交互- SUBNET：采用子带交互的语音增强 N

Jun Chen[1, 2, †], *Wei Rao*[2, *], *Zilin Wang*[1], *Jiuxin Lin*[1], *Zhiyong Wu*[1, 3, *], *Yannan Wang*[2], *Shidong Shang*[2], *Helen Meng*[1,3]

[1]清华大学深圳国际研究生院，中国深圳[2]腾讯天籁实验室，腾讯，中国深圳[3]香港中文大学，中国香港特别行政区y-chen21@mails.tsinghua.edu.cn, ellenwrao@tencent.com, zywu@sz.tsinghua.edu.cn

arXiv:2305.05599v1 [cs.SD] 9 May 2023

摘要

基于子带的方法通过共享参数在模型中并行处理子带，以学习局部频谱的共性，从而实现降噪。通过这种方式，它们以更少的参数取得了显著成果。然而，在一些复杂环境中，缺乏全局频谱信息会对这些基于子带的方法的性能产生负面影响。为此，本文引入子带交互，作为一种利用跨频带依赖关系和全局频谱模式等全局频谱信息补充子带模型的新方法，并提出了一种名为交互式子带网络（Inter-SubNet）的新型轻量级单通道语音增强框架。在 DNS 挑战赛 - Interspeech 2021 数据集上的实验结果表明，所提出的 Inter-SubNet 相较于子带模型取得了显著提升，并优于其他最先进的语音增强方法，证明了子带交互的有效性。

索引词— 子带交互、全局频谱信息、Inter-SubNet、语音增强

1. 引言

环境噪声造成的干扰是妨碍助听器、音频通信和自动语音识别的主要因素之一。因此，人们迫切需要单通道语音增强方法，以消除含噪语音中的干扰噪声，并提高语音的清晰度和可理解度。传统语音增强方法利用统计信号理论有效抑制平稳噪声，但在低信噪比（SNR）条件下表现不佳。近年来，借助深度神经网络的强大能力，基于深度学习的语音增强方法取得了卓越成果，尤其是在应对低信噪比、混响等具有挑战性的情形时。深度神经网络可以在时域或频域对含噪音频进行降噪。时域方法 [1–5]直接从带噪音频波形生成纯净语音波形。相比之下，频域方法 [6–9]通常采用含噪频谱图特征作为模型的输入特征，其预测目标则是纯净频谱图特征或频谱图掩码，包括理想二值掩码（IBM） [10], 、理想比率掩码（IRM） [11]和复数理想比率掩码（cIRM） [12], 仅举几例。总体而言，考虑到系统的鲁棒性和计算复杂度，频域方法更受青睐 [13]。

基于子带的方法 [14, 15] 对单通道频域语音增强十分有效。这类方法将输入频谱图划分为多个子带，然后通过共享参数的模型并行处理这些子带，这有助于模型学习局部频谱的共性以及逐频信号平稳性，从而抑制噪声 [14]。因此，这些基于子带的方法在保持模型性能的同时，减少了参数数量和计算负担。

然而，由于它们仅对局部信息敏感，当遇到局部频谱模式不明确的复杂声学环境时，基于子带的模型很难恢复纯净语音。借助全局频谱信息可以应对这一挑战，例如长距离跨频带依赖关系和全局频谱模式 [16]: 模型能够根据全局频谱的模式以及其他相对纯净的子带，推断当前子带的局部模式。FullSubNet [16] 和 FullSubNet+ [17]通过将全频带模型与子带模型串联，为子带模型提供全局频谱信息。它们在降噪任务中表现出色，这证明了为子带模型补充全局频谱信息的价值。然而，由于引入了全频带模型，这些方法导致模型整体参数数量大幅增加，同时也提高了模型的复杂度。

本文提出了一种新颖的子带交互（SubInter）模块，用于语音增强，以在保持模型复杂度较低的同时提取全局频谱信息，例如跨频带依赖关系和全局频谱模式。我们将新设计的 SubInter 模块集成到子带模型中 [14] ，以补充模型的全局频谱信息。这种新型轻量级语音增强框架称为 Inter-SubNet，由交替排列的 SubInter 模块和长短期记忆网络（LSTM）层组成，能够在保留关注局部频谱模式能力的同时捕获全局频谱信息。在 DNS 挑战赛 - Interspeech 2021 数据集上的实验结果表明，与子带模型相比，我们的 Inter-SubNet 性能大幅提升。此外，无论是参数数量还是性能，Inter-SubNet+ 均优于 FullSubNet，也超过了其他先进的语音增强方法。这些结果表明，子带交互是一种用于建模全局频谱信息的轻量且高效的方法。

† 第一作者在腾讯实习期间完成的工作。

* 通信作者。

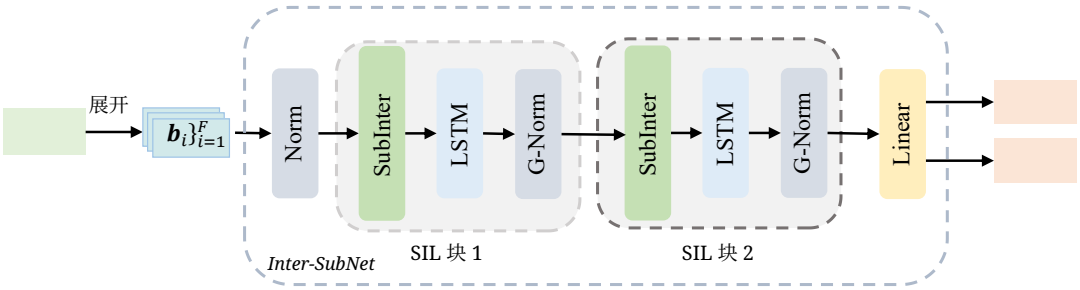


图 1。所提出的 Inter-SubNet 的总体示意图，其中，“SIL 块 1”和“SIL 块 2”分别指第一个和第二个 SubInter-LSTM 块。“G-norm”表示组归一化 [18]。Inter-SubNet G_{is} 主要由两个堆叠的 SIL 块和一个全连接层组成。模型以子带单元 $\{\mathbf{b}_i\}_{i=1}^{F}$ 为输入，G_{is} 生成最终输出复数理想比率掩码 (cIRM) $\mathbf{M}^r$ 和$\mathbf{M}^i$。

2. 方法

我们考虑对采样率为 16 kHz 的单通道信号进行短时傅里叶变换（STFT）：

$$X_{t,f} = S_{t,f} + N_{t,f}. \tag{1}$$

其中，$t = 1, ..., T$和 $f = 0, ..., F-1$ 分别表示帧索引和频率索引。$X_{t,f}$、$S_{t,f}$和 $N_{t,f}$分别是含噪语音、无噪语音（麦克风拾取的混响信号）和干扰噪声在复数域中的 STFT 因子。本文仅关注频域（STFT）中的去噪任务，目标是滤除噪声，并以 16 kHz 采样率恢复无噪语音信号。

为实现这一目标，我们提出了一种名为 Inter-SubNet 的框架。该框架通过 SubInter 模块提取跨频带依赖关系和全局频谱模式等全局频谱信息，同时保留基于子带的模型专注于局部频谱模式的能力。Inter-SubNet 的整体示意图 G_{is} 如图 1 所示。首先，通过“展开”操作将输入幅度谱 $\mathbf{X} \in \mathbb{R}^{F\times T}$转换为子带单元 $\{\mathbf{b}_i\}_{i=1}^{F}$，其中 F和 T分别表示频率点和帧的总数。模型 G_{is} 以子带单元为输入，生成最终输出复数理想比率掩码 $\mathbf{M}^r$ 和 $\mathbf{M}^i$。该模型 G_{is} 主要由两个堆叠的 SubInter-LSTM（SIL）模块和一个全连接层构成。我们将在以下小节中详细介绍各个部分。

2.1. 子带单元

子带单元通过“展开”操作从幅度谱$\mathbf{X}$ 中获取，并作为 Inter-SubNet 的输入。具体而言，对于幅度谱$\mathbf{X}$中的每个频率 i，我们取一个频率点向量 $\mathbf{X}_i \in \mathbb{R}^T$ 以及其相邻的 $2\times n$ 个频率点向量，构成一个子带单元$\mathbf{b}_i$:

$$\mathbf{b}_i = [\mathbf{X}_{i-n}, \cdots, \mathbf{X}_i, \cdots, \mathbf{X}_{i+n}] \in \mathbb{R}^{F_s\times T}. \tag{2}$$

其中，“[,]”表示拼接操作，且 $F_s = 2n+1$。此外，具有 $i+n > F-1$ 或 $i-n < 0$ 的边界频率被指定为循环傅里叶频率。通过上述过程，我们最终可以获得总计 F 个这样的子带单元，具有维度为（F_s, T），记作 $\{\mathbf{b}_i\}_{i=1}^{F} \in \mathbb{R}^{F\times F_s\times T}$。

2.2. SubInter 模块

在多通道语音分离中， [19] 通过音频通道之间的交互，将全局信息与各个通道的信息相结合。不同通道。受此启发，我们引入 SubInter 模块，通过并行子带单元之间的交互，利用全局频谱信息（包括跨频带依赖关系和全局频谱模式），并将全局频谱信息融入这些子带单元中。具体而言，按照第 2.1 节，我们有一组 F子带单元 $\{\mathbf{b}_i\}_{i=1}^{F}$，其中 $\mathbf{b}_i \in \mathbb{R}^{F_s\times T}$，其中 F_s 表示一个子带单元中的频率点数。SubInter 模块应用于这些子带单元，以补充全局频谱信息：

$$\{\hat{\mathbf{b}}_i\}_{i=1}^{F} = \mathcal{H}(\{\mathbf{b}_i\}_{i=1}^{F}). \tag{3}$$

其中，$\hat{}\mathbf{b}_i \in \mathbb{R}^{F_s\times T}$ 是包含全局频谱信息和局部频谱模式的子带特征，而 $\mathcal{H}(\cdot)$ 是用于描述 SubInter 模块的映射函数。

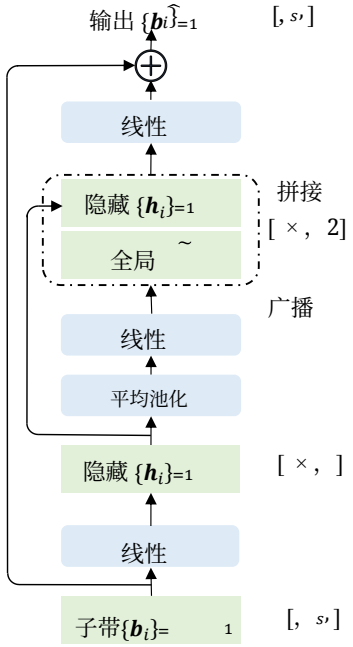


图 2。SubInter 模块的详细结构。“Hidden”表示隐藏表示。“Global”表示第二个线性层的输出，其中包含全局频谱信息。

关于具体 SubInter 模块的设计，如图 2 所示，对于子带单元集合 $\{\mathbf{b}_i\}_{i=1}^{F}$，我们对维度进行变换，

然后应用映射函数为 $\mathcal{F}(\cdot)$ 的线性层，以获得隐藏表示$\mathbf{h}_i \in \mathbb{R}^{T\times H}$:

$$\{\mathbf{h}_i\}_{i=1}^{F} = \mathcal{F}(transform(\{\mathbf{b}_i\}_{i=1}^{F})). \quad (4)$$

其中，$\{\mathbf{h}_i\}_{i=1}^{F}$ 在内部表示并行子带单元的信息。接下来，对所有$\mathbf{h}_i$s求平均，并将其输入第二个线性层:

$$\tilde{\mathbf{h}} = \mathcal{R}(\frac{1}{F}\sum_{i=1}^{F}\mathbf{h}_i). \quad (5)$$

其中，$\mathcal{R}(\cdot)$表示由第二个线性层定义的映射函数。这种操作使原本并行的子带单元能够进行交互，因此该模块提取出包含全局频谱信息的$\tilde{\mathbf{h}} \in \mathbb{R}^{T\times H}$，例如跨子带依赖关系和全局频谱模式。最后，将每个$\tilde{\mathbf{h}}_i$ 与$\mathbf{h}$ 拼接，以补充全局频谱信息。随后，将它们输入第三个线性层，通过映射函数 $\mathcal{P}(\cdot)$进行特征融合。经过维度变换，并从模块输入$\mathbf{b}_i$添加残差连接后，得到输出$\hat{}\mathbf{b}_i \in \mathbb{R}^{F_s\times T}$:

$$\{\hat{\mathbf{b}}_i\}_{i=1}^{F} = \{\mathbf{b}_i\}_{i=1}^{F} + transform(\mathcal{P}(\{[\mathbf{h}_i, \tilde{\mathbf{h}}]\}_{i=1}^{F})). \quad (6)$$

其中，“[,]”是拼接操作。

通过上述模块，我们让原本并行处理的子带单元相互交互，从而提取全局频谱信息，再将全局信息融合到先前的子带单元中。因此，输出子带特征$\{\hat{}\mathbf{b}_i\}_{i=1}^{F}$同时包含全局频谱信息和局部频谱模式。

2.3. 模型架构与工作流程

原始子带模型由两个堆叠的 LSTM 层组成，它们并行处理子带单元，以建模信号平稳性和局部频谱模式。为提取并整合全局频谱信息，我们提出了 Inter-SubNet，通过 SubInter 模块扩展原始子带模型。具体而言，我们采用 SIL 块替换原始子带模型中的 LSTM 层。如图 1 所示，Inter-SubNet G_{is} 主要由两个堆叠的 SIL 块和一个全连接层组成。每个 SIL 块中都堆叠了 SubInter 模块、LSTM 网络和组归一化（G-norm） [18] 层。

根据第 2.1 节，共有 F个子带单元 $\{\mathbf{b}_i\}_{i=1}^{F}$ 作为输入送入 G_{is}。在 SIL 块内部，$\{\mathbf{b}_i\}_{i=1}^{F}$ 首先由 SubInter 模块进行交互，以提取并融合全局频谱信息，输出子带特征 $\{\hat{}\mathbf{b}_i\}_{i=1}^{F}$。随后，LSTM 学习这些子带特征中包含的局部频谱模式和互补的全局频谱信息。之后，G-norm 层对 LSTM 的输出进行归一化。最后，全连接层输出作为学习目标的复数理想比率掩码。

3. 实验设置

3.1. 数据集

Inter-SubNet 使用 DNS Challenge - Interspeech 2021 数据集的一个子集进行训练，该数据集的采样率为 16 kHz，以下章节中将简称为 DNS Challenge 数据集。干净语音数据集包含来自 2,150 位说话人的 563 小时语音片段；噪声数据集则包含来自 150 类噪声的 181 小时、超过 60,000 个片段。为充分利用数据集，采用动态混音来模拟模型训练期间的含噪语音。具体而言，从 openSLR26 和 openSLR28 [25]，中随机选取房间脉冲响应（RIR），并将其添加到 75% 的干净语音中。随后，以随机选取的 -5 至 20 dB 信噪比，将干净语音或混响语音与噪声混合，动态生成含噪语音。DNS Challenge 还提供了一个可公开访问的测试集，其中包含两类合成片段 [26], ，分别为有混响和无混响。每类均有 150 个含噪片段，信噪比范围为 0 至 20 dB。不同模型的有效性均在该测试集上进行了验证。

3.2. 训练设置与基线模型

采用长度为 32 ms、帧移为 16 ms 的汉宁窗，将信号转换到频域。我们采用学习率为 1e-3 的 Adam 优化器。对于子带单元，$n = 15$按照[14], 设定，这意味着每个输入频率箱两侧各取 15 个相邻频率。训练过程中，输入—目标序列对被生成固定长度为 $T = 192$ 帧的序列（约 3 秒）。

为验证子带交互方法的有效性，我们对比了以下模型。为确保比较公平，所有模型均采用相同的实验设置和预测目标（复数理想比率掩码）。

子带模型：如 [14], 所述，子带模型包含 2 个 LSTM 层，每层由 384 个隐藏单元组成。子带模型的参数总数为 1.82 M。

子带模型^L： 为了验证 Inter-SubNet 的性能提升是由 SubInter 模块而非参数数量增加所致，我们设计了一个名为子带模型的大型模型^L。通过增加 LSTM 层数，我们增大了模型的参数量。子带模型^{L}由 3 个 LSTM 层组成，每层包含 384 个隐藏单元，共有 3.00 M 个参数。

Inter-SubNet： 在第一个 SIL 块中，SubInter 模块包含 102 个隐藏单元，而 LSTM 包含 384 个隐藏单元。在第二个块中，SubInter 模块包含 307 个隐藏单元，LSTM 包含 384 个隐藏单元。Inter-SubNet 的参数量为 2.29 M。

4. 结果与讨论

4.1. 与基线方法和当前最先进方法的比较

表 1 展示了不同语音增强模型在 DNS Challenge 数据集上的性能。表中，“带混响”和“不带混响”分别表示包含和不包含混响的测试集。“# Para”表示模型的参数总量，以百万为单位。“Feat.”表示模型的输入特征，其中“波形”指时域波形，“幅度”表示频域幅度谱，而“RI”表示频域实部和虚部频谱。需要注意的是，“Mag+RI”表示模型将所有幅度谱、实部频谱和虚部频谱都作为输入特征。

在表 1 的最后三行中，我们比较了子带模型、子带模型^{L}和所提出的 Inter-SubNet 的性能。结果表明，Inter-Sub Net 在所有评估指标上均优于子带模型。此外，结果还表明，子带模型^L 的性能优于子带模型。

表 1。DNS 挑战赛测试数据集上的性能指标，包括宽带感知语音质量评估 [MOS], NB-PESQ [MOS], STOI [%], 和尺度不变信号失真比 [dB]。

模型	特征	参数量 (M)	含混响				不含混响			
			宽带感知语音质量评估	NB-PESQ	STOI 尺度不变信号失真比		宽带感知语音质量评估	NB-PESQ	STOI 尺度不变信号失真比	
含噪	-	-	1.822	2.753	86.62	9.033	1.582	2.454	91.52	9.07
DCCRN-E [20]	RI	3.70	-	3.077	-	-	-	3.266	-	-
Conv-TasNet [2]	波形	5.08	2.750	-	-	-	2.730	-	-	-
PoCoNet [21]	RI	50.0	2.832	-	-	-	2.748	-	-	-
DCCRN+ [15]	RI	3.30	-	3.300	-	-	-	3.330	-	-
TRU-Net [22]	Mag*	0.38	2.740	3.350	91.29	14.87	2.860	3.360	96.32	17.55
CTS-Net [23]	幅度+实部和虚部	4.99	3.020	3.470	92.70	15.58	2.940	3.420	96.66	17.99
FullSubNet [16]	Mag	5.64	3.057	3.584	92.11	16.04	2.882	3.428	96.32	17.30
FullSubNet+ [17]	幅度+实部和虚部	8.67	3.177	3.648	93.64	16.44	**3.002**	3.503	**96.67**	18.00
子带模型 [14]	Mag	1.82	2.885	3.475	92.10	15.54	2.552	3.242	95.10	16.46
子带模型^L	Mag	3.00	2.953	3.504	92.44	16.02	2.593	3.265	95.36	16.94
Inter-SubNet	Mag	2.29	**3.207**	**3.659**	**93.98**	**16.76**	2.997	**3.504**	96.61	**18.05**

*更准确地说，TRU-Net 将逐通道能量归一化（PCEN） [24] 特征作为输入。

增加参数数量后也是如此。然而，与子带模型^{L}相比，Inter-SubNet以更少的参数取得了更优的结果，证明Inter-SubNet的性能提升归功于SubInter 模块，而非参数数量的增加。

此外，从表格中可以比较 Inter-SubNet 与 FullSubNet [16] 和 FullSubNet+ [17]。可以看出，尽管参数数量少得多，我们的模型在所有指标上都优于 FullSubNet。此外，与更复杂的 FullSubNet+相比，后者将幅度、实部和虚部频谱图全部作为输入特征，而 Inter-SubNet 仅使用幅度频谱图作为输入，其表现仍优于或持平于 FullSubNet+。这些结果表明，与将全频带模型和子带模型串联相比，我们提出的 SubInter 模块是补充全局频谱信息的一种更高效、更轻量的方式。

我们还在表格 中，将提出的 Inter-SubNet 与 DNS Challenge 数据集[20 23] [2, 15, 1] 上的其他一些最先进的时域和频域方法进行了比较。可以得出结论：与最新方法相比，Inter-SubNet 在无混响去噪任务中表现出色；而在存在混响时，其性能提升更加显著。这表明，所提出的 Inter-SubNet 延续了文献 [14], 报告的子带模型在应对混响影响方面的卓越能力，并通过 SubInter 模块对全局频谱信息进行建模，大幅提升了去噪能力。

4.2. 消融研究

本节进行了消融研究，以进一步探究 SubInter 模块的作用。表 2 展示了消融实验的结果，其中“1st SubInter”和“2nd SubInter”分别指第一个和第二个 SIL 块中的 SubInter 模块。“Δ”表示移除模块后的数值与移除模块前的数值之差。可以看出，移除第一个和第二个 SubInter 模块都会降低模型在信号（尺度不变信号失真比，SI-SDR）和感知质量（宽带感知语音质量评估，WB-PESQ）方面的表现。这表明，在每层子带处理 LSTM 网络之前进行子带交互能够带来性能提升，而且这些提升可以叠加，彼此不会产生冲突。此外，性能下降

表 2。无混响测试集上的消融研究中 WB-PESQ、 ΔWB-PESQ、尺度不变信号失真比和 Δ尺度不变信号失真比的性能。

模型	WB-PESQ	ΔWB-PESQ	SI-SDR	ΔSI-SDR
Inter-SubNet	**2.997**	-	**18.05**	-
– 2nd SubInter	2.867	-0.130	17.37	-0.68
– 1st SubInter	2.552	-0.315	16.46	-0.91

移除第一个 SubInter 模块造成的性能下降大于移除第二个 SubInter 模块。这可能是因为移除第二个 SubInter 模块后，模型仍可通过第一个 SubInter 模块获取全局频谱信息；而移除第一个 SubInter 模块后，模型则退化为缺乏全局频谱信息的子带模型，因此在复杂声学环境中表现不佳。总的来说，使用 SubInter 模块确实能够有效提升 Inter-SubNet 的降噪能力。

5. 结论

本文提出了一种新颖的子带交互（SubInter）模块，作为一种用于语音增强的全局频谱信息建模新方法，并提出了一种名为 Inter-SubNet[1] 的新框架。这种轻量级语音增强框架旨在利用全局频谱信息，同时保留聚焦局部频谱模式的能力。实验结果证明了所提出的 SubInter 模块的有效性。我们还在 DNS Challenge 数据集上将 Inter-SubNet 与其他排名靠前的方法进行了比较，结果表明，所提出的 Inter-SubNet 性能卓越。

未来，我们将探索子带交互是否适用于其他基于子带的语音增强方法。实现因果子带交互也将是我们未来工作的重点。

致谢：本研究获得国家自然科学基金（62076144）、腾讯人工智能实验室 Rhino-Bird 重点研究计划（RBFR2022005）以及清华大学-腾讯联合实验室的支持。

[1] Co代码和示例：https://github.com/RookieJunChen/Inter-SubNet

6. 参考文献

[1] Ashutosh Pandey 和 DeLiang Wang，“采用扩张卷积的密集连接神经网络：时域实时语音增强”，载于 *ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2020，第 6629–6633 页。[2] Yuichiro Koyama、Tyler Vuong、Stefan Uhlich 和 Bhiksha Raj，“探索用于基于深度神经网络的低延迟语音增强的时序卷积网络的最佳损失函数”，*arXiv* 预印本 *arXiv:2005.11611*，2020。[3] Xiang Hao、Xiangdong Su、Zhiyu Wang、Hui Zhang 等，“Unetgan：一种适用于极低信噪比条件的时域鲁棒语音增强方法”，*arXiv* 预印本 *arXiv:2010.15521*，2020。[4] Ashutosh Pandey 和 DeLiang Wang，“Tcnn：用于时域实时语音增强的时序卷积神经网络”，载于 *ICASSP 2019-2019 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2019，第 6875–6879 页。[5] Zhifeng Kong、Wei Ping、Ambrish Dantrey 和 Bryan Catanzaro，“利用自注意力在波形域进行语音去噪”，载于 *ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022，第 7867–7871 页。[6] Yong Xu、Jun Du、Li-Rong Dai 和 Chin-Hui Lee，“基于深度神经网络的语音增强回归方法”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 23 卷，第 1 期，第 7–19 页，2014。[7] DeLiang Wang 和 Jitong Chen，“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。[8] Ke Tan 和 DeLiang Wang，“利用门控卷积循环网络学习复数频谱映射，实现单声道语音增强”，*IEEE/ACM*音频、语音与语言处理汇刊，第 28 卷，第 380–390 页，2019。[9] Jitong Chen 和 DeLiang Wang，“用于监督式语音分离中说话人泛化的长短期记忆网络”，美国声学学会杂志，第 141 卷，第 6 期，第 4705–4714 页，2017。[10] DeLiang Wang，“关于作为听觉场景分析计算目标的理想二值掩码”，载于人类与机器的语音分离，第 181–197 页。施普林格，2005。[11] Lei Sun、Jun Du、Li-Rong Dai 和 Chin-Hui Lee，“基于 LSTM-RNN 的语音增强的多目标深度学习”，载于*2017*年免提语音通信与麦克风阵列（*HSCMA*）。IEEE，2017，第 136–140 页。[12] Donald S Williamson、Yuxuan Wang 和 DeLiang Wang，“单声道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。[13] Dacheng Yin、Chong Luo、Zhiwei Xiong 和 Wenjun Zeng，“PHASEN：一种感知相位与谐波的语音增强网络”，载于*AAAI*人工智能会议论文集，2020，第 34 卷，第 9458–9465 页。[14] Xiaofei Li 和 Radu Horaud，“利用延迟子带长短期记忆网络进行在线单声道语音增强”，载于*INTERSPEECH*论文集，2020。

[15] Shubo Lv、Yanxin Hu、Shimin Zhang 和 Lei Xie，“DCCRN+：用于语音增强的通道级子带 DCCRN 与信噪比估计”，*Proc. Interspeech 2021*，第 2816–2820 页，2021 年。[16] Xiang Hao、Xiangdong Su、Radu Horaud 和 Xiaofei Li，“FullSubNet：用于实时单通道语音增强的全频带与子频带融合模型”，载于*ICASSP 2021-2021 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2021 年，第 6633–6637 页。[17] Jun Chen、Zilin Wang、Deyi Tuo、Zhiyong Wu、Shiyin Kang 和 Helen Meng，“FullSubNet+：用于语音增强的复数频谱图通道注意力 FullSubNet”，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2022 年，第 7857–7861 页。[18] Yuxin Wu 和 Kaiming He，“组归一化”，载于欧洲计算机视觉会议 *(ECCV)* 论文集，2018 年，第 3–19 页。[19] Yi Luo、Zhuo Chen、Nima Mesgarani 和 Takuya Yoshioka，“端到端麦克风排列与数量不变的多通道语音分离”，载于*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2020 年，第 6394–6398 页。[20] Yanxin Hu、Yun Liu、Shubo Lv、Mengtao Xing、Shimin Zhang、Yihui Fu、Jian Wu、Bihong Zhang 和 Lei Xie，“DCCRN：用于相位感知语音增强的深度复数卷积循环网络”，*Proc. Interspeech2020*，2476，2020 年 2472pp. – 。[21] Umut Isik、Ritwik Giri、Neerad Phansalkar、Jean-Marc Valin、Karim Helwani 和 Arvindh Krishnaswamy，“Poconet：借助频率位置嵌入、半监督对话数据和偏置损失实现更优语音增强”，*Proc. Interspeech2020*，第 2487–2491 页，2020 年。[22] Hyeong-Seok Choi、Sungjin Park、Jie Hwan Lee、Hoon Heo、Dongsuk Jeon 和 Kyogu Lee，“采用微型循环 U-Net 的实时降噪与去混响”，载于*ICASSP 2021- 2021 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2021 年，第 5789–5793 页。[23] Andong Li、Wenzhe Liu、Chengshi Zheng、Cunhang Fan 和 Xiaodong Li，“两个头胜过一个：用于单耳语音增强的两阶段复数频谱映射方法”，*IEEE/ACM*音频、语音与语言处理汇刊，第 29 卷，第 1829–1843 页，2021 年。[24] Yuxuan Wang、Pascal Getreuer、Thad Hughes、Richard F Lyon 和 Rif A Saurous，“用于鲁棒远场关键词检测的可训练前端”，载于*2017 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2017 年，第 5670–5674 页。[25]Tom Ko、Vijayaditya Peddinti、Daniel Povey、Michael L Seltzer 和 Sanjeev Khudanpur，“混响语音数据增强在鲁棒语音识别中的研究”，载于*2017 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2017 年，第 5220–5224 页。[26] Chandan KA Reddy、Vishak Gopal、Ross Cutler、Ebrahim Beyrami、Roger Cheng、Harishchandra Dubey、Sergiy Matu- sevych、Robert Aichner、Ashkan Aazami、Sebastian Braun 等，“INTERSPEECH 2020 深度噪声抑制挑战赛：数据集、主观测试框架与挑战结果”，*Proc. Interspeech 2020*，第 2492–2496 页，2020 年。