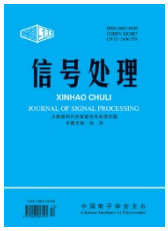

信号处理
Journal of Signal Processing
ISSN 1003-0530,CN 11-2406/TN

《信号处理》网络首发论文

题目：基于可学习图比率掩码估计的图频域语音增强方法
作者：王景润，郭海燕，王婷婷，杨震
网络首发日期：2024-01-29
引用格式：王景润，郭海燕，王婷婷，杨震．基于可学习图比率掩码估计的图频域语音增强方法[J/OL]．信号处理.
https://link.cnki.net/urlid/11.2406.TN.20240126.1647.010

网络首发时间：2024-01-29 09:30:02
网络首发地址：https://link.cnki.net/urlid/11.2406.TN.20240126.1647.010

基于可学习图比率掩码估计的图频域语音增强方法

王景润[1]，郭海燕[1]，王婷婷[1]，杨震*[1,2]
1. 南京邮电大学通信与信息工程学院，江苏 南京 210003；
2. 南京邮电大学通信与网络技术国家地方联合工程研究中心，江苏 南京 210003

摘　要：在基于深度神经网络（deep neural network, DNN）的时频域语音增强方法中，通常将短时傅里叶变换（short-time Fourier transform , STFT）得到的复数域含噪语音时频谱作为 DNN 输入，以估计纯净语音的幅度和相位。此类方法由于会涉及对复数的运算，计算复杂度和模型参数量较大。针对此问题，本文利用图信号处理（graph signal processing, GSP）理论，提出了基于 DNN 的图频域语音增强方法。首先，基于语音信号样点间的位置关系定义实对称的邻接矩阵，将语音信号以无向图形式的图信号进行表示，在此基础上利用对应的图傅里叶变换（graph Fourier transform, GFT）提取实数域的语音图频域特征。由于 GFT 基与邻接矩阵密切相关，该图频域特征隐式地利用了信号样点间的关系，并且可在实数网络中进行处理。然后，构建基于卷积增强 transformer（convolution-augmented transformer, conformer）的网络 GFT-conformer，分别从时间维度和图频率维度捕获图频域特征的局部和全局依赖关系，训练基于掩码的目标，以实现语音增强。最后，考虑到语音和噪声在不同图频率分量上的特性差异，提出可学习图比率掩码（learnable graph ratio mask, LGRM），对不同图频率分量的掩码范围分别进行控制，实现对不同图频率分量的精细化去噪，进一步提升 GFT-conformer 模型的增强性能。在 Voice Bank+DEMAND 数据集和 Deep Xi 数据集上的实验结果表明，所提出的方法在五种常用的评价指标上，优于基于 DNN 的时域和时频域对比方案。

关键词：深度神经网络；语音增强；图傅里叶变换
中图分类号：TN912.35　　**文献标识码：**A

Learnable graph ratio mask based speech enhancement in the graph frequency domain

WANG Jingrun[1], GUO Haiyan[1], WANG Tingting[1], YANG Zhen*[1,2]
1. School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. National Local Joint Engineering Research Center for Communications and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: Deep neural network (DNN) based speech enhancement methods in the time-frequency domain generally work on the complex-valued time-frequency representations obtained through short-time Fourier transform (STFT). With the fine-detailed structures of noisy speech in terms of complex-valued spectrogram as the input, both the magnitude and phase of clean speech can be estimated by DNNs. Such methods mainly employ complex neural networks to deal with complex-valued inputs directly or two-path networks to deal with real and imaginary parts separately, resulting in high computational complexity and a large number of model parameters. To address this problem, this paper proposes a DNN-based speech enhancement method in the graph frequency domain by utilizing the theory of graph signal processing (GSP) to obtain real-valued inputs instead of complex-valued inputs. Specifically, a novel real symmetric adjacency matrix is defined based on the positional relationships among the samples of speech signals so that the speech signals are represented as undirected graph signals. Through eigenvalue decomposition of the real symmetric adjacency

*通讯作者：杨震　yangz@njupt.edu.cn　　*Corresponding Author: YANG Zhen　yangz@njupt.edu.cn
基金项目：国家自然科学基金(62071242)
Fund Project: National Natural Science Foundation(62071242)

matrix, the graph Fourier transform (GFT) basis is obtained and then utilized to extract the real-valued features of the speech graph signals in the graph frequency domain. Since the GFT basis is closely related to the adjacency matrix, these real-valued features in the graph frequency domain implicitly exploit the relationships among speech samples. Furthermore, by combining the convolution-augmented transformer (conformer) and the convolutional recurrent network (CRN), this paper constructs the GFT-conformer model, which is an essentially convolutional encoder-decoder (CED) with four two-stage conformer blocks (TS-conformers) to capture both local and global dependencies of the features in both the time and graph-frequency dimensions, for estimating the targets based on masking to achieve better speech enhancement. Moreover, considering the differences in characteristics between speech and noise across various graph frequency components, this paper introduces the learnable graph ratio mask (LGRM), which allows separate control over the mask ranges for different graph frequency components, enabling fine-grained denoising of various graph frequency components to further improve the speech enhancement performance of the GFT-conformer. We evaluate the performance of the proposed GFT-conformer with LGRM on the Voice Bank+DEMAND dataset and Deep Xi dataset in terms of five commonly used metrics. Experimental results show that the proposed GFT-conformer with LGRM achieves a better performance with the smallest model size of 1.4M parameters, as compared to several other state-of-the-art DNN-based time-domain and time-frequency domain methods.

Keywords: deep neural network; speech enhancement; graph Fourier transform

1 引言

语音增强旨在从受到噪声或者其他干扰的含噪语音中提取出清晰可懂的语音，以提高语音的质量和可理解性。由于在现代自动语音识别、视频会议和助听器设备等实际应用中，感知到的语音质量直接取决于底层语音增强系统的性能，因此不断提升当前语音增强技术的性能具备重要的实际价值。受益于深度学习的快速发展，基于深度神经网络（deep neural network, DNN）的语音增强方法在处理非平稳和复杂噪声环境时明显优于传统的信号处理方法[1]。

将语音增强描述为一个监督学习问题，含噪语音可以通过 DNN 在时域或者时频域进行增强。时域增强[2-3]方法通过 DNN 学习从含噪语音波形到纯净语音波形的映射，不需要进行频谱转换。然而，直接生成高分辨率的波形是一项相对较困难的任务[4]。相比之下，时频域方法更具优势。时频域增强[5-7]方法先采用短时傅里叶变换（short-time Fourier transform, STFT）或其他时频分析方法将语音信号转换为频谱表示，再通过 DNN 在时频域上进行去噪，获得增强后的语音信号。

在时频域语音增强方法中，训练目标通常可以分为两类：基于映射的目标[8]和基于掩码的目标[9]。前者旨在直接通过 DNN 输出干净语音的时频谱，后者旨在通过掩码描述干净语音和背景噪声之间的时频关系。基于掩码的训练目标包括理想二值掩码（ideal binary mask, IBM）[10]、理想比率掩码（ideal ratio mask, IRM）[11]和谱幅掩码（spectral magnitude mask, SMM）[12]等，主要依赖于干净语音和背景噪声之间的幅度信息，忽略了相位信息。相比之下，相位敏感掩码（phase-sensitive mask, PSM）[13]是第一个利用相位信息的方法，展示了相位估计的可行性。随后，文献[14]提出了复数比率掩码（complex ratio mask, CRM），通过同时估计干净语音和含噪语音频谱之比的实部和虚部，来提升重构语音的质量。文献[15]和[16]采用复数神经网络对复数域的时频谱进行处理，以极坐标形式的 CRM 对干净语音的幅度和相位进行同步估计，结果表明基于复数神经网络的方法优于以往仅对幅度进行估计或采用实数神经网络的方法。然而，采用复数神经网络，意味着将 DNN 中的卷积层或循环层从实数扩展到复数操作，这会导致参数量和计算量分别增大为原来的两倍和四倍。

为了避免对复数目标进行估计，一些研究者使用离散余弦变换（discrete cosine transform, DCT）来代替 STFT 提取含噪语音的时频域特征[17-18]。然而，经过 DCT 之后，大部分信号的能量集中在少部分频率分量上，这就可能会导致一些细微的频率特征被模糊化，进而影响增强后语音的细节还原。与 DCT 不同，基于无向图的图傅里叶变换（graph Fourier transform, GFT）[19]能够相对均匀地保留语音信号的所有频率信息。具体而言，将每一帧语音信号视为一个无向图，其样点

视为图上的节点，通过对应的 GFT 就可以得到语音信号的实数图频谱。并且，相比于传统的时频域分析方法，GFT 基由于与邻接矩阵密切相关，因而隐式地利用了信号样点间的关系。此外，文献[20]提出了一种用于语音增强的图维纳滤波器，文献[21]提出了一种图频域内的多通道语音增强方法，文献[22]对比了 STFT 和 GFT 在 U-Net 网络下的性能，均证明了 GFT 在语音增强上的有效性。

另一方面，transformer[23]可以有效地捕获输入特征中的长距离依赖关系，能够提升基于 DNN 的语音增强方法的性能[24]。与 transformer 相比，卷积增强 transformer（convolution-augmented transformer, conformer）[25]将卷积神经网络（convolutional neural network, CNN）和 transformer 的特点相结合，可以有效捕获特征的局部和全局依赖关系，可进一步改善时域[3]和时频域[26-27]语音增强方法的性能，展现出了比长短期记忆（long short-term memory, LSTM）更强的时间建模能力。

在此背景下，本文提出利用 GFT 提取含噪语音的实数图频域特征，采用 conformer 来捕获图频域特征之间的依赖关系，估计实数图频率掩码，来实现语音增强。本文主要的创新工作如下。

1) 基于语音样点间位置关系提出了含噪语音的无向图表示，在此基础上通过对应的 GFT 得到含噪语音的实数图频域特征，并构建基于 conformer 的实数网络 GFT-conformer，用于处理图频域特征以训练基于掩码的目标。

2) 考虑到语音和噪声在不同图频谱分量上的特性差异，提出扩展的可学习图比率掩码（extended learnable graph ratio mask, LGRM-E），对不同图频率分量的掩码范围分别进行控制，以更精细化地去除噪声，进一步提升 GFT-conformer 模型的增强性能。

3) 对所提出采用 LGRM-E 的 GFT-conformer 模型在 Voice Bank +DEMAND[28]数据集和 Deep Xi[29]数据集上进行了全面的评估。实验结果表明，所提出方法在五种常用评价指标上总体优于基于 DNN 的时域和时频域对比方案。

2 相关工作

2.1 图傅里叶变换

一个加权的图信号 $\boldsymbol{x}\in\mathbf{R}^{N\times 1}$ 可以由 $\mathcal{G}=(\mathcal{V},\boldsymbol{A})$ 来表示，其中 $\mathcal{V}=\{v_1,v_2,...,v_N\}$ 表示图上的 N 个节点集合，矩阵 $\boldsymbol{A}\in\mathbf{R}^{N\times N}$ 表示邻接矩阵或边权重矩阵[30]。$\boldsymbol{A}$ 中的第 i 行第 j 列的元素 A_{ij} 代表两个节点 v_i 和 v_j 之间的关联程度，若这两个节点之间存在关联，则 $A_{ij}\neq 0$，否则 $A_{ij}=0$。

为了更好地分析加权图信号 $\boldsymbol{x}$ 的图频域特性，$\boldsymbol{x}$ 的 GFT 定义为

$$\boldsymbol{X}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x} \tag{1}$$

其中 $\boldsymbol{U}$ 是将 $\boldsymbol{A}$ 进行特征值分解所得到的图傅里叶基[19]，即 $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}$，$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,...,\lambda_N)$ 表示特征值矩阵，特征值 $\lambda_1,\lambda_2,...,\lambda_N$ 表示图频率。相应地，$\boldsymbol{x}$ 的逆图傅里叶变换（inverse graph Fourier transform, IGFT）定义为

$$\boldsymbol{x}=\boldsymbol{U}\boldsymbol{X} \tag{2}$$

2.2 CRN 网络框架

卷积循环网络（convolutional recurrent network, CRN）[31]是一种主要用于语音增强的网络结构，它结合了 CNN 和循环神经网络（recurrent neural network, RNN）的特点。CRN 通常包括由 CNN 构成的编码器和解码器部分，以及由 RNN 构成的增强块。通过使用卷积编解码器从含噪语谱图中提取高维特征，并用 LSTM 层对其进行处理，以有效捕获时间信息，从而改善语音质量。为了提高 CRN 的相位感知能力，DCCRN（deep complex convolutional recurrent network）[16]采用 STFT 后的复数域时频谱作为网络的输入，并将 CRN 中的 CNN 和 RNN 从实数操作扩展到复数操作。这使得网络能够更好地处理幅度和相位信息，从而进一步提升了语音增强的性能。

3 本文方法

3.1 图频域特征

一段含噪语音信号 $\boldsymbol{y}$ 经过分帧之后可以表示为 $[\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_T]^{\mathrm{T}}\in\mathbf{R}^{T\times N}$，其中 T 和 N 分别表示语音帧数和每帧语音的样点数。将每个语音帧视为一个无向图，帧内的每个样点视为图上一个节点，则每个语音帧均可以构建为一个包含 N 个节点的图

信号。与文献[22]中定义的邻接矩阵不同，考虑到语音样点之间的关系通常随着样点间隔的增大而减弱，本文中邻接矩阵 $\boldsymbol{A}\in\mathbf{R}^{N\times N}$ 中第 i 行第 j 列的元素 A_{ij} 定义为

$$A_{ij}=\begin{cases}0, i=j\\ a_{ij}, i\neq j\end{cases}, a_{ij}=N-|i-j| \tag{3}$$

其中 $i,j=1,2,...,N$ 。

由于对每帧语音所定义的是一个无向图，则其对应邻接矩阵 $\boldsymbol{A}$ 是一个实对称矩阵。对 $\boldsymbol{A}$ 进行特征值分解，可得到正交的图傅里叶基 $\boldsymbol{U}$ ，再根据式（1），对各含噪语音帧对应的图信号进行 GFT，可以得到 $\boldsymbol{y}$ 对应的实数域图频谱 $\boldsymbol{Y}\in\mathbf{R}^{T\times G}$ ，其中 G 表示图频率索引数。在本文中，所描述的 GFT 是一种正交变换，不会丢失语音信号的有用信息。正交变换是可逆的，因此可以通过逆变换将变换后的信号完全恢复回原始域。

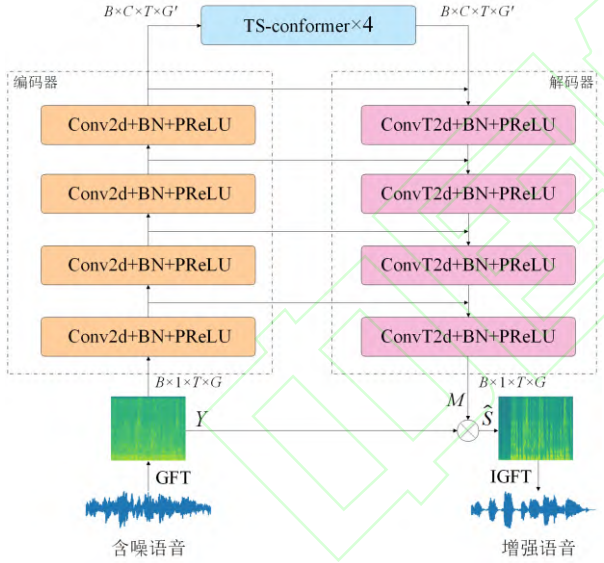

图 1　GFT-conformer 总体结构

Fig.1　Overall structure of the GFT-conformer

3.2　GFT-conformer 模型

如图 1 所示，GFT-conformer 与 CRN 类似，采用了编码器和解码器的结构，而增强块使用了 conformer 代替 CRN 中常用的 LSTM 层。Conformer 结合了 transformer 和 CNN 的优点，能够捕获长距离依赖关系，同时也能高效地利用局部特征。文献[27]提出两阶段 conformer 块，（two-stage conformer block, TS-conformer），分别对时频域特征的时间维度和频率维度进行捕获局部和全局的依赖关系。本文借鉴了 TS-conformer，采用 4 个连续的 TS-conformer，分别对图频域特征的时间维度和图频率维度进行建模。

编码器由 4 个卷积块组成，旨在从输入特征中提取高级特征并降低分辨率。随后，解码器将低分辨率特征映射回输入的维度，使得编码器和解码器形成对称的结构。编码器中的二维卷积块包括二维卷积（two-dimensional CNN, Conv2d）层，批量归一化（batch normalization, BN）和 PReLU 激活函数组成，而解码器中则使用反卷积（two-dimensional transposed CNN, ConvT2d）层代替 Conv2d 层。此外，编码器和解码器之间采用了跳过连接，以有利于梯度的流动，加速训练过程。

3.3　可学习的图比率掩码

在训练过程中，GFT-conformer 估计一个实数掩码 $\boldsymbol{M}\in\mathbf{R}^{T\times G}$ ，将其与含噪语音的图频谱 $\boldsymbol{Y}$ 对应位置元素相乘，得到增强语音的图频谱 $\boldsymbol{S}$ ，即

$$\hat{\boldsymbol{S}}=\boldsymbol{Y}\odot\boldsymbol{M} \tag{4}$$

进一步由式（2）定义的逆 GFT 即可得到增强语音 $\hat{\boldsymbol{s}}$ 。文献[22]基于 IRM 给出了一种用于图频谱的掩码，称为理想图比率掩码（ideal graph ratio mask, IGRM），其定义为

$$M_{t,g}=\frac{S_{t,g}}{Y_{t,g}} \tag{5}$$

其中，$Y_{t,g}$ 和 $S_{t,g}$ 分别表示第 t 帧含噪语音和干净语音在第 g 图频率索引下的图频谱值。

鉴于 IRM 的取值范围为 0 到 1，以及较大的取值范围可能会使 IGRM 的估计复杂化，本文提出一种可学习的图比率掩码（learnable graph ratio mask, LGRM），其定义为

$$\mathrm{LGRM}_{t,g}=k\cdot\tanh(c\cdot M_{t,g})+b \tag{6}$$

其中，k，c 和 b 均为可学习的参数，它们与网络模型的参数一样，在整个训练过程中通过反向传播一起更新和学习。在式（6）中，采用斜率 c 和 tanh 函数先将掩码的取值限制在-1 到 1 之间，再经过线性调整，将整体的掩码值控制在 $[-k+b,k+b]$ 之间。

图 2 展示了 Voice Bank+DEMAND 测试集中各条含噪语音与其对应的纯净语音计算得到的 IGRM 在不同图频率索引下的值分布，其中矩形上下的直线段表示 IGRM 值的范围，矩形的底边表示第一四

分位数，矩形的顶边表示第三四分位数，黄线表示IGRM值的中位数。从图2可以观察到，由于语音和噪声之间特性的差异，IGRM在不同图频率索引下的值分布均有所不同。因此，将LGRM扩展到每一个图频率索引，得到扩展的可学习图比率掩码（extended learnable graph ratio mask, LGRM-E），定义第t帧LGRM-E在第g图频率索引为

$$\text{LGRM-E}_{t,g} = k_g \cdot \tanh(c_g \cdot M_{t,g}) + b_g \tag{7}$$

其中，k_g，c_g和b_g为控制第g图频率索引掩码值范围的可学习参数。通过对各图频率分量训练最佳的掩码值范围，有望实现对不同图频率分量的细粒度去噪，进一步提升增强语音的质量。

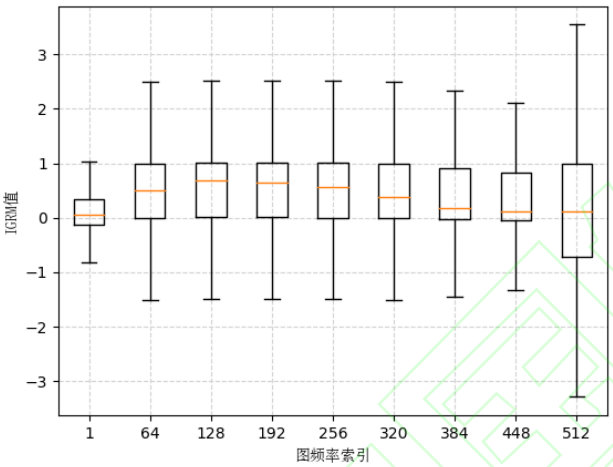

图2　IGRM在Voice Bank+DEMAND测试集中的值分布

Fig.2　The distribution of IGRM values in the Voice Bank+DEMAND test set

3.4　损失函数

本文使用尺度不变信噪比（scale-invariant source-to-noise ratio, Si-SNR）损失函数$L_{\text{Si-SNR}}$[32]来评估语音增强效果。$L_{\text{Si-SNR}}$作为一种时域损失函数，综合考虑了信号的幅度和相位信息，并且在处理多变的噪声信号和强度时具有较好的适应能力[33]。

由于没有在图频域直接计算损失值，增强语音的图频谱$\hat{\boldsymbol{S}}$和原始纯净语音图频谱$\boldsymbol{S}$之间存在差异[22]。为此，本文额外采用了一个基于客观语音质量评估（perceptual evaluation of speech quality, PESQ）的损失函数L_{PMSQE}[34]，以直接对增强语音的PESQ进行优化，有望进一步减少$\hat{\boldsymbol{S}}$和$\boldsymbol{S}$之间的差异。L_{PMSQE}定义为

$$L_{\text{PMSQE}} = \frac{1}{T}\sum_{t}(0.1D_t^{(s)} + 0.0309D_t^{(a)}) \tag{8}$$

其中，$D_t^{(s)}$和$D_t^{(a)}$分别表示文献[34]中定义的对称和非对称扰动，根据人耳的掩蔽效应在Bark谱中逐帧计算得到。

结合（8）和（9），构建总的损失函数L为

$$L = \gamma_1 L_{\text{Si-SNR}} + \gamma_2 L_{\text{PMSQE}} \tag{9}$$

其中，γ_1和γ_2为对应损失的权重系数。

4　实验数据及评估指标

4.1　数据集

为了验证所提方法的优越性，本文分别在Voice Bank+DEMAND[28]数据集和Deep Xi[29]数据集上进行实验。在Voice Bank+DEMAND数据集中，训练集包含来自28个说话人的11 572条纯净语音，与2种人工噪声和来自DEMAND[35]数据集的8种真实噪声录音混合。每个纯净语音片段都会随机选择一个噪声，以0 dB、5 dB、10 dB或15 dB的信噪比进行混合。训练集以9:1的比例进行分割，分别用于训练和验证。测试集包含来自2个说话人的824条纯净语音，与来自DEMAND数据集且未在训练集种使用的5种真实噪声录音之一混合，信噪比为2.5 dB、7.5 dB、12.5 dB或17.5 dB。

Deep Xi数据集包含69 708条纯净语音和17 458条噪声录音。从中随机选择1 000条纯净语音和噪声录音用于构建验证集，其余纯净语音和噪声录音用于训练。每条纯净语音都与一条随机选择的噪声录音以随机的信噪比进行混合，范围从-5 dB到20 dB，增量为1 dB。测试集包含200条含噪语音，由10个纯净语音片段与4种噪声录音，分别以-5 dB、0 dB、5 dB、10 dB和15 dB的信噪比进行混合得到。测试集中的四种噪声录音分别为嘈杂人声、街道音乐声、F16噪声和工厂噪声。

为了进一步验证所提出的方法对于不同口音说话人的鲁棒性，本文还在TIMIT[36]数据集上进行实验。TIMIT数据集包含来自八个不同地区的630个说话人的总共6300条纯净语音片段，这些语音片段涵盖了不同的口音、方言和语速变化。训练集包含4620条纯净语音，与来自DEMAND数据集的12种真实噪声录音进行混合，混合信噪比为0 dB、5 dB、10 dB或15 dB。训练集以9:1的比例进行分割，分别用于训练和验证。测试集包含来自1680

条纯净语音，与来自 DEMAND 数据集且未在训练集种使用的 4 种真实噪声录音进行混合，信噪比为 2.5 dB、7.5 dB、12.5 dB 或 17.5 dB。

在实验中，所有语音都以 16 000 Hz 的采样率重新采样，训练集和验证集的语音都被切成 2 s 的片段，而测试集的语音不进行任何裁剪，测试语音长度不固定。

表 1　在 Voice Bank+ DEMAND 数据集上的结果对比

Tab.1　Comparison of the results on the Voice Bank+DEMAND dataset

模型	分类	PESQ	CSIG	CBAK	COVL	STOI	参数/兆
Noisy	-	1.97	3.35	2.44	2.63	0.92	-
SADNUNet	时域	2.82	4.18	3.47	3.51	0.95	2.63
CleanUNet	时域	2.91	4.34	3.42	3.65	**0.956**	46.07
DEMUCS	时域	3.07	4.31	3.40	3.63	0.95	128
DCCRN	时频域	2.68	3.88	3.18	3.27	0.935	3.67
DCCRN+	时频域	2.84	-	-	-	-	3.3
S-DCCRN	时频域	2.84	4.03	2.97	3.43	0.94	2.34
DCTCN	时频域	2.83	3.91	3.37	3.37	-	9.7
CTS-Net	时频域	2.92	4.25	3.46	3.59	0.947	4.35
GaGNet	时频域	2.94	4.26	3.45	3.59	-	5.94
ResTCN+TFA-Xi	时频域	3.02	4.32	3.52	3.68	0.942	1.72
SA-TCN	时频域	3.02	4.29	3.50	3.67	0.944	9.91
DeepMMSE	时频域	3.03	4.35	3.52	3.71	0.941	-
iDeepMMSE	时频域	3.09	4.25	**3.56**	3.67	0.95	-
WMPNet	时域+时频域	3.05	4.27	3.53	3.68	-	7.63
WSFNet	时域+时频域	3.09	4.32	3.51	3.72	-	2.14
GFT-conformer	图频域	**3.14**	**4.50**	3.33	**3.89**	0.951	**1.40**

4.2　评估指标

本文采用五个广泛使用的评价标准，具体为 PESQ、短时可懂度（short-time objective intelligibility, STOI）和三个基于平均意见分（mean opinion score, MOS）的评价标准，对所提方法的性能进行全面评估。PESQ 的值范围在-0.5 到 4.5 之间，反映了语音信号的自然度和质量，更高的得分意味着更好的语音质量。STOI 用于衡量语音的可理解度，其值在 0 到 1 之间，更高的 STOI 值意味着更好的语音可理解度。基于 MOS 的评价标准包括以下三个指标：反映语音清晰度的 CSIG，较高的 CSIG 值表示较低的语音失真程度；用于评估背景噪声水平的 CBAK，较高的 CBAK 值表示背景噪声较少；表示整体听觉体验质量的 COVL，较高的 COVL 值表示更好的整体听觉体验。这三个指标的取值范围均都在 0 到 5 之间。

4.3　实验设置

实验语音使用 32 ms 的矩形窗函数进行分帧，重叠率为 75%，即帧移为 8 ms。对于损失函数权重系数，取 $\gamma_1 = 0.1$，$\gamma_2 = 0.95$。本文采用 AdamW 优化器进行训练，数据批大小设置为 4。初始学习率为 0.001，采用学习率衰减策略。如果连续三个 epoch 在验证集上的损失值超过最佳损失值，则将学习率减半。为了减轻过模型拟合，训练过程采用早停策略。如果在 10 个 epoch 内，在验证集上的损失值没有任何改善，则停止训练。在 GFT-conformer 模型中，编码器和解码器的卷积层通道大小为{64, 64, 64, 64}，卷积核大小为（5, 2），卷积步长为（2, 1）。

5　实验结果及分析

如表 1 所示，将本文提出的采用 LGRM-E 的 GFT-conformer 的模型其他用于语音增强的 DNN 模型在 Voice Bank+ DEMAND 数据集进行了性能比较。其中，SADNUNet[37]、CleanUNet[2] 和 DEMUCS[38]为时域方法，DCCRN[16]、DCCRN+[39]、S-DCCRN[40]、DCTCN[6]、CTS-Net[8]、GaGNet[7]、

ResTCN+TFA-Xi[41]、SA-TCN[5]、DeepMMSE[29]和iDeepMMSE[26]为时频域方法，WMPNet[42]和WSFNet[43]为时域和时频域结合的方法。从表 1 可以看出，与上述对比方案相比，本文所出的 GFT-conformer 整体上性能最优，其中 PESQ、CSIG 和 COVL 的得分均最高，STOI 值稍低于 CleanUNet。虽然 CBAK 得分略低于 WMPNet 等方法，但模型参数量远低于 WMPNet 等方法。这验证了本文所提出的采用 LGRM-E 的 GFT-conformer 模型在语音增强上的有效性，不仅能够提高增强语音的质量和可理解性，而且模型参数量小。

表 2　在 Deep Xi 数据集上的结果对比

Tab.2　Comparison of the results on the Deep Xi dataset

模型	PESQ	CSIG	CBAK	COVL	STOI
Noisy	1.24	2.26	1.81	1.67	0.78
DCCRN	1.94	3.10	2.19	2.54	0.88
DeepMMSE	1.97	3.30	2.62	2.59	0.87
iDeepMMSE	2.07	3.32	**2.72**	2.66	0.87
ResTCN+TFA-Xi	2.03	3.35	2.69	2.65	-
GFT-conformer	**2.18**	**3.55**	2.45	**2.88**	**0.896**

表 2 给出了本文提出的采用 LGRM-E 的 GFT-conformer 的模型与 DCCRN[16]、DeepMMSE[29]、iDeepMMSE[26]和 ResTCN+TFA-Xi[41]在 Deep Xi 数据集上的实验结果。如表 2 所示，本文提出的采用 LGRM-E 的 GFT-conformer 模型整体性能优于 DeepMMSE、iDeepMMSE 和 ResTCN+TFA-Xi，特别是 PESQ、CSIG 和 COVL，相比 iDeepMMSE 分别提高了 0.11、0.23 和 0.22。表 3 给出了本文提出的采用 LGRM-E 的 GFT-conformer 的模型与 DCCRN[16]和 ResTCN+TFA-Xi[41]在 Deep Xi 数据集上针对不同噪声类型和不同信噪比条件下的实验结果。如表 3 所示，本文提出的采用 LGRM-E 的 GFT-conformer 模型在不同噪声和不同信噪比条件下的 PESQ 得分均优于 ResTCN +TFA-Xi。以嘈杂人声噪声为例，相比 ResTCN +TFA-Xi，本文所提出的方法在 PESQ 得分上分别提高了 0.03、0.10、0.22、0.27 和 0.26，这进一步验证了本文所提出方法的有效性。

表 3　在 Deep Xi 数据集上不同噪声类型和不同信噪比下的 PESQ 对比

Tab.3　Comparison of the PESQ scores under different noise types and SNR levels on the Deep Xi dataset

噪声类型	信噪比/dB	Noisy	DCCRN	ResTCN+TFA-Xi	GFT-conformer
嘈杂人声	-5	1.07	1.18	1.22	**1.25**
	0	1.12	1.40	1.52	**1.62**
	5	1.23	1.81	1.93	**2.15**
	10	1.47	2.33	2.48	**2.75**
	15	1.89	2.83	2.95	**3.21**
街道音乐声	-5	1.03	1.20	1.24	**1.30**
	0	1.05	1.50	1.48	**1.67**
	5	1.10	1.94	1.86	**2.11**
	10	1.25	2.38	2.32	**2.58**
	15	1.56	2.85	2.77	**3.05**
F16 噪声	-5	1.04	1.30	1.38	**1.44**
	0	1.06	1.54	1.69	**1.80**
	5	1.11	1.91	2.10	**2.27**
	10	1.27	2.36	2.60	**2.78**
	15	1.58	2.77	3.01	**3.20**
工厂噪声	-5	1.05	1.21	1.26	**1.32**
	0	1.05	1.45	1.53	**1.64**
	5	1.10	1.81	1.93	**2.07**
	10	1.24	2.29	2.43	**2.51**
	15	1.52	2.69	2.83	**2.97**

表 4　在 TIMIT 数据集上针对不同地区说话人条件下的 PESQ 对比

Tab.4　Comparison of the PESQ scores across speakers from different regions on the TIMIT dataset

地区	Noisy	DCCRN	GFT-conformer
DR1	2.00	3.11	**3.26**
DR2	2.07	3.21	**3.34**
DR3	2.12	3.19	**3.38**
DR4	1.96	3.10	**3.20**
DR5	2.09	3.23	**3.35**
DR6	2.15	3.31	**3.43**
DR7	2.19	3.35	**3.46**
DR8	2.01	3.18	**3.39**
平均	2.07	3.21	**3.35**

表 4 给出了本文提出的采用 LGRM-E 的 GFT-conformer 的模型与 DCCRN[16]在 TIMIT 数据集上针对不同地区说话人条件下的 PESQ 结果。如表 4 所示，本文提出的采用 LGRM-E 的 GFT-conformer 模型在来自八个不同地区的说话人条件下均优于 DCCRN，PESQ 得分平均提升了 0.14。这表明本文所提出的方法对于不同说话人口音的数

据具有一定的鲁棒性。

5.1 不同前端特征对比及符号校正

为了验证本文提出的图频域特征在语音增强中的有效性，本小节实验比较了分别采用 DCT、STFT 和不同 GFT 来提取语音特征作为网络输入时的语音增强性能，在 Voice Bank+ DEMAND 数据集上实验结果如表 5 所示。考虑到文献[16]不仅提出了基于复数神经网络的 DCCRN，同时也给出了其相应的实数版本 DCCRN，因此本小节实验在 DCCRN 模型上进行对比，损失函数为 Si-SNR。在表 5 中，DCCRN-DCT 和 DCCRN-STFT 分别表示采用 DCT 和 STFT 提取语音时频域特征，DCCRN-GFT[22]表示提取图频域特征的 GFT 基由文献[22]中所定义的邻接矩阵进行特征值分解得到，DCCRN-GFT 表示提取的图频域特征的 GFT 基由本文式(3)定义的邻接矩阵进行特征值分解得到。值得提出的是，由于 GFT 和 DCT 所提取的特征是实数形式，因此 DCCRN-DCT、DCCRN-GFT[22]和 DCCRN-GFT 采用的是实数版本的 DCCRN，模型的参数量和计算复杂度均更低，训练参数设置与文献[16]中保持一致。此外，表 5 还给出了使用 GFT 得到的增强语音经过符号校正后的各项评价指标得分。增强语音的图频谱 $\boldsymbol{S}$ 经过符号校正后得到的图频谱 $\boldsymbol{S}_r$ 为

$$\boldsymbol{S}_r = \mathrm{abs}(\boldsymbol{S}) \odot \mathrm{sign}(\boldsymbol{S}) \tag{10}$$

其中，$\mathrm{sign}(\boldsymbol{S})$ 表示纯净语音图频谱 $\boldsymbol{S}$ 对应位置的正负符号所构成的矩阵。

表 5　不同前端特征在 DCCRN 模型上的结果对比

Tab.5　Comparison of different front-end features on the DCCRN model

模型	符号校正	PESQ	CSIG	CBAK	COVL	STOI	参数/兆	浮点数运算/吉
Noisy	-	1.97	3.35	2.44	2.63	0.92	-	-
DCCRN-DCT	-	2.43	3.71	2.64	3.09	0.933	2.16	10.9
DCCRN-STFT	-	2.68	3.88	3.18	3.27	0.937	3.67	22.2
DCCRN-GFT[22]	×	2.49	3.85	3.11	3.21	0.936	2.16	10.9
	√	2.63	4.00	3.25	3.36	0.943		
DCCRN-GFT	×	2.57	4.00	3.18	3.33	0.947	**2.16**	**10.9**
	√	**2.71**	**4.14**	**3.32**	**3.47**	**0.949**		

如表 5 所示，DCCRN-GFT 在五种评价指标 PESQ、CSIG、CBAK、COVL 和 STOI 上的得分均高于 DCCRN-DCT 和 DCCRN-GFT[22]。相较于 DCCRN-STFT，DCCRN-GFT 仍具优势，除 PESQ 外的其他四个评价指标均有所提升，这说明了采用本文所提出的 GFT 提取图频域特征，用于语音增强具有很好的潜力，能够大幅减少模型参数与计算量，且性能与传统的 STFT 相当。此外，从表 5 可以看出，经过符号校正后的 DCCRN-GFT 在所有五种评价指标中均取得了最高分，这也表明了采用 Si-SNR 作为损失函数的 DCCRN-GFT 得到增强语音图频谱 $\boldsymbol{S}$ 与纯净语音图频谱具有一定的差异，需要采取措施进一步减少两者的差异，提升增强语音的质量。

表 6 给出了本文提出的采用 LGRM-E 的 GFT-conformer 模型得到的增强语音及其经过符号校正后的各项性能指标。从表 6 的结果中可以观察到，相较于使用符号校正的情况，不使用符号校正的 PESQ 得分提升了 0.07，这是由于本文采用的损失函数能够直接对增强语音的 PESQ 进行优化，从而减少增强语音图频谱 $\boldsymbol{S}$ 和原始纯净语音图频谱 $\boldsymbol{S}$ 的差异，使得本文提出的方法不需要额外的符号校正。

表 6　符号校正在 GFT-conformer 模型上的影响

Tab.6　The impact of symbol correction on the GFT-conformer model

模型	符号校正	PESQ	CSIG	CBAK	COVL	STOI
GFT-conformer	×	**3.14**	**4.50**	3.33	**3.89**	0.951
	√	3.07	4.50	**3.44**	3.85	**0.952**

5.2 不同基于掩码的训练目标对比

为了验证本文提出的 LGRM-E 对 GFT-conformer 模型性能的进一步提升，本小节在 Voice Bank+ DEMAND 数据集上比较了 GFT-conformer 模型采用不同基于掩码训练目标的实验结果，如表 7 所示。其中，IGRM+tanh 表示在 IGRM 的基础上仅使用了 tanh 函数进行调整，将掩码的取值限制在 -1 到 1 之间。

表 7　不同基于掩码的训练目标在 GFT-conformer 模型上的对比

Tab.7　Comparison of different mask-based training objectives on the GFT-conformer model

模型	PESQ	CSIG	CBAK	COVL	STOI
Noisy	1.97	3.35	2.44	2.63	0.92
IGRM	2.98	4.37	3.00	3.73	0.949
IGRM+tanh	3.05	4.42	3.28	3.78	0.949
LGRM	3.09	4.49	3.32	3.86	0.949
LGRM-E	**3.14**	**4.50**	**3.33**	**3.89**	**0.951**

从表 7 可以看出，与 IGRM 相比，IGRM+tanh 和 LGRM 下的 PESQ 得分分别提高了 0.07 和 0.11，这表明在实际训练中，对掩码的取值范围进行适度的约束有助于网络进行训练，以实现更好的语音增强性能。同时，从表 7 中还可以看出，与 LGRM 相比，LGRM-E 下的 PESQ 得分提升了 0.05，这表明 LGRM-E 能够更精细化地去除噪声，进一步提升图频域语音增强的性能。

5.3　模型架构的超参数选择

本小节设计实验研究所提 GFT-conformer 模型中编码器和解码器卷积通道数的设置和增强块中 TS-conformer 的个数，对模型性能的影响。表 8 给出了编码器和解码器中不同卷积通道数下所提 GFT-conformer 模型的性能。从表 8 可以看出，随着卷积通道数的增大，PESQ、CSIG、CBAK 和 COVL 的得分逐渐提高，同时模型参数量也成倍增长。然而，卷积通道数为{128,128,128,128}相较于{64, 64, 64, 64}只提升了 0.02 的 PESQ 得分，但是模型参数量却增加至近 4 倍。因此，综合考虑模型性能和模型参数量，GFT-conformer 模型卷积通道数设置为{64, 64, 64, 64}。

表 9 给出了增强块中不同 TS-conformer 的个数下所提 GFT-conformer 模型的性能。如表 9 所示，随着 TS-conformer 个数的增多，模型的性能逐渐提高，当 TS-conformer 的个数大于 4 后，模型的性能未见进一步改善。因此，本文中 GFT-conformer 模型采用 4 个 TS-conformer 作为增强块。

表 8　不同卷积通道数对 GFT-conformer 模型的影响

Tab.8　The impact of different numbers of convolutional channels in the GFT-conformer model

卷积通道数	PESQ	CSIG	CBAK	COVL	参数/兆
Noisy	1.97	3.35	2.44	2.63	-
{32,32,32,32}	3.02	4.43	3.26	3.79	**0.38**
{64,64,64,64}	3.14	4.50	**3.33**	3.89	1.40
{128,128,128,128}	**3.16**	**4.51**	3.30	**3.90**	5.25

表 9　TS-conformer 个数对 GFT-conformer 模型的影响

Tab.9　The impact of the numbers of TS-conformers in the GFT-conformer model

TS-conformer 个数	PESQ	CSIG	CBAK	COVL	参数/兆
Noisy	1.97	3.35	2.44	2.63	-
1	2.93	4.22	2.86	3.59	**0.63**
2	3.09	4.47	3.34	3.85	0.89
3	3.11	4.48	3.30	3.86	1.15
4	**3.14**	**4.50**	3.33	**3.89**	1.40
5	3.14	4.48	**3.35**	3.88	1.66

6　结论

本文基于语音信号样点间的位置关系，将语音信号表示为无向图形式的图信号，利用对应 GFT 提取语音信号的图频域特征，并构建了 GFT-conformer 模型从输入特征的时间维度和图频率维度分别进行建模。同时，提出用于图频率语音增强的 LGRM，进一步提高了图频率语音增强性能。在 Voice Bank +DEMAND 数据集和 Deep Xi 数据集上的实验结果表明，与基于 DNN 的时域和时频域对比方案相比，本文所提出的方法不仅性能更优，而且参数量更少。

参考文献：

[1] GANNOT S, VINCENT E, MARKOVICH-GOLAN S, et al. A consolidated perspective on multimicrophone speech enhancement and source separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(4): 692-730.

[2] KONG Zhifeng, PING Wei, DANTREY A, et al. Speech denoising in the waveform domain with self-attention[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 7867-7871.

[3] KIM E, SEO H. SE-conformer: Time-domain speech enhancement using conformer[C]//Interspeech 2021. ISCA: ISCA, 2021: 2736-2740.

[4] LU Yexin, AI Yang, LING Zhenhua. MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra[C]//Interspeech 2023. ISCA: ISCA, 2023: 3834-3838.

[5] LIN Ju, DE LIND VAN WIJNGAARDEN A J, WANG K C, et al.

Speech enhancement using multi-stage self-attentive temporal convolutional networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3440-3450.

[6] REN Jigang, MAO Qirong. DCTCN: deep complex temporal convolutional network for long time speech enhancement[C]//Interspeech 2022. ISCA: ISCA, 2022: 5478-5482.

[7] LI Andong, ZHENG Chengshi, ZHANG Lu, et al. Glance and gaze: A collaborative learning framework for single-channel speech enhancement[J]. Applied Acoustics, 2022, 187: 108499.

[8] LI Andong, LIU Wenzhe, ZHENG Chengshi, et al. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1829-1843.

[9] XU Ziyi, ELSHAMY S, FINGSCHEIDT T. Using separate losses for speech and noise in mask-based speech enhancement[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 7519-7523.

[10] WANG Deliang. On ideal binary mask as the computational goal of auditory scene analysis[M]//Speech Separation by Humans and Machines. Boston: Kluwer Academic Publishers, 2006: 181-197.

[11] NARAYANAN A, WANG Deliang. Ideal ratio mask estimation using deep neural networks for robust speech recognition[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, BC, Canada. IEEE, 2013:7092-7096.

[12] WANG Yuxuan, NARAYANAN A, WANG Deliang. On training targets for supervised speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849-1858.

[13] ERDOGAN H, HERSHEY J R, WATANABE S, et al. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia. IEEE, 2015: 708-712.

[14] WILLIAMSON D S, WANG Yuxuan, WANG Deliang. Complex ratio masking for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(3): 483-492.

[15] CHOI H S, KIM J H, HUH J, et al. Phase-aware speech enhancement with deep complex U-net[EB/OL]. 2019: arXiv:1903.03107. https://arxiv.org/abs/1903.03107.pdf.

[16] HU Yanxin, LIU Yun, LV Shubo, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[C]//Interspeech 2020. ISCA: ISCA, 2020: 2472-2476.

[17] GENG Chuang, WANG Lei. End-to-end speech enhancement based on discrete cosine transform[C]//2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). Dalian, China. IEEE, 2020: 379-383.

[18] 徐峰, 李平. DVUGAN：基于 STDCT 的 DDSP 集成变分 U-Net 的语音增强[J]. 信号处理, 2022, 38(3): 582-589.
XU Feng, LI Ping. DVUGAN: DDSP integrated variational U-net speech enhancement based on STDCT[J]. Journal of Signal Processing, 2022, 38(3): 582-589. (in Chinese)

[19] SANDRYHAILA A, MOURA J M F. Discrete signal processing on graphs: Graph Fourier transform[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, BC, Canada. IEEE, 2013: 6167-6170.

[20] WANG Tingting, GUO Haiyan, YAN Xue, et al. Speech signal processing on graphs: The graph frequency analysis and an improved graph Wiener filtering method[J]. Speech Communication, 2021, 127: 82-91.

[21] 杨洋, 郭海燕, 王婷婷, 等. 基于联合时空图拓扑结构的多通道语音 MVDR 增强算法[J]. 信号处理, 2023, 39(3): 540-549.
YANG Yang, GUO Haiyan, WANG Tingting, et al. Multichannel speech MVDR enhancement algorithm based on joint spatial-temporal graph topology[J]. Journal of Signal Processing, 2023, 39(3): 540-549. (in Chinese)

[22] ZHANG Chenhui, PAN Xiang. Single-channel speech enhancement using graph Fourier transform[C]//Interspeech 2022. ISCA: ISCA, 2022: 946-950.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems (NIPS 2017). Los Angeles, USA, 2017: 5998-6008.

[24] YU Guochen, LI Andong, ZHENG Chengshi, et al. Dual-branch attention-in-attention transformer for single-channel speech enhancement[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 7847-7851.

[25] GULATI A, QIN J, CHIU C C, et al. Conformer: convolution-augmented transformer for speech recognition[C]//Interspeech 2020. ISCA: ISCA, 2020: 5036-5040.

[26] KIM M, SONG H, CHEONG S, et al. iDeepMMSE: An improved deep learning approach to MMSE speech and noise power spectrum estimation for speech enhancement[C]//Interspeech 2022. ISCA: ISCA, 2022: 181-185.

[27] CAO Ruizhe, ABDULATIF S, YANG Bin. CMGAN: conformer-based metric GAN for speech enhancement[C]//Interspeech 2022. ISCA: ISCA, 2022: 936-940.

[28] VALENTINI-BOTINHAO C, WANG Xin, TAKAKI S, et al.

Investigating RNN-based speech enhancement methods for noise-robust text-to-speech[C]//9th ISCA Workshop on Speech Synthesis Workshop (SSW 9). ISCA: ISCA, 2016: 146-152.

[29] NICOLSON A, PALIWAL K K. Deep learning for minimum mean-square error approaches to speech enhancement[J]. Speech Communication, 2019, 111: 44-55.

[30] SHUMAN D I, NARANG S K, FROSSARD P, et al. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains[J]. IEEE Signal Processing Magazine, 2013, 30(3): 83-98.

[31] TAN Ke, WANG Deliang. A convolutional recurrent neural network for real-time speech enhancement[C]//Interspeech 2018. ISCA: ISCA, 2018: 3229-3233.

[32] LE ROUX J, WISDOM S, ERDOGAN H, et al. SDR – half-baked or well done? [C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK. IEEE, 2019: 626-630.

[33] KOLBÆK M, TAN Zhenghua, JENSEN S H, et al. On loss functions for supervised monaural time-domain speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 825-838.

[34] MARTIN-DOÑAS J M, GOMEZ A M, GONZALEZ J A, et al. A deep learning loss function based on the perceptual evaluation of the speech quality[J]. IEEE Signal Processing Letters, 2018, 25(11): 1680-1684.

[35] THIEMANN J, ITO N, VINCENT E. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings[J]. The Journal of the Acoustical Society of America, 2013, 133(5): 3591.

[36] GAROFOLO J S, LAMEL L F, FISHER W M, 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.

[37] XIANG Xiaoxiao, ZHANG Xiaojuan, CHEN Haozhe. A nested U-net with self-attention and dense connectivity for monaural speech enhancement[J]. IEEE Signal Processing Letters, 2022, 29: 105-109.

[38] DÉFOSSEZ A, SYNNAEVE G, ADI Y. Real time speech enhancement in the waveform domain[C]//Interspeech 2020. ISCA: ISCA, 2020: 3291-3295.

[39] LV Shubo, HU Yanxin, ZHANG Shimin, et al. DCCRN+: channel-wise subband DCCRN with SNR estimation for speech enhancement[C]//Interspeech 2021. ISCA: ISCA, 2021: 2816-2820.

[40] LV Shubo, FU Yihui, XING Mengtao, et al. S-DCCRN: Super wide band DCCRN with learnable complex feature for speech enhancement[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 7767-7771.

[41] ZHANG Qiquan, QIAN Xinyuan, NI Zhaoheng, et al. A time-frequency attention module for neural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 462-475.

[42] XIANG Xiaoxiao, ZHANG Xiaojuan. Joint waveform and magnitude processing for monaural speech enhancement[J]. Applied Acoustics, 2022, 200:109077.

[43] YU Runxiang, CHEN Wenzhuo, YE Zhongfu. A novel target decoupling framework based on waveform-spectrum fusion network for monaural speech enhancement[J]. Digital Signal Processing, 2023, 141: 104150.

作者简介

王景润 男，2000 年出生，湖北咸宁人。南京邮电大学硕士生，主要研究方向为语音图信号处理。
E-mail: 1022010426@njupt.edu.cn

郭海燕 女，1983 年出生，湖北钟祥人。南京邮电大学副教授、硕士生导师，主要研究方向为语音信号处理与现代语音通信、协作通信、无线安全传输等。
E-mail: guohy@njupt.edu.cn

王婷婷 女，1992 年出生，安徽六安人。南京邮电大学讲师，主要研究方向为图信号处理、语音信号处理等。
E-mail: tingting_wang@njupt.edu.cn

 **杨震**　男，1961 年出生，江苏苏州人。南京邮电大学教授、博士生导师，主要研究方向为语音信号处理与现代语音通信、无线通信中的通信与信号处理技术等。

E-mail: yangz@njupt.edu.cn