, et al. Interleaved group convolutions [C] //Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 4383−4392

[34] Garofolo J S, Lamel L F, Fisher W M, et al. TIMIT acoustic-phonetic continuous speech corpus [DB/OL]. Philadelphia, PA: Linguistic

Data Consortium. 1993[2022-06-16]. https://catalog.ldc.upenn.edu/LDC93S1

[35] Varga A, Steeneken H. 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

[36] Thiemann J, Ito N, Vincent E. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings[J]. Acoustical Society of America, 2013, 133(5): 3591−3597

[37] Taal C H, Hendriks R C, Heusdens R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech [C] // Proc of the 31st IEEE Int Conf on Acoustics Speech and Signal Processing. Piscataway, NJ: IEEE, 2010: 4214−4217

[38] Rix A W, Beerends J G, Hollier M P, et al. Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs [C] //Proc of the 26th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2001: 749−752

[39] Roux L, Wisdom J, Erdogan S, et al. SDR − half-baked or well done? [C] //Proc of the 44th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2019: 626−630

[40] Tan Ke, Chen Jitong, Wang Deliang. Gated residual networks with dilated convolutions for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(1): 189−198

[41] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-Net: Learning where to look for the pancreas[J]. arXiv preprint, arXiv: 1804. 03999, 2018

[42] Li Andong, Zheng Chengshi, Fan Cunhang, et al. A recursive network with dynamic attention for monaural speech enhancement[J]. arXiv preprint, arXiv: 2003. 12973, 2020

Kang Hongbo, born in 1997. Master. Her main research interests include speech signal processing and machine learning.

康宏博, 1997 年生. 硕士. 主要研究方向为语音信号处理和机器学习.

Feng Yujia, born in 1997. Master. Her main research interests include speech signal processing and machine learning.

冯雨佳, 1997 年生. 硕士. 主要研究方向为语音信号处理和机器学习.

Tai Wenxin, born in 1997. PhD candidate. His main research interests include speech signal processing and machine learning.

台文鑫, 1997 年生. 博士研究生. 主要研究方向为语音信号处理和机器学习.

Lan Tian, born in 1977. PhD, professor, master supervisor. His main research interests include medical image processing, speech enhancement, and natural language processing.

蓝　天, 1977 年生. 博士, 研究员, 硕士生导师. 主要研究方向为医学图像处理、语音增强和自然语言处理.

Wu Zufeng, born in 1978. PhD, associate professor, master supervisor. His main research interests include speech signal processing and machine learning.

吴祖峰, 1978 年生. 博士, 副研究员, 硕士生导师. 主要研究方向为医学语音信号处理和机器学习.

Liu Qiao, born in 1974. PhD, professor, PhD supervisor. His main research interests include natural language processing, machine learning, and data mining.

刘　峤, 1974 年生. 博士, 教授, 博士生导师. 主要研究方向为自然语言处理、机器学习和数据挖掘.