全频带语音增强的局部频谱注意力

Zhongshu Hou[1,2,3], Qinwen Hu[1,2,3], Kai Chen[1,2,3], Jing Lu[1,2,3]

[1] 现代声学重点实验室，南京大学，中国南京 210093 [2] NJU-Horizon智能音频实验室，地平线机器人，中国北京 100094 [3] 南京先进人工智能研究院，中国南京 210014

{zhongshu.hou, qinwen.hu}@smail.nju.edu.cn, {chenkai, lujing}@nju.edu.cn

摘要

注意力机制已广泛应用于语音增强（SE）中，因为从理论上讲，它能够有效建模信号在时域和频谱域中的内在联系。通常，注意力在时域中的跨度受到限制，而频域中的注意力则覆盖整个频率范围。在本文中，我们注意到，覆盖整个频率范围的注意力会妨碍全频带语音增强的推理，并可能导致过多的残余噪声。为缓解这一问题，我们通过限制注意力跨度，将局部频谱注意力（LSA）引入全频带语音增强模型。在最先进水平（SOTA）的全频带语音增强模型上进行的消融测试表明，局部频率注意力能够有效提升整体性能。改进后的模型在全频带 VoiceBank+DEMAND 数据集上取得了最佳客观指标得分。

索引词— 局部频谱注意力、全频带语音增强、深度学习

1. 引言

语音增强（SE）对于电信系统、现代自动语音识别（ASR）和助听设备至关重要[1,2] , , 旨在提升受背景干扰影响而劣化的语音信号的感知质量、可懂度以及语音识别性能。近年来，基于数据驱动的深度神经网络（DNN）的语音增强方法已展现出显著优于传统基于规则的信号处理方法的性能，尤其是在具有高度非平稳噪声的低信噪比（SNR）条件下[3]。

基于时域深度神经网络的方法通过编码器-增强器-解码器框架直接增强原始语音波形 [4-6], , 而大多数语音增强方法则在时频（T-F）域中运行，以估计干净频谱与失真频谱之间的掩码 [7,8], , 或根据带噪语音直接预测目标干净复频谱的实部和虚部 [9,10]。时频模型采用了不同的网络结构来提取时域和频域特征，包括卷积神经网络（CNN） [9-15], 、循环神经网络（RNN） [14-18], 和自注意力（SA）机制 [19-26]。

近年来，基于自注意力机制的模型在语音增强任务中展现出优越性，在宽带（采样率为 16kHz）[25] （CMGAN）和全频带（采样率为 48kHz）[26] （MTFAA）场景中均达到最先进水平。时间注意力通过指定在线处理时参与注意力计算的帧数，或限制离线处理时训练序列的最大长度，高效地对有限范围内的语音时间依赖关系进行建模；而频率维度的注意力始终作用于整个频率范围，在全局范围内计算注意力权重。本文发现，全局频谱注意力会妨碍全频带语音增强的推理，并可能导致残余噪声过多。这可能是由于语音不同频带之间存在特征差异所致，即低频带中的结构性谐波与高频带中近乎随机分布的成分之间存在差异。这种较弱的相似性可能导致注意力机制中产生大量冗余，使语音增强模型无法聚焦于关键特征，进而造成性能不佳。

自然语言处理中的局部注意力机制在[27], 中被引入，其每次关注一部分源词，以避免为实现有效的机器翻译而进行昂贵的计算。我们发现，该机制有潜力在某一频带内通过有限范围促进频谱建模。本文提出一种局部频谱注意力（LSA）机制，该机制仅关注某一频率处的相邻频带，以缓解全频带注意力带来的问题。我们基于 MTFAA 模型和双路径注意力循环网络（DPARN） [20] 开展实验，验证了所提出 LSA 机制的优势，并在公开的全频带 Voice-Bank+DEMAND [28] 数据集上，通过改进的 MTFAA 取得了最先进水平（SOTA）的结果。

2. 方法描述

2.1. 频谱局部注意力

令 $\mathcal{X} \in \mathrm{R}^{T\times F\times 2}$ 表示具有逐频率注意力的时频模型的输入频谱图，其中 T 和 F 分别表示时间维度和频率维度。则传递给传统频谱注意力模块的键、查询和得分张量可表示为 $\mathcal{Q}_F$、$\mathcal{K}_F$、$\mathcal{V}_F \in \mathrm{R}^{T\times F'\times C}$，其中 F' 表示经过卷积神经网络下采样或

上采样后的逐频率维度， C 表示通道数。则传统注意力输出 $\mathcal{A}_F \in \mathrm{R}^{T\times F'\times C}$ 可表示为

$$\mathbf{A}_{F,t} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{F,t}\mathbf{K}_{F,t}^{\mathrm{T}}}{\sqrt{F'\times C}}\right)\mathbf{V}_{F,t}, \tag{1}$$

$$\mathcal{A}_F = \mathrm{concat}\left(\mathbf{A}_{F,1}, \mathbf{A}_{F,2}, \ldots, \mathbf{A}_{F,T}\right), \tag{2}$$

其中，$\mathbf{Q}_{F,\ t}$、$\mathbf{K}_{F,\ t}$、$\mathbf{V}_{F,\ t} \in \mathrm{R}^{F'\times C}$ 分别表示帧 t处的键、查询和得分矩阵，concat(∗)表示沿时间维度拼接输出注意力矩阵。由$\mathbf{Q}_{F,\ t}$ 与$\mathbf{K}_{F,\ t}^{\mathrm{T}}$ 的点积计算得到的相似度矩阵构成注意力机制[19], 的权重，反映了整体频带之间的相关性。我们认为，由于低频带与高频带之间的特征分布相似度较低，长距离频谱相关性估计可能会误导模型使用无关信息，在全频带场景下尤其如此。为有效提取局部频谱依赖，我们将局部注意力掩码 $\mathbf{M}_F \in \mathrm{R}^{F'\times F}$，应用于相似度矩阵，其定义如下：

$$M_{F,i,j} = \begin{cases} 0, & |i-j| \le N_l \\ -\infty, & |i-j| > N_l' \end{cases} \tag{3}$$

其中，$M_{F,i,j}$ 表示 $\mathbf{M}_F$ 在第 i^{th} 行、第 j^{th} 列的元素，$N_l \le F'$ 为局部交互维度。因此，所提出的 LSA 输出 $\tilde{\mathcal{A}}_F$ 可计算为

$$\widetilde{\mathbf{A}}_{F,t} = \mathrm{softmax}\left(\frac{\mathbf{Q}_{F,t}\mathbf{K}_{F,t}^{\mathrm{T}}}{\sqrt{F'\times C}} + \mathbf{M}_F\right)\mathbf{V}_{F,t}, \tag{4}$$

$$\tilde{\mathcal{A}}_F = \mathrm{concat}\left(\widetilde{\mathbf{A}}_{F,1}, \widetilde{\mathbf{A}}_{F,2}, \ldots, \widetilde{\mathbf{A}}_{F,T}\right). \tag{5}$$

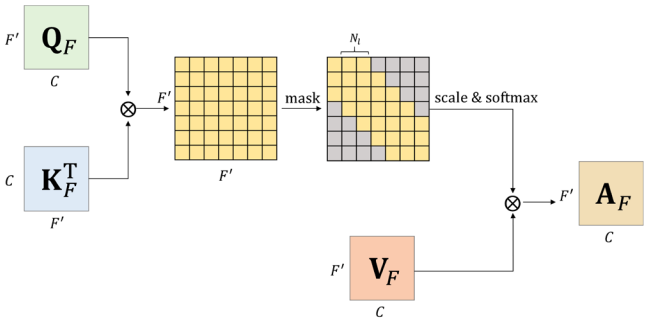


图1。所提出的局部频谱注意力机制的可视化示意图。

所提出的局部频谱注意力机制的可视化结果如图1所示。在实际网络实现中，$\mathbf{M}_F$ 的 $-\infty$ 可以设为较大的负值，例如 -1e9。所提出方法的源代码可在 *https://github.com/ZhongshuHou/LSA*获取。

2.2. 模型架构

MTFAA 网络是4th DNS 挑战赛中表现最佳的全频带模型，该挑战赛于 ICASSP 2022 [26], 举办。该模型采用逐频率自注意力进行频谱建模。它包含一个相位编码器模块，用于将复数频谱特征映射为实数分量。随后，在频带合并模块中，根据 ERB尺度映射实数频谱输入，以减少高频处的冗余。频带拆分是频带合并的逆过程。主网络由若干相似的处理单元构成，其中频率下采样（FD）或上采样（FU）通过二维卷积实现。时频卷积模块采用空洞卷积，以扩大时域上的感受野；轴向自注意力（ASA）则沿时间轴和频率轴计算注意力矩阵，以进行时频特征建模。在掩码估计与应用流程中，第一阶段将深度滤波器掩码应用于幅度频谱，随后通过复数掩码进一步细化幅度频谱和相位频谱。关于 MTFAA 的详细介绍见[26]。在原始的轴向自注意力模块中，注意力以全局方式沿频率方向运行。我们将 ASA 中逐频率的全局注意力替换为所提出的局部频谱注意力机制，并在第3节中对模型性能进行消融研究。

为了进一步验证 LSA 的有效性，我们还测试了它在另一种名为 DPARN 的全频带语音增强模型上的表现。该模型源自 DPCRN（在 DNS-3 中排名第 3rd），[17] 但用多头自注意力取代了内部循环神经网络模块，以进行频谱建模。

2.3. 训练目标和损失函数

幂压缩损失 [29] 函数用于更好地处理低功率时频点中的信息：

$$\mathcal{L}_{RI}(\tilde{S}, S) = \left\|S_{real}^{\mathcal{C}} - \tilde{S}_{real}^{\mathcal{C}}\right\|_F^2 + \left\|S_{imag}^{\mathcal{C}} - \tilde{S}_{imag}^{\mathcal{C}}\right\|_F^2, \tag{6}$$

$$\mathcal{L}_{Mag}(\tilde{S}, S) = \left\||S|^{\gamma} - |\tilde{S}|^{\gamma}\right\|_F^2, \tag{7}$$

$$\mathcal{L}_{Ovrl}(\tilde{S}, S) = \mathcal{L}_{Mag}(\tilde{S}, S) + \mathcal{L}_{RI}(\tilde{S}, S), \tag{8}$$

$$S_{real}^{\mathcal{C}} = |S|^{\gamma}cos\theta_S, \qquad S_{imag}^{\mathcal{C}} = |S|^{\gamma}sin\theta_S, \tag{9}$$

其中，S 和 $\tilde{S}$ 分别表示时频域中的目标干净频谱图和增强后的频谱图，γ 表示压缩参数，上标 $\mathcal{C}$ 表示幂压缩形式，θ_S 表示复数

频谱图的相位角，而 $||\cdot||_F$ 表示矩阵的弗罗贝尼乌斯范数。MTFAA 网络和 DPARN 均通过总体损失函数 L$_{ovrl}$$\hat{S}, S$进行优化。

3. 实验

3.1. 数据集

我们在常用的公开全频带 VoiceBank- +DEMAND 数据集 [28]上验证了所提出方法的有效性。干净语音片段选自 VoiceBank 语料库 [30]，训练集中包含 28 名说话人，测试集中包含 2 名未见过的说话人。训练集包含 11,572 条语音，测试集包含 872 条语音。在训练集中，干净语音与背景噪声（来自 DEMAND 数据库 [31] 的 8 类真实噪声录音和 2 类人工噪声）以 0 dB、5 dB、10 dB 和 15 dB 的信噪比混合。对于测试集，干净语音则叠加了来自 DEMAND 数据库的 5 类未见噪声，信噪比分别为 2.5 dB、7.5 dB、12.5 dB 和 17.5 dB。所选噪声类型大多具有挑战性，例如家庭噪声（客厅和厨房）、公共场所噪声（餐厅、咖啡厅和办公室）以及交通/街道噪声（汽车、公交车、地铁、繁忙路段、地铁站和公共广场）。在我们的实验中，所有语音均以 48 kHz 采样。

3.2. 参数设置与训练策略

短时傅里叶变换（STFT）的窗长和帧移分别为 32 毫秒和 8 毫秒。离散傅里叶变换长度为 1536，并使用汉宁窗进行重叠相加波形重建。MTFAA 中的局部交互维度 N_l [8, 4, 2, 2, 2, 4, 8, 16] 在每个 ASA 模块中分别设为 N_l，DPARN 中的局部交互维度则在每个多头自注意力模块中设为 64。MTFAA 和 DPARN 的其他参数设置分别见 [26] 和 [20]，。

我们的训练批量大小为 4，输入音频的时间长度为 4 秒。预热策略 [19] 对于训练基于自注意力的模型至关重要，其中学习率α 按照以下规则更新：$\alpha = \frac{1}{\sqrt{C}} \times \min(\varphi^{-1}\sqrt{\varphi}, \varphi\sqrt{\Psi^3})$，其中 $C = 64$、预热步数 $\Psi = 3000$ 和 φ 表示训练步数。我们使用基于预热的 Adam 优化器训练模型，参数为 $\beta_1 = 0.9$、$\beta_2 = 0.98$、$\epsilon = 10^{-9}$。压缩参数 γ–为 1 3。

3.3. 消融研究与评估指标

为验证所提出的 LSA 相较于传统全局注意力机制的有效性，我们在因果型 MTFAA 和 DPARN 模型上进行了消融研究，其中因果性由沿时间维度的填充膨胀卷积、掩码注意力和单向循环神经网络实现。为全面评估性能，我们还将修改后的 MTFAA 模型与先前的全频带最先进方法在 VoiceBank+DEMAND 测试集上进行比较。

我们选取了一组常用的语音指标来评估综合性能，即语音质量感知评估（PESQ）[32], 、短时客观可懂度（STOI）[33], 、尺度不变信号失真比（SiSDR）以及基于综合平均意见分（MOS）的指标 [34]: 信号失真的 MOS 预测（CSIG）、背景噪声侵扰程度的 MOS 预测（CBAK）以及总体效果的 MOS 预测（COVL）。所有指标的值越高，表示性能越好。由于缺少针对全频带语音的 PESQ 和综合 MOS 评估，我们将增强后的结果降采样至 16 千赫兹（宽带）并计算分数，而 STOI 和 SiSDR 则基于全频带结果进行测量。

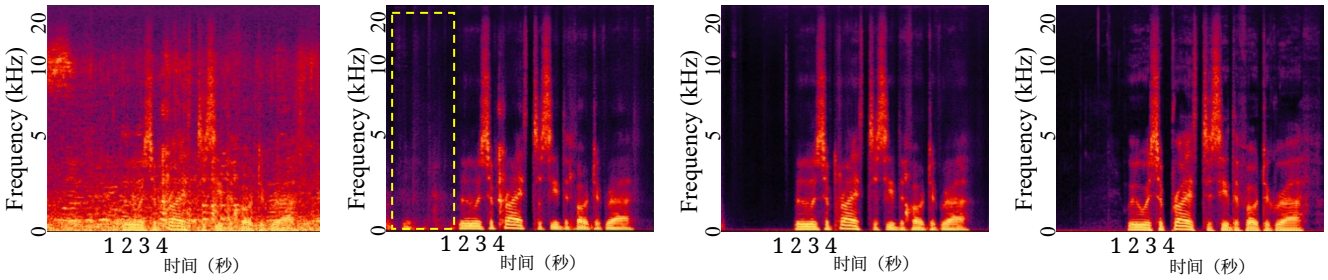
3.4. 实验结果与分析

3.4.1. 消融研究

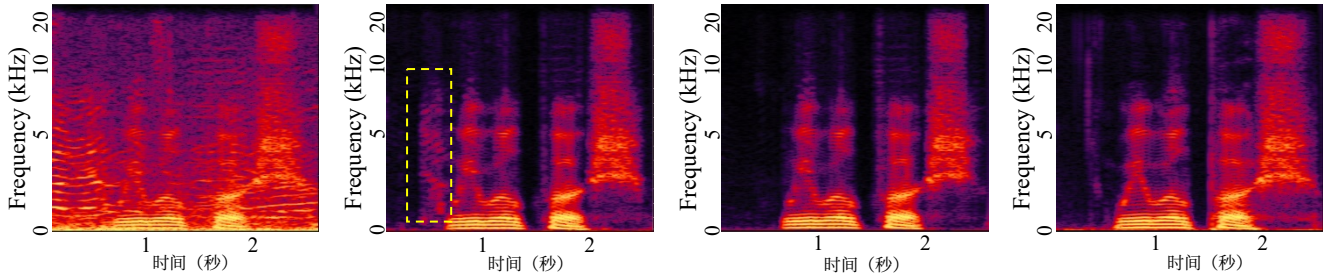
消融研究结果如表 1 所示。所提出的 LSA 在所有客观指标上均提升了因果 DPARN 和 MTFAA 模型的增强性能。进一步考察增强后的信号发现，频域中的全局注意力更容易产生过多的残余噪声，尤其是在非语音片段中，而所提出的 LSA 能有效缓解这一问题。图 2 展示了两个典型示例，可以清楚地看到 LSA 带来的益处。一种可能的解释是，局部注意力能够更充分地利用频谱中的语音模式，并更有效地区分语音和噪声成分，尤其是在低信噪比环境中。

表 1. 消融研究结果

配置。		宽带指标				全频带指标	
模型 LSA	PESQ	CSIG	CBAK	COVL	短时客观可懂度(%)	SiSDR(dB)	
MTFAA	X	3.13	4.33	3.54	3.75	94.6	17.7
	√	**3.16**	**4.35**	**3.61**	**3.78**	**94.7**	**18.8**
DPARN	X	2.92	4.26	3.61	3.65	94.2	18.3
	√	**2.96**	**4.29**	**3.63**	**3.68**	**94.2**	**18.7**



(a) 含噪 (b) 原始 MTFAA (c) 基于 LSA 的 MTFAA (d) 干净



(e) 含噪 (f) 原始 DPARN (g) 基于 LSA 的 DPARN (h) 干净

图 2. 典型频谱图。(a,e) 带噪语音，(b,f) 原始非 LSA 网络增强后的语音，(c,g) 基于 LSA 的网络增强后的语音，(d,h) 干净的参考语音

表 2.客观指　　标：VoiceBank+DEMAND 测试集上的结果。“-”表示原论文未报告的分数　　rs

模型	年份	参数量（M）	宽带指标				全频带指标
			PESQ	CSIG	CBAK	COVL	STOI（%）
含噪	-	-	1.97	3.34	2.44	2.63	92.1
S-DCCRN[35]	2022	2.34	2.84	4.03	3.43	2.97	94.0
FullSubNet+[36]	2022	8.67	2.88	3.86	3.42	3.57	94.0
GaGNet[37]	2022	5.95	2.94	4.26	3.45	3.59	-
DMF-Net[38]	2022	7.84	2.97	4.26	3.52	3.62	94.4
SF-Net[13]	2022	6.98	3.02	4.36	3.54	3.67	94.5
DeepFilterNet2[39]	2022	2.31	3.08	4.30	3.40	3.70	94.3
MTFAA（因果，LSA）	2022	1.50	**3.16**	**4.35**	**3.61**	**3.78**	**94.7**

3.4.2. 与全频带最先进方法的比较

我们在 VoiceBank+DEMAND 数据集上，将修改后的 MTFAA 模型与以往的全频带最先进方法进行了比较，结果列于表 2。可以看出，修改后的因果 MTFAA 模型在所有评估指标上均优于以往的最先进方法，PESQ 得分达到 3.16，短时客观可懂度值接近 95%。此外，由于在频域中采用局部处理，LSA 机制所需的参数更少，计算复杂度也更低。

4. 结论

本文提出一种面向基于注意力的全频带语音增强模型的局部频谱注意力机制，有助于更有效地提取频谱信息。我们将所提出的 LSA 融入包括 DPARN 和 MTFAA 在内的全频带模型，并验证了该方法在恢复语音时能取得更好的效果，同时减少残余噪声。搭载 LSA 的 MTFAA 在 VoiceBank+DEMAND 测试集上也取得了最佳性能。此外，所提出的方法还带来了更小的模型和更低的计算负担。

5. 致谢

本研究得到国家自然科学基金委员会资助（项目编号：11874219和12274221）。

6. 参考文献

[1] F. Weninger和H. Erdogan，“利用LSTM循环神经网络进行语音增强及其在抗噪语音识别中的应用”，载于国际潜变量分析与信号分离会议。施普林格，2015年，第91–99页。 [2] J. L. Desjardins和K. A. Doherty，“助听器降噪对听力受损成年人聆听努力的影响”，《耳与听力》，第35卷，第600–610页，2014年。[3] DeLiang Wang和陈吉通，“基于深度学习的监督式语音分离：综述”，*IEEE/ACM*音频、语音与语言处理汇刊，第26卷，第10期，第1702–1726页，2018年。[4] 罗毅和N. Mesgarani，“Conv-TasNet：超越语音分离的理想时频幅度掩码”，*IEEE/ACM*音频、语音与语言处理汇刊，第27卷，第8期，第1256–1266页，2019年。[5] A. Defossez、G. Synnaeve和Y. Adi，“波形域实时语音增强”，载于 *Interspeech*，2020年，第 3291–3295 页。[6] E. Kim 和 H. Seo，《SE-Conformer：使用 Conformer 的时域语音增强》，载于 *Interspeech*，2021 年，第 2736–2740 页。[7] D. S. Williamson、王宇轩和 DeLiang Wang，《用于单耳语音分离的复数比率掩码》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 24 卷，第 3 期，第 483–492 页，2015 年。[8] C. Hummersone、T. Stokes 和 T. Brookes，《以理想比率掩码为计算听觉场景分析目标》，载于盲源分离，第 349 368 2014 – 页。施普林格，Ke Tan 和 DeLiang Wang，《用于实时语音增强的卷积循环神经网络》，载于 *Interspeech*，2018 年，第 2018 卷，第 3229–3233 页。[10] Andong Li、刘文哲和罗晓雪，《一种同步去噪与去混响框架及目标解耦方法》，载于 *Interspeech*，2021 年，第 2801–2805 页。[11] A. Pandey 和 DeLiang Wang，《TCNN：用于时域实时语音增强的时间卷积神经网络》，载于 *ICASSP*. IEEE，2019 年，第 6875–6879 页。[12] 杨显、杨孙，《用于单声道语音增强的卷积融合网络》，《神经网络》，第 143 卷，第 97–107 页，2021 年。[13] Guochen Yu、Andong Li，《协同优化：用于实时全频带语音增强的协同子带融合模型》，*arXiv* 预印本 *arXiv:2203.16033*，2022 年，[14] Ke Tan 和 DeLiang Wang，“使用卷积循环网络进行复数谱映射的单声道语音增强”，载于 *ICASSP*。IEEE，2019 年，第 6865–6869 页。[15] Yanxin Hu、Yun Liu，“Dccrn：用于相位感知语音增强的深度复数卷积循环网络”，*arXiv* 预印本 *arXiv:2008.00264*，2020 年。[16] Zhong-Qiu Wang、S. Cornell，“Tf-gridnet：让时频域模型再次适用于单声道说话人分离”，*CoRR*，第 abs/2209.03952 卷，2022 年。[17] Xiaohuai Le、Hongsheng Chen，“DPCRN：用于单通道语音增强的双路径卷积循环网络”，载于 *Interspeech*，2021 年，第 2811–2815 页。[18] Shengkui Zhao、Bin Ma，“Frcrn：利用频率递归增强单声道语音增强中的特征表示”，载于 *ICASSP*。IEEE，2022 年，第 9281–9285 页。[19] A. Nicolson 和 K. K. Paliwal，“用于因果语音增强的掩码多头自注意力”，*Speech Communication*，第 125 卷，第 80–96 页，2020 年。[20] Qinwen Hu、Zhongshu Hou，“轻量级全频带语音增强模型”，*ArXiv*，第 abs/2206.14524 卷，2022 年。[21] C. Subakan、M. Ravanelli，“语音分离中所需的一切皆为注意力”，载于 ICASSP。IEEE，2021 年，第 21–25 页。[22] A. Pandey 和 DeLiang Wang，“用于实时语音增强的双路径自注意力 RNN”，arXiv 预印本 arXiv:2010.12713，2020 年。[23] A. Pandey、Buye Xu，“Tparn：用于时域多通道语音增强的三路径注意力循环网络”，载于 ICASSP。IEEE，2022 年，第 6497–6501 页。[24] Jngjing Chen、Qirong Mao 和 Dong Liu，“双路径 Transformer 网络：端到端单声道语音分离的直接上下文感知建模”，arXiv 预印本 arXiv:2007.13975，2020 年。[25] S. Abdulatif、Ruizhe Cao 和 Bin Yang，“Cmgan：基于 Conformer 的度量 GAN，用于单声道语音增强”，arXiv 预印本 arXiv:2209.11112，2022 年。[26] Guochang Zhang、Libiao Yu、Chunliang Wang 和 Jianqiang Wei，“带轴向注意力的多尺度时频卷积网络，用于语音增强”，载于 ICASSP。IEEE，2022 年，第 9122–9126 页。[27] T. Luong、H. Pham 和 C. D. Manning，“基于注意力的神经机器翻译的有效方法”，arXiv 预印本 Xin W ”，载于 Chengshi Zheng，“功率压缩和相位估计在单声道语音去混响中的重要性”，JASA Express Letters，第 1 卷，014802，2021 年 1 月，无页码。[30] C. Veaux、J. Yamagishi 和 S. King，“Voice Bank 语料库：大型区域口音语音数据库的设计、收集与数据分析”，载于与 2013 年亚洲口语语言研究与评估会议联合举办的国际东方 CO-COSDA 会议。IEEE，2013 年，第 1–4 页。[31] J. Thiemann、N. Ito 和 E. Vincent，“多样环境多通道声学噪声数据库（DEMAND）：多通道环境噪声录音数据库”，Journal of the Acoustical Society of America，第 133 卷，第 3591–3591 页，2013 年。[32] A. W. Rix、J. G. Beerends 和 M. P. Hollier，“语音质量感知评估（PESQ）——一种用于电话网络和编解码器语音质量评估的新方法”，载于 ICASSP。IEEE，2001 年，第 2 卷，第 749–752 页。[33] C. H. Taal、R. C. Hendriks 和 R. Heusdens，“一种用于时频加权噪声语音的短时客观可懂度度量”，载于 ICASSP。IEEE，2010 年，第 4214–4217 页。[34] Yi Hu 和 P. C. Loizou，“语音增强客观质量度量的评估”，IEEE Transactions on Audio, Speech, and Language Processing，第 16 卷，第 1 期，第 229–238 页，2007 年。[35] Shubo Lv、Yihui Fu，“S-dccrn：带可学习复数特征的超宽带 DCCRN，用于语音增强”，载于 ICASSP。IEEE，2022 年，第 7767–7771 页。[36] Jun Chen、Zilin Wang，“Fullsubnet+：带复数频谱图的通道注意力FullSubNet，用于语音增强”，载于 ICASSP。IEEE，2022 年，第 7857–7861 页。[37] Andong Li、Chengshi Zheng，“Glance and Gaze：用于单通道语音增强的协作学习框架”，Applied Acoustics，第 187 卷，第 108499 页，2022 年。[38] Guochen Yu、Yuansheng Guan，“Dmf-net：用于实时全频带语音增强的解耦式多频带融合模型”，arXiv 预印本 arXiv:*2203 00472*, 2022 。[39] H. Schroter、T. Rosenkranz，“Deepfilternet：面向嵌入式设备全频带音频实时语音增强”，arXiv 预印本 arXiv:2205.05474，2022 年。