计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):60−68 [doi: 10.15888/j.cnki.csa.009458]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 深度复数轴向自注意力卷积循环网络的语音增强①

曹　洁[1,2], 王　乔[1], 梁浩鹏[1], 王宸章[1], 李晓旭[1], 于　泓[3]

[1](兰州理工大学 计算机与通信学院, 兰州 730050)
[2](兰州城市学院 信息工程学院, 兰州 730020)
[3](鲁东大学 信息与电气工程学院, 烟台 264025)
通信作者: 曹　洁, E-mail: haop1115@163.com

**摘　要**: 单通道语音增强任务中相位估计不准确会导致增强语音的质量较差, 针对这一问题, 提出了一种基于深度复数轴向自注意力卷积循环网络 (deep complex axial self-attention convolutional recurrent network, DCACRN) 的语音增强方法, 在复数域同时实现了语音幅度信息和相位信息的增强. 首先使用基于复数卷积网络的编码器从输入语音信号中提取复数表示的特征, 并引入卷积跳连模块用以将特征映射到高维空间进行特征融合, 加强信息间的交互和梯度的流动. 然后设计了基于轴向自注意力机制的编码器-解码器结构, 利用轴向自注意力机制来增强模型的时序建模能力和特征提取能力. 最后通过解码器实现对语音信号的重构, 同时利用混合损失函数优化网络模型, 提升增强语音信号的质量. 实验在公开数据集 Valentini 和 DNS Challenge 上进行, 结果表明所提方法相对于其他模型在客观语音质量评估 (perceptual evaluation of speech quality, PESQ) 和短时客观可懂度 (short-time objective intelligibility, STOI) 两项指标上均有提升, 在非混响数据集中, PESQ 比 DCTCRN (deep cosine transform convolutional recurrent network) 提高了 12.8%, 比 DCCRN (deep complex convolutional recurrent network) 提高了 3.9%, 验证了该网络模型在语音增强任务中的有效性.

**关键词**: 单通道语音增强; 复数卷积循环网络; 卷积跳连; 轴向自注意力机制

引用格式: 曹洁,王乔,梁浩鹏,王宸章,李晓旭,于泓.深度复数轴向自注意力卷积循环网络的语音增强.计算机系统应用,2024,33(4):60–68. http://www.c-s-a.org.cn/1003-3254/9458.html

## Speech Enhancement Based on Deep Complex Axial Self-attention Convolutional Recurrent Network

CAO Jie[1,2], WANG Qiao[1], LIANG Hao-Peng[1], WANG Chen-Zhang[1], LI Xiao-Xu[1], YU Hong[3]

[1](School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China)
[2](School of Information Engineering, Lanzhou City University, Lanzhou 730020, China)
[3](School of Information and Electrical Engineering, Ludong University, Yantai 264025, China)

**Abstract**: Inaccurate phase estimation in single-channel speech enhancement tasks will cause poor quality of the enhanced speech. To this end, this study proposes a speech enhancement method based on a deep complex axial self-attention convolutional recurrent network (DCACRN), which enhances speech amplitude information and phase information in the complex domain simultaneously. Firstly, a complex convolutional network-based encoder is employed to extract complex features from the input speech signal, and a convolutional hopping module is introduced to map the features into a high-dimensional space for feature fusion, which enhances the information interaction and the gradient flow. Then an encoder-decoder structure based on the axial self-attention mechanism is designed to enhance the model's timing modeling ability and feature extraction ability. Finally, the reconstruction of the speech signals is realized by the

① 基金项目: 甘肃省重点研发计划 (22YF7GA130)
收稿时间: 2023-10-07; 修改时间: 2023-11-09; 采用时间: 2023-11-24; csa 在线出版时间: 2024-01-18
CNKI 网络首发时间: 2024-01-19

decoder, while the hybrid loss function is adopted to optimize the network model to improve the quality of enhanced speech signals. Meanwhile, the mixed loss function is utilized to optimize the network model and improve the quality of enhanced speech signals. The experiments are conducted on the public datasets Valentini and DNS Challenge, and the results show that the proposed method improves both the perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) metrics compared to other models. In the non-reverberant dataset, PESQ is improved by 12.8% over DCTCRN and 3.9% over DCCRN, which validates the effectiveness of the proposed model in speech enhancement tasks.

**Key words**: single-channel speech enhancement; complex convolutional recurrent network; convolution jump; axial self-attention mechanism

语音信号被不同噪声干扰时, 语音可懂度和语音质量都会受到很大影响. 语音增强要解决的问题是从噪声背景中提取有用的语音信号, 抑制噪声干扰, 改善人们的听觉感受或提高机器识别准确率. 语音增强问题因通道数不同主要分为单通道语音增强和多通道语音增强. 多通道方法包括波束形成与独立成分分析等方法, 单通道方法包括信号处理方法以及掩模估计方法, 其中掩膜估计方法包括模型化方法以及近几年兴起的有监督学习方法. 单通道语音采集条件要求较低, 应用前景更广, 难度更大, 对模型要求更高, 研究更具挑战性, 因此本文主要关注单通道的语音增强方法.

目前常见的语音增强方法主要有两类: 一类是传统的语音增强方法, 主要包括谱减法[1]、维纳滤波法[2]、基于子空间语音增强算法[3]等, 传统方法主要基于高斯噪声假设来进行建模, 这可能会忽略噪声与语音信号之间的相关性, 影响听觉上的感知质量. 另一类是基于深度学习的语音增强方法, 如卷积神经网络 (convolutional neural network, CNN)[4]、循环神经网络 (recurrent neural network, RNN)[5]、生成对抗网络 (generative adversarial network, GAN)[6]等, 这些方法可以学习到更丰富的特征表达, 在训练时往往能够更好地应对噪声、混响和其他语音环境的变化, 具有比传统语音增强方法更高的性能.

作为监督学习问题, 神经网络可以在时域或时频 (time-frequency, TF) 中增强噪声语音. 目前, 通过短时傅立叶变换 (short-time Fourier transform, STFT) 研究 TF 域表示更为常见. 在 TF 域中定义的训练目标主要分为两类, 一类是描述干净语音与背景噪声时频关系的掩模目标, 另一类是对应于干净语音频谱表示的映射目标. 在掩蔽家族中, 理想二进制掩模 (IBM)[7]、理想比值掩模 (IRM)[8]和频谱幅度掩模 (SMM)[9]只使用干净语音和混合语音之间的幅度, 大多数都忽略了相位信息, 导致目标语音细节部分丢失的问题. 相位敏感掩模 (phase-sensitive mask, PSM)[10]是第 1 个展示相位信息可行性的方法, 复数比值掩模 (complex ratio mask, CRM) [11]可以通过实部掩模和虚部掩模的分别学习来进行语音重建.

近年来, 深度复数 U-Net (deep complex U-Net, DCUNET)[12]结合了深度复数网络[13]和 U-Net[14]的优点来处理复值谱图, DCUNET 经过训练, 估计 CRM, 并用短时傅里叶逆变换 (inverse short-time Fourier transform, iSTFT) 将输出 TF 域谱图转换为时域波形后, 优化尺度不变的源噪声比 (scale-invariant source-to-noise ratio, SI-SNR)[15]损失. 针对语谱图中相位信息的表达不充分会影响目标语音估计的问题, Hu 等人[16]提出了模拟复数运算的深度复数卷积循环网络 (deep complex convolutional recurrent network, DCCRN), 通过复数运算结构保留更多的目标语音相位信息, 实现了相位信息的有效捕获. 在提升语音增强任务的性能的同时, 采用复数运算导致模型规模大、复杂性高, 限制了其语音增强效果的进一步提升.

本文以最小化参数量并最大化提升语音增强效果为原则, 提出了基于深度复数轴向自注意力卷积循环网络 (deep complex axial self-attention convolutional recurrent network, DCACRN) 的语音增强方法, 该方法首先设计了一种卷积跳连模块, 通过将编码层的输出与解码层的输入相加更好地进行梯度传递, 并利用卷积跳连过滤噪声, 更好地保留原始信号. 同时, 在编码器-解码器之间插入轴向自注意力机制, 有效捕捉时间序列信号中的长期依赖关系, 增加模型对信号特征的

感知能力, 减少噪声干扰. 最后, 采用短时傅里叶内核初始化卷积/反卷积模块, 在送入网络和计算损失函数之前, 来分析和生成波形, 从而有效提高语音的感知质量. 实验结果表明, 所提出的 DCACRN 模型在小型 Valentini 数据集和大型公开数据集 DNS-2020 上实现了更好的语音增强性能.

## 1 深度复数轴向自注意力卷积循环网络模型

### 1.1 深度复数卷积循环网络架构

卷积循环网络 (convolutional recurrent network, CRN)[17]是一种编码器-解码器架构, Hu 等人[16]在 CRN 模型的基础上, 引入复数卷积和长短期记忆网络 (long short term memory, LSTM), 提出深度复数卷积循环网络 (DCCRN), 该模型中编码器-解码器的接受域有限, 跨频率上下文的特征表示存在一定的局限性, 以及复杂网络结构的计算量非常大, 这将面临很大的挑战. 针对这些问题, 本文设计了一种深度复数轴向自注意力卷积循环网络 (DCACRN), 整体结构如图 1 所示. DCACRN 模型包括对称的编码器-解码器、复数 LSTM、卷积跳连模块、轴向自注意力机制. 为了提高运算速度, 特征提取过程中的短时傅里叶变换 (STFT), 使用参数固定的短时傅里叶卷积 (Conv-STFT) 和短时傅里叶逆卷积 (Conv-iSTFT) 操作来实现.

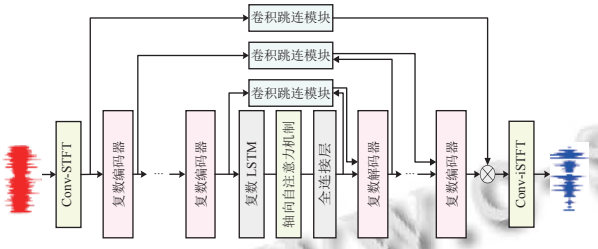


图 1 DCACRN 网络

编码器/解码器由二维卷积/反卷积、批量归一化和 PReLU 激活函数组成. 首先, 通过 Conv-STFT 将时域音频信号转换为时频域表示, 编码器对输入数据进行逐层处理, 增大通道数, 减少频域维度, 解码器将频率分辨率的特征重构. 在编码器-解码器之间插入复数 LSTM, 旨在对语音中的时间依赖性进行建模. 然后, 卷积跳跃连接操作用于加强信息间的交互和梯度的传递, 同时轴向自注意力机制位于复数 LSTM 之后, 用于关注其重要信息. 最后, 通过 Conv-iSTFT 将时频域表示转换回时域信号, 以进行语音任务的预测或输出.

本文与 DCCRN 有以下区别.

(1) DCACRN 模型不是简单地将编码器层的输出和相应的解码器层的输出相拼接, 而是添加一种基于注意力机制的卷积跳连模块, 将编码器的输出和解码器的输入进行信息融合, 提取多尺度信息和更全局的上下文信息, 实现更为有效的信息传递.

(2) 传统的 DCCRN 是用简单的 LSTM 进行时序建模, 本文在编码器-解码器之间插入轴向自注意力机制, 分别经过频域和时域的预卷积, 得到注意力权重与相应位置的特征向量, 捕获不同维度之间的依赖性, 使其对帧内或帧间的信息进行更为有效的融合, 之后结合复数 LSTM 模块一起进行时序建模, 进而增强模型的时序建模能力和特征提取能力.

### 1.2 复数编码器模块

复数编码器模块包括复数二维卷积、复数批归一化[18]和实值 PReLU[19]. 复数二维卷积包含两个二维卷积层 $W_r$ 和 $W_i$, 通过 4 次卷积操作实现复数卷积流程. 接着在复数域上进行复数归一化处理以增强网络的稳定性, 并引入实值 PReLU 作为非线性激活函数以促进模型的非线性建模能力. 复数二维卷积如图 2(a) 所示, 复数编码器模块如图 2(b) 所示.

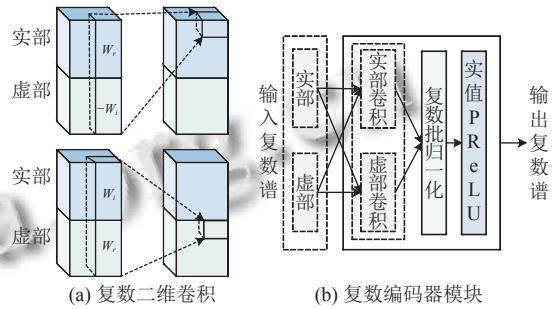


图 2 复数模块

将复数卷积层 $W$ 定义为 $W = W_r + jW_i$, 其中 $W_r$ 和 $W_i$ 分别表示复数卷积核的实部和虚部, 同时定义输入复数矩阵 $X = X_r + jX_i$, 因此, 通过复数卷积运算 $X \otimes W$ 得到复数输出:

$$F_{\text{out}} = (X_r \otimes W_r - X_i \otimes W_i) + j(X_r \otimes W_i + X_i \otimes W_r) \quad (1)$$

其中, $F_{\text{out}}$ 表示一个复数卷积层的输出特征. 与复数卷积类似, 复数 LSTM 同样包含实部 $LSTM_r$ 和虚部 $LSTM_i$, 复数 LSTM 输出 $F_{\text{out}}$ 可以定义为:

$$F_{rr} = LSTM_r(X_r);\ F_{ir} = LSTM_r(X_i) \quad (2)$$

$$F_{ri} = LSTM_i(X_r);\ F_{ii} = LSTM_i(X_i) \quad (3)$$

$$F_{\text{out}} = (F_{rr} - F_{ii}) + j(F_{ri} + F_{ir}) \tag{4}$$

其中, $F_{ri}$ 是由输入 $X_r$ 和 $LSTM_i$ 计算得到的.

### 1.3 卷积跳连模块

在 DCACRN 模型中, 基于注意力机制的编码器和解码器之间的卷积跳连, 通过将编码器层的输出与对应解码器层的输入进行拼接, 卷积跳连的作用是将编码器的输出 $U_i$ 和所对应解码器的输入 $C_i$ 进行信息融合, 获取更加丰富的上下文信息. 如图 3 所示, 将 $U_i$ 和 $C_i$ 二者进行信息融合, 学习得到注意力权重 $A_i$, 利用 $A_i$ 对 $C_i$ 进行加权, 得到更为有效的解码器输入信息. 注意力权重 $A_i$ 的学习过程就是将两个二维卷积相加之后进行 *Sigmoid*, 然后再进行卷积和 *Sigmoid*, 最后得到卷积跳连模块的输出 $B_i$.

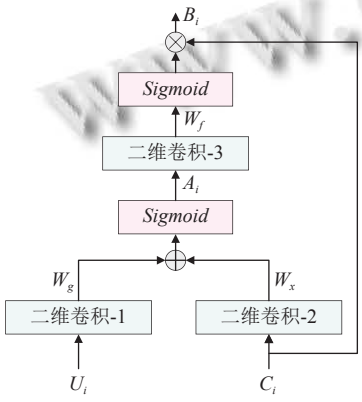


图 3 卷积跳连模块

图 3 中的两个二维卷积, 核大小为 1, $W_g$ 是输出通道, $W_x$ 是输入通道, $W_g$ 是 $W_x$ 的 2 倍, 用于将 $U_i$ 和 $C_i$ 映射到高维空间特征进行信息融合, 学习得到注意力权重 $A_i$:

$$A_i = \sigma\left(W_g \otimes U_i + W_x \otimes C_i\right) \tag{5}$$

其中, $\sigma$ 是 *Sigmoid* 函数, $W_f$ 表示另一个二维卷积层. 卷积跳连模块的输出 $B_i$ 为:

$$B_i = \sigma\left(W_f \otimes B_i\right) \cdot C_i \tag{6}$$

### 1.4 轴向自注意力机制

自注意力可以提高网络捕捉特征之间长期关系的能力. 与计算机视觉中的像素级注意力不同[20], 本文中语音的轴向自注意力 (axial self-attention, ASA) 分别在水平和竖直方向进行自注意计算, 行和列组合使用才能更好地融合全局信息, 减少对内存和计算的需求, 也更适合于语音等长序列信号.

轴向自注意力机制的结构如图 4 所示, 输入特征维度是 $C_i \times T \times F$, $C_i$ 表示通道数, 首先经过频域的预卷积, 生成频域的自注意力参数 $Q^F$、$K^F$ 和 $V$, $Q^F$、$K^F$ 和 $V$ 的通道数是 $C$, 然后, 再通过时域的预卷积, 得到用来进行时域注意力的参数 $Q^T$、$K^T$. $Q$、$K$、$V$ 分别表示注意力中的查询 (Query)、键 (Key) 和值 (Value), 它们在计算注意力权重和调节不同位置/特征的关注程度起着关键作用.

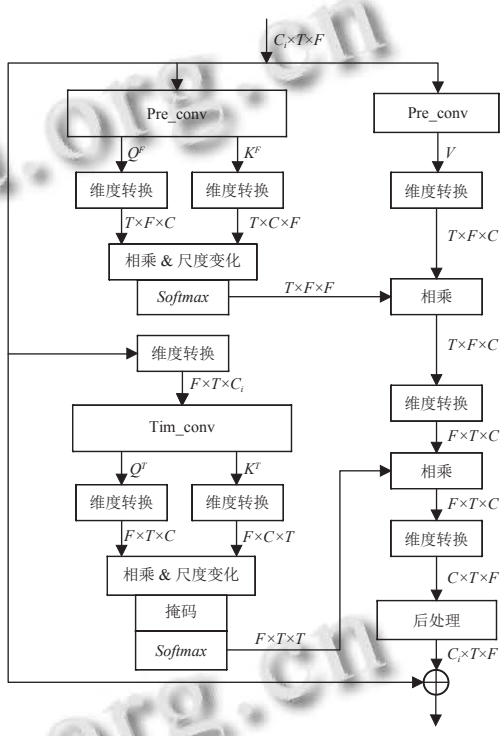


图 4 轴向自注意力模块

轴向自注意力的注意得分矩阵沿频率轴和时间轴计算, 分别称为频率-注意力 (F-attention) 和时间-注意力 (T-attention). 分数矩阵可以表示为:

$$M_F(t) = Softmax\left(Q_f(t) K_f^T(t)\right) \tag{7}$$

$$M_T(f) = Softmax\left(Mask\left(Q_t(f) K_t^T(f)\right)\right) \tag{8}$$

其中, $Q_f(t)$, $K_f(t) \in R^{T \times C}$, $M_F(t) \in R^{F \times F}$ 表示 F-attention 在第 $t$ 帧的键值、查询和得分矩阵. $Q_t(f)$, $K_t(f) \in R^{F \times C}$, $M_T(f) \in R^{T \times T}$ 表示频率波段 T-attention 的键值、查询和得分矩阵. *Softmax* 将沿着最后一个维度计算, T-attention 中的 $Mask(*)$ 通过调整 ASA 更好地捕获时序依赖关系, 保证因果性.

### 1.5 损失函数

尺度不变信噪比 (SI-SNR) 是噪声抑制中常用的

评价指标, SI-SNR 是一种信号级损耗, 直接作用于信号本身. 时域 SI-SNR 损失$\mathcal{L}_{\text{SI-SNR}}(y,\hat{y})$定义为:

$$\begin{cases} y_{\text{target}} = (\langle \hat{y}, y\rangle \cdot y)/\|y\|_2^2 \\ e_{\text{noise}} = \hat{y} - y_{\text{target}} \\ \mathcal{L}_{\text{SI-SNR}}(y,\hat{y}) = 10\lg\left(\dfrac{\|y_{\text{target}}\|_2^2}{\|e_{\text{noise}}\|_2^2}\right) \end{cases} \tag{9}$$

首先得到$y_{\text{target}}$语音分量, 然后得到噪声分量$e_{\text{noise}}$, 最后结合两者计算出 SI-SNR. 其中, $y$和$\hat{y}$分别为干净和估计的时域波形, $\langle\cdot,\cdot\rangle$表示两个向量的点积, $\|\cdot\|_2$为欧氏范数 (L2 范数).

此外, 还使用复数比值掩模 (CRM) 估计的均方误差 (MSE) 损失[21]来指导信噪比估计器的学习. 掩码损失函数定义为:

$$\mathcal{L}_{\text{Mask}}\left(M,\hat{M}\right) = \sum_{t,f}[(\hat{M}_r - M_r)^2 + (\hat{M}_i - M_i)^2] \tag{10}$$

其中, $M_r$和$M_i$表示目标频谱掩码的实部和虚部, $\hat{M}_r$和$\hat{M}_i$表示估计频谱掩码的实部和虚部, 计算所有时间点和频率点的平方差之和, 使估计的频谱掩码尽可能接近目标频谱掩码.

本文将 SI-SNR 损失与 CRM 估计的均方误差 (MSE) 损失集成, 以此来优化神经网络模型:

$$\mathcal{L}(y,\hat{y}) = \lambda_{\text{SI-SNR}}\mathcal{L}_{\text{SI-SNR}}(y,\hat{y}) + \lambda_{\text{Mask}}\mathcal{L}_{\text{Mask}}(M,\hat{M}) \tag{11}$$

其中, $\lambda$为加权因子, 用于平衡两个损失项的权重系数. $(y,\hat{y})$表示目标信号和估计信号之间的损失, $(M,\hat{M})$表示目标频谱掩码和估计的频谱掩码之间的损失. SI-SNR 损失项有助于提高源信号的还原质量, 而 Mask 损失项有助于确保频谱掩码的准确性.

本文使用短时傅里叶变换内核初始化卷积/反卷积模块对波形进行分析/合成[22]. 对于一个给定的波形信号, 首先使用初始化的短时傅里叶变换内核对其进行特定的信号处理或增强操作, 将时域信号转换为频域信号; 接着, 使用反卷积模块将处理后的频域信号进行合成, 将其还原为时域信号. 最后, 将语音信号发送到网络并计算损失函数.

## 2 实验配置

### 2.1 实验设置

本文实验所用的 DCACRN 模型如图 1 所示, 其中各网络层的参数见表 1. 实验环境为 15 vCPU AMD EPYC 7543 32-Core Processor, GPU 采用 RTX 3090 显存 24 GB, 内存 80 GB. 在此基础上, 服务器采用 Ubuntu 20.24 系统, Python 3.8, CUDA 11.3, PyTorch 1.11.0 的开发环境.

表 1 DCACRN 网络模型的参数

| 层 | 输入大小 | 超参数 | 输出大小 |
|---|---|---|---|
| c-conv2d_1 | 2×256×$T$ | 5×2, (2, 1), 16 | 16×128×$T$ |
| c-conv2d_2 | 16×128×$T$ | 5×2, (2, 1), 32 | 32×64×$T$ |
| c-conv2d_3 | 32×64×$T$ | 5×2, (2, 1), 64 | 64×32×$T$ |
| c-conv2d_4 | 64×32×$T$ | 5×2, (2, 1), 128 | 128×16×$T$ |
| c-conv2d_5 | 128×16×$T$ | 5×2, (2, 1), 128 | 128×8×$T$ |
| Reshape_1 | 128×8×$T$ | — | $T$×64×8 |
| c-LSTM | $T$×64×8 | 128 | $T$×64×8 |
| Reshape_2 | $T$×64×8 | — | 128×8×$T$ |
| ASA | 128×8×$T$ | 1×1, (1, 1), 128 | 128×8×$T$ |
| c-deconv2d_5 | 256×8×$T$ | 5×2, (2, 1), 128 | 128×16×$T$ |
| c-deconv2d_4 | 256×16×$T$ | 5×2, (2, 1), 64 | 64×32×$T$ |
| c-deconv2d_3 | 64×32×$T$ | 5×2, (2, 1), 32 | 32×64×$T$ |
| c-deconv2d_2 | 32×64×$T$ | 5×2, (2, 1), 16 | 16×128×$T$ |
| c-deconv2d_1 | 64×128×$T$ | 5×2, (2, 1), 1 | 8×128×$T$ |

### 2.2 数据集

为了全方位评估所提出方法的降噪表现, 本文分别选择在小型公开数据集 Valentini[23]和大型公开数据集 DNS Challenge[24]上进行实验. 数据集 1 利用语音增强领域常用的噪声库和语音库合成制作, 用于验证各信噪比条件下和有无混响情况下的降噪表现. 数据集 2 的训练集和验证集是基于 DNS-2020 挑战赛官方开源数据合成, 测试集是采用 DNS-2020 官方合成的无混响测试集, 用于和当前先进的降噪模型进行比较. 以下将对两个数据集的具体细节进行描述.

数据集 1 的采样率是 48 kHz, 合成带噪信号之前需要将语音信号重采样为 16 kHz. 干净语音的数据集是来自各种文本段落的句子录音, 并从 Voice Bank 语料库[25]中选择了 30 个英语演讲者, 包括具有各种口音的男性和女性, 28 个和 2 个说话人分别被分配到训练集和测试集. 噪声数据来自于噪声库 NOISEX-92[26], 该噪声库中包含了白噪声, 粉红噪声, 高频通道噪声, 工厂底层噪声等 15 种噪声. 本文利用以上干净语音和噪声合成了 50 h 的训练集, 其具体设置为: 带噪语音的信噪比 (signal-to-noise ratio, SNR) 为 0–20 dB, 其中 40% 的语音数据不带有混响, 剩余的 60% 数据带有混响 (T60 为 0.3–1.3 s). 房间脉冲响应 (room impulse response, RIR) 是从 DNS RIR 数据集中随机选择的. 为了验证该模型在不同的信噪比和混响或无混响情况下的噪声抑制性能, 生成了两个测试集: 混响和非混响测试集, 两个测试集语音信噪比都被设置为 0 dB、5 dB、10 dB、15 dB、20 dB.

数据集 2 是基于 Interspeech 2020 年 DNS 挑战赛数据集生成的, 所有波形的采样频率为 16 kHz. 该挑战赛的干净语音数据集来源于公共有声读物数据集 Librivox. Librivox 是一个非营利性组织, 旨在将公共领域的书籍录制成有声读物, 让人们可以免费获取和使用这些资源. 该组织录制了超 10 000 本不同语言的有声读物, 其中大部分为英文. 该项目中包含来自 2 150 名发言者的超过 500 h 的语音. DNS 挑战赛的噪声数据集由 180 小时的噪声集组成, 涵盖 150 个类别和 65 000 个噪声片段, 这些片段从 Audioset2 和 Freesound3 中选出. 使用随机选取的语音片段和噪声片段生成了一个 500 h 的噪声训练集, 其信噪比范围为−10 dB 到 20 dB. 每个选定的音频片段长度被设置为 10 s. 为了与挑战赛中的先进模型的降噪性能进行比较, 本文使用了由 DNS-2020 合成的无混响测试集来评估所提出的模型.

### 2.3 评价指标

为客观评价不同网络的语音增强性能, 分别采用不同网络对测试集含噪语音进行语音增强, 并比较不同网络增强后语音的平均语音质量和平均可懂度, 其中, 语音质量的评价指标为语音质量的感知评估 (perceptual evaluation of speech quality, PESQ)[27], 其得分范围为−0.5–4.5, 得分越高代表语音质量越好; 语音可懂度的评价指标短时客观可懂度 (short-time objective intelligibility, STOI)[28], 其得分范围为 0–1, 得分越高代表语音可懂度越高.

### 2.4 模型训练和基线方法

本文使用一个周期性汉宁窗, 窗口长度和帧移动分别为 25 ms 和 6.25 ms, 特征长度为 512. 对于所有模型的训练, 本文使用 PyTorch 平台和 Adam 优化器[29], 将初始学习率设置为 0.001, 批处理大小为 8, 网络训练周期为 100. 当验证损失增加时, 学习率将衰减 0.5, 对于损耗参数, 设置 $\lambda_{\text{SI-SNR}}=0.5$, $\lambda_{\text{Mask}}=0.5$. 实验中, 对所有的输入音频信号降采样至 16 kHz, 通过提前停止选择模型. 将提出的模型与深度复数卷积循环网络 (DCCRN)[16]和深度余弦变换卷积循环网络 (deep cosine transform convolutional recurrent network, DCTCRN)[30]进行比较, 遵循原始参数设置如下.

(1) DCCRN: 窗口长度为 25 ms, 帧移动为 6.25 ms, 特征长度为 512. DCCRN 的通道数为{32, 64, 128, 128, 256, 256}, 卷积核大小和步长设置为 (5, 2) 和 (2, 1). 设置 2 层 LSTM 节点数量为 256 个, LSTM 后有一个 1024×256 的全连通层. 在编码器模块中, 在每个卷积编码器层的时间维度前填充一个零帧. 在解码器模块中, 在每个卷积层中查看一帧.

(2) DCTCRN: 窗口长度为 32 ms, 帧移动为 8 ms, 特征长度为 512. DCTCRN 的通道数为{8, 16, 32, 64, 128, 128, 256}, 卷积核大小和步长设置为 (5, 2) 和 (2, 1). 设置 2 层 LSTM 节点数量为 256 个. 在编码器模块中, 在每个 conv2d 的时间维度前面垫一个零帧. 在解码器模块中, 移除每个转置卷积解码器的最后一个时间帧.

(3) Baseline: 窗口长度为 25 ms, 帧移动为 6.25 ms, 特征长度为 512. 基线的通道数为{16, 32, 64, 128, 128}, 卷积核大小和步长设置为 (5, 2) 和 (2, 1). F-T-LSTM 节点个数设置为 128, 与 DCCRN-E 一样, 在编码器模块中, 在每个卷积编码器层的时间维度前填充一个零帧. 在解码器模块中, 在每个卷积层中查看一帧.

### 2.5 损失收敛曲线

在语言增强任务中, 使用尺度不变信噪比 SI-SNR 和均方误差 MSE 联合损失函数是一种有效的方法. 这种联合使用的目标是在保留语音质量的同时, 优化语音增强模型的输出, 使其更好地适应原始语音信号. 图 5 是损失函数的收敛曲线, 横坐标表示训练周期, 纵坐标表示损失值, 负值越低效果越好. 蓝色是训练集的损失曲线, 黄色是验证集的损失曲线. 在训练的早期, 模型参数随机初始化, 损失迅速下降. 在训练过程中, 模型通过梯度下降不断地调整参数, 逐渐学到更好的表示, 从而提高语音增强的性能. 最后阶段, 模型已经学到了数据的关键特征, 损失几乎不再下降, 进一步的训练对性能的提升不会太大, 有可能还会导致过拟合.

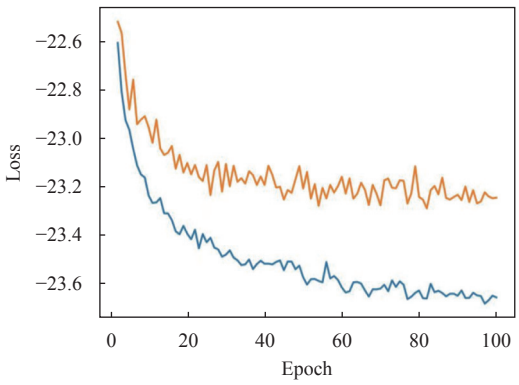


图 5 损失收敛曲线图

## 3 实验分析

实验主要包括 2 个方面.

1) 基于 DNS-2020 大数据集开展消融实验, 以探究卷积跳连和轴向自注意力对语音增强性能的影响.

2) 针对 Valentini 数据集, 在不同信噪比的情况下, 将本文所提模型与现有的语音增强网络做性能对比.

### 3.1 消融实验

消融实验结果仅使用客观语音质量评价指标 PESQ 得分在数据集 2 上验证 DCTCRN、DCCRN 和本文所提出模型 DCACRN 的降噪性能, Noisy 表示未经语音增强方法处理的带噪语音的评估结果, 与 Noisy 比较可以直观地看到增强方法对信号的改善效果. 表 2 是客观语音质量 PESQ 的评估结果.

表 2 DNS Challenge 数据集上的客观语声质量评估结果

| 模型 | Param (M) | GMacs | Look ahead (ms) | PESQ |
|---|---|---|---|---|
| Noisy | — | 0 | 0 | 2.45 |
| DCTCRN[30] | 2.86 | 2.69 | 0 | 3.24 |
| DCCRN[16] | 2.74 | 3.96 | 37.5 | 3.26 |
| Baseline | 2.51 | 3.34 | 31.25 | 3.25 |
| +skip-conv | 2.74 | 3.96 | 31.25 | **3.33** |
| +ASA (DCACRN) | 2.77 | 3.99 | 31.25 | **3.41** |

实验结果表明, 本文提出的模型相比其他模型具有更高的 PESQ 得分. 卷积跳连将低维特征和高维特征相融合, 实现更为有效的信息传递; 轴向自注意力能够捕获序列数据中的全局依赖关系, 提高模型的表达能力, 使其能够更准确地建模输入数据的特征和关系. 从表 2 中可以看出, 在加入卷积跳连和轴向自注意力模块之后, PESQ 指标分别增加了 0.08 和 0.16. 与 DCCRN 相比, 本文模型用相对少的参数达到了更好的降噪效果, 反映了添加这两个模块的有效性, 也证实了该结构可以达到增强网络性能的效果.

### 3.2 Valentini 数据集实验结果分析

本文使用语音质量的感知语音质量评价 (PESQ)、短时客观可理解度 (STOI) 作为客观指标. 为了更好地证明所提算法的性能, 利用 Valentini 数据集来验证在有无混响情况下不同信噪比的降噪性能, 各类网络模型的 PESQ 和 STOI 的得分如下所示. 表 3 和表 4 分别表示为无混响情况下测试集上的客观结果, 表 5 和表 6 分别表示为有混响条件下的结果. 在每种情况下, 最佳结果都用加粗数字突出显示.

从非混响数据集的结果可以发现, 在不同信噪比的情况下, 本文的模型在 PESQ 和 STOI 两个指标上都优于 DCCRN 和 DCTCRN, 证明 DCACRN 达到了最先进的性能. 从表 3 和表 4 可以看出, 信噪比为 20 dB 时, 最佳 PESQ 得分为 3.404, STOI 得分为 0.968, 与原来的 DCCRN 相比有明显提高.

表 3 不同模型在非混响数据集 2 上的 PESQ 得分

| 模型 | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB | Avg |
|---|---|---|---|---|---|---|
| Noisy | 1.559 | 1.876 | 2.222 | 2.560 | 2.870 | 2.217 |
| DCTCRN | 1.872 | 2.143 | 2.483 | 2.731 | 3.016 | 2.449 |
| DCCRN | 2.362 | 2.690 | 2.909 | 3.098 | 3.274 | 2.867 |
| DCACRN | **2.419** | **2.743** | **2.996** | **3.222** | **3.404** | **2.957** |

表 4 不同模型在非混响数据集 2 上的 STOI 得分

| 模型 | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB | Avg |
|---|---|---|---|---|---|---|
| Noisy | 0.735 | 0.820 | 0.886 | 0.932 | 0.960 | 0.867 |
| DCTCRN | 0.762 | 0.852 | 0.890 | 0.934 | 0.963 | 0.880 |
| DCCRN | 0.809 | 0.871 | 0.899 | 0.934 | 0.965 | 0.896 |
| DCACRN | **0.823** | **0.883** | **0.917** | **0.936** | **0.967** | **0.905** |

表 5 不同模型在混响数据集 2 上的 PESQ 得分

| 模型 | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB | Avg |
|---|---|---|---|---|---|---|
| Noisy | 1.687 | 1.980 | 2.299 | 2.628 | 2.911 | 2.301 |
| DCTCRN | 1.895 | 2.064 | 2.345 | 2.735 | 3.093 | 2.426 |
| DCCRN | 2.126 | 2.513 | 2.809 | 3.046 | 3.190 | 2.737 |
| DCACRN | **2.234** | **2.613** | **2.891** | **3.077** | **3.194** | **2.802** |

表 6 不同模型在混响数据集 2 上的 STOI 得分

| 模型 | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB | Avg |
|---|---|---|---|---|---|---|
| Noisy | 0.723 | 0.829 | 0.905 | 0.953 | 0.978 | 0.878 |
| DCTCRN | 0.772 | 0.804 | 0.895 | 0.955 | 0.979 | 0.881 |
| DCCRN | 0.781 | 0.865 | 0.912 | 0.956 | 0.981 | 0.899 |
| DCACRN | **0.793** | **0.877** | **0.921** | **0.958** | **0.982** | **0.906** |

本文的模型通过在编码器和解码器之间添加卷积跳连模块, 使低层和高层的特征信息能够相互交流和利用, 有助于学习更加丰富的语音特征表示, 提高语音增强网络对语音信号的建模和还原能力. 而轴向自注意力机制模块可以使网络能够更加关注语音信号中的重要信息, 提高对语音内容的理解和处理能力. 综上所述, DCACRN 模型相比其他主流网络有着更好的性能.

在混响测试集上, 表 5 和表 6 分别展示了本文模型与 DCCRN 和 DCTCDN 两种对比模型测评得到的 PESQ 分数和 STOI 分数. 从中可看出: 在 0 dB 时, PESQ 提高了 0.108, STOI 提高了 0.012; 在 20 dB 时, PESQ 提高 0.004, STOI 提高 0.001, 说明 DCACRN 模型在低信噪比条件下可以表现出更好的性能. 本文引入轴向自注意力机制可以自适应地对不同时间的语音信号进行加权, 增强对语音信号的突出和保留; 卷积跳连模块过滤了一些从编码器层连接到解码器层的噪声特征, 提高语音信号的清晰度和准确性.

在语音增强实验中划分有混响和无混响的数据集, 是为了模拟真实世界中存在的不同环境. 对模型来说, 训练时使用有混响和无混响的数据集可以帮助模型学习到更广泛的语音特征, 使其能够更好地应对不同环境下的语音增强任务. 因此, 本文提出的 DCACRN 模型可以改善语音增强结果的质量, 使语音增强性能变得更加优越.

### 3.3 可视化分析

为了更加直观地看出所提方法对语音增强的效果, 图 6 展示了混合噪声之前的纯净语音、输入的带噪语音和增强后语音的语谱图. 增强后的语谱图与 DCTCRN 和 DCCRN 两个模型做了对比, 第 3 行为 DCTCRN 的语谱图, 第 4 行为 DCCRN 的语谱图, 第 5 行为本文模型 DCACRN 的语谱图.

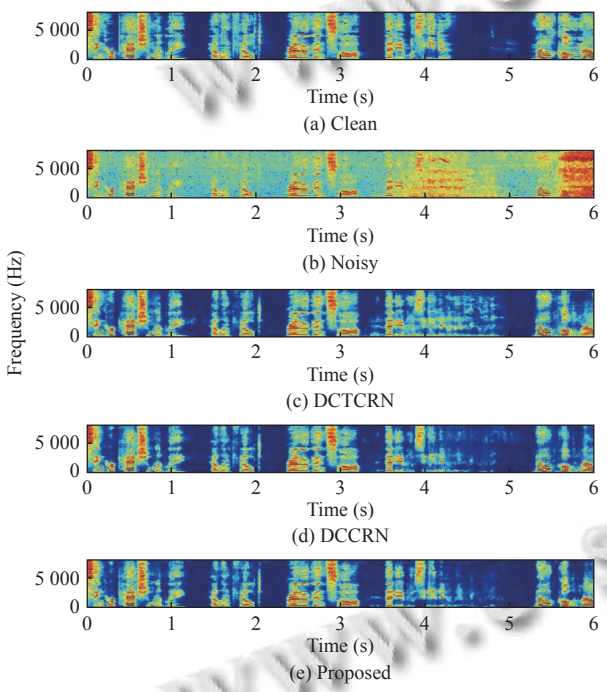


图 6 语音信号语谱图

从图 6 可以看出, DCTCRN 的效果比 DCCRN 低, DCACRN 模型增强语音的语谱图很接近纯净语音的语谱图, 也表明了 DCACRN 模型能有效去除噪声信息, 达到了语音增强目的. 这也进一步验证了, 所提出的深度复数轴向自注意力卷积循环网络模型, 可以有效加强目标语音信息而抑制语音噪声, 增强目标语音的清晰度.

## 4 结论与展望

针对现有基于深度学习的语音增强模型中参数规模大、计算复杂度高、相位估计不准确等问题, 本文提出了一种深度复数轴向自注意力卷积循环网络模型, 该模型由卷积编码器-解码器、卷积跳连模块、复数 LSTM、轴向自注意力构成, 卷积跳连过滤了一些从编码器层连接到解码器层的噪声特征, 轴向自注意力机制可以减少对内存和计算的需求, 更适合于语音等长序列信号, 复数模块通过模拟复乘法来模拟幅值和相位之间的关系. 实验结果表明, 本文提出的 DCACRN 网络在模型参数配置相似的情况下, PESQ 和 STOI 评分优于其他模型. 在未来, 本文将在不忽略相位信息的同时, 把复数网络转换成实数网络, 在效果相似的情况下降低参数量和计算量.

## 参考文献

1 Lim J, Oppenheim A. All-pole modeling of degraded speech. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(3): 197–210. [doi: 10.1109/TASSP.1978.1163086]

2 Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113–120. [doi: 10.1109/TASSP.1979.1163209]

3 Ephraim Y, Van Trees HL. A signal subspace approach for speech enhancement. IEEE Transactions on Speech and Audio Processing, 1995, 3(4): 251–266. [doi: 10.1109/89.397090]

4 Park SR, Lee J. A fully convolutional neural network for speech enhancement. Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm: 2017. 1993–1997.

5 Gao T, Du J, Dai LR, *et al*. Densely connected progressive learning for LSTM-based speech enhancement. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5054–5058.

6 Ye SS, Hu XH, Xu XK. TDCGAN: Temporal dilated convolutional generative adversarial network for end-to-end speech enhancement. arXiv:2008.07787, 2020.

7 Wang DL. On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi P, ed. Speech Separation by Humans and Machines. Boston: Springer, 2005. 181–197.

8 Narayanan A, Wang DL. Ideal ratio mask estimation using deep neural networks for robust speech recognition. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 7092–7096.

9 Wang YX, Narayanan A, Wang DL. On training targets for

supervised speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849–1858. [doi: 10.1109/TASLP.2014.2352935]
10 Erdogan H, Hershey JR, Watanabe S, *et al*. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane: IEEE, 2015. 708–712.
11 Williamson DS, Wang YX, Wang DL. Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(3): 483–492. [doi: 10.1109/TASLP.2015.2512042]
12 Choi HS, Kim JH, Huh J, *et al*. Phase-aware speech enhancement with deep complex U-Net. Proceedings of the 7th International Conference on Learning Representations. New Orleans: OpenReview.net, 2019.
13 Trabelsi C, Bilaniuk O, Zhang Y, *et al*. Deep complex networks. Proceedings of the 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.
14 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.
15 Luo Y, Mesgarani N. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256–1266. [doi: 10.1109/TASLP.2019.2915167]
16 Hu YX, Liu Y, Lv SB, *et al*. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai: IEEE, 2020. 2472–2476.
17 Tan K, Wang DL. A convolutional recurrent neural network for real-time speech enhancement. Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, 2018. 3229–3233.
18 Pandey A, Wang DL. Exploring deep complex networks for complex spectrogram enhancement. Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 6885–6889.
19 He KM, Zhang XY, Ren SQ, *et al*. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1026–1034.
20 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10022.
21 Zhao SK, Nguyen TH, Ma B. Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021. 6648–6652.
22 Gu RZ, Wu J, Zhang SX, *et al*. End-to-end multi-channel speech separation. arXiv:1905.06286, 2019.
23 Valentini-Botinhao C. Noisy speech database for training speech enhancement algorithms and TTS models. https://datashare.ed.ac.uk/handle/10283/1942. (2016-03-22).
24 Reddy CKA, Gopal V, Cutler R, *et al*. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, 2020. 2492–2496.
25 Veaux C, Yamagishi J, King S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. Proceedings of the 2013 International Conference Oriental COCOSDA Held Jointly with the 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). Gurgaon: IEEE, 2013. 1–4.
26 Varga A, Steeneken HJM. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication, 1993, 12(3): 247–251.
27 Rix AW, Beerends JG, Hollier MP, *et al*. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City: IEEE, 2001. 749–752.
28 Taal CH, Hendriks RC, Heusdens R, *et al*. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125–2136. [doi: 10.1109/TASL.2011.2114881]
29 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego, 2015.
30 Li QL, Gao F, Guan HX, *et al*. Real-time monaural speech enhancement with short-time discrete cosine transform. arXiv:2102.04629, 2021.

(校对责编: 孙君艳)