arXiv:2603.07471v1 [eess.AS] 8 Mar 2026

面向真实环境下语音增强模型的轻量化自适应

Longbiao Cheng、*Shih-Chii Liu*

神经信息学研究所，苏黎世大学与苏黎世联邦理工学院

摘要

近期研究表明，部署后自适应能够提升语音增强模型在未见噪声条件下的鲁棒性。然而，现有方法往往带来过高的计算与内存开销，限制了其在设备端部署的适用性。在本工作中，我们研究了存在动态声学场景变化的真实场景下的模型自适应，并提出一种轻量化框架：在冻结骨干网络上增配低秩适配器，并通过自监督训练进行更新。在涵盖37种噪声类型、111种环境以及三个信噪比范围（包括具有挑战性的 [-8,0] 分贝范围）的序列场景评估实验中，结果表明，我们的方法仅更新基础模型不到1%的参数，同时在每个场景仅20次更新内平均实现1.51分贝的尺度不变信号失真比提升。与最先进方法相比，我们的框架在感知质量上达到了相当或更优的水平，且收敛更平滑、更稳定，证明了其在真实声学条件下对语音增强模型进行轻量化设备端自适应的实用性。

索引词— 语音增强、自监督自适应、低秩自适应、轻量级自适应

1. 引言

语音增强（**语音增强**）在嘈杂环境中，对于使用助听器及其他听觉辅助设备的听者而言，在提升语音可懂度和减轻聆听负担方面发挥着重要作用 [1, 2, 3, 4]。近期，基于神经网络的方法在广泛的声学条件下均取得了显著进展 [5, 6]。尽管取得了这些进展，深度学习模型的泛化能力有限仍然是一项挑战。在训练条件下表现良好的系统，部署于不匹配环境时往往会出现性能下降，例如遇到未见噪声类型、不匹配麦克风或不同语音特征时。

为了提高网络的泛化能力，训练过程中通常会采用大量数据增强，例如纳入多样化的噪声源和说话人群体 [7, 8]。与此同时，人们开发了更先进的模型架构，包括在大规模音频数据集上预训练的自监督学习骨干网络 [9, 10, 11] ，以及显式建模语音与噪声分布的生成模型（如基于扩散的方法） [12, 13, 14]。尽管这些方法能够提高网络的泛化能力，但通常以增加模型复杂度为代价，因此难以将这些模型部署到边缘设备上。

由于轻量级模型在处理广泛声学场景方面的能力本身有限，一种替代方法是探索持续学习 [15, 16] 以实现部署后自适应。例如，基于RemixIT的方法 [17, 18, 19, 20], 利用预训练网络生成伪目标，以训练更适合部署环境的学生网络。近期工作探索了测试时训练，即引入噪声目标辅助任务，在自适应过程中更新共享骨干网络层 [21]。然而，这些方法通常需要大幅增加参数以支持教师—学生框架或辅助任务，并涉及对预训练模型参数中很大一部分进行微调。除了带来超出低资源平台可用范围的内存与计算需求外，网络训练过程中的收敛也需要更长时间。

此外，大多数既有自适应研究旨在提升语音增强在分布外（**OOD**数据集上的表现，这类数据集通常涵盖多种说话人、噪声类型以及广泛的信噪比（**信噪比**）范围。此类数据集虽有助于进行基准测试，但与真实世界环境有显著差异。在实际应用中，声学条件按场景组织，系统在任一时刻都只在单一场景中运行。然而，随着时间推移，场景会不断变化，从而导致连续场景变化。与针对多样但静态的数据集进行一次性自适应相比，针对这类场景变化处理自适应过程更具挑战性，但对于实现可靠、实用的使用体验至关重要。

在这项工作中，我们致力于解决真实声学条件下轻量级语音增强模型自适应的问题。我们的贡献如下：

- 我们针对真实世界声学场景中的语音增强模型，形式化定义了一种自适应设定，其中声学场景会随时间变化。
- 我们提出了一种基于低秩适配器的轻量级自监督自适应框架 [22], ，无需对整个模型进行微调。
- 我们在 111 种语音增强骨干网络上评估了所提出的框架，结果显示，对于一个场景，仅需 20 个自适应步骤，尺度不变信号失真比（**尺度不变信号失真比**）平均提升 1.51 分贝，同时更新的预训练参数不到 1%。

这些结果表明，在动态的真实世界语音增强声学环境中，轻量级自适应具有实用性。

2. 问题表述

我们考虑单通道语音增强任务，其中观测信号$y(t)$ 建模为

$$y(t) = s(t) + n(t), \tag{1}$$

这项工作部分由瑞士国家科学基金会 CA-DNNEdge 项目（208227）资助。

© 2026 IEEE。允许个人使用本材料。对于所有其他用途，无论是在任何当前或未来的媒介中，包括为广告或宣传目的重印/再版本材料、创作新的汇编作品、用于转售或再分发至服务器或列表，或在其他作品中重用本作品的任何受版权保护的组成部分，均须获得 IEEE 的许可。

其中 $s(t)$ 表示干净语音信号，而 $n(t)$ 表示加性干扰。语音增强的目标是估计$\hat{s}(t)$，使其尽可能接近 $s(t)$，并从带噪输入 $y(t)$。

令 f_θ 表示在成对带噪与干净语音源 语料库上训练的深度神经网络，其中 θ_0 表示预训练基础模型的参数。在部署时，系统将在不同的声学场景中运行，每个场景对应于若干因素相对稳定的组合，例如背景噪声类型、说话人和 信噪比 范围。我们将给定场景的自适应数据集表示为

$$\mathcal{D}_{\text{adapt}}^{(m)} = \{y_i^{(m)}\}_{i=1}^{N_m}, \tag{2}$$

其中，m 为场景索引，且没有可用的干净参考。目标是更新模型参数 θ_m，使适配后的模型 f_{θ_m} 在场景 m中实现更好的增强性能，并在相同条件下收集的不相交测试集 $\mathcal{D}^{(m)}$ 测试集上进行评估。

然而，在实际使用中，声学环境并非固定不变。随着时间推移，系统会遇到场景变化，声学条件会从 m 转变为新的场景 $m+1$（例如，出现新的噪声源、信噪比发生变化信噪比，或说话人发生变化）。因此，模型必须持续自适应，仅使用 θ_m 的数据更新 θ_{m+1}，以利用 $\mathcal{D}^{(m+1)}$ 进行自适应。这种在场景变化下的持续自适应带来了额外挑战，有别于针对 分布外（OOD）数据集的一次性自适应；在这种情况下，所有场景的数据都可在自适应期间同时获得，模型可以在这些数据上联合优化。

3. 方法

由于第2节所设定的自适应场景中无法获得带噪-干净音频对，传统的监督微调无法应用。因此，我们提出一种自监督框架，利用预训练基础模型 f_{θ_0} 生成伪目标，并采用轻量级低秩适配器，以便在场景演变时实现高效且稳定的更新。

3.1. 自监督适应框架

遵循先前工作 [17, 18, 19, 20], ，自适应过程首先使用预训练基础模型 f_{θ_0}，从带噪输入 $y \in \mathcal{D}^{(m)}$ 自适应 （场景 m）生成伪干净估计：

$$\hat{x} = f_{\theta_0}(y). \tag{3}$$

该估计$\hat{x}$在自适应过程中用作教师目标。为构建自监督学习信号，我们采样一段独立的噪声片段 $n \in \mathcal{D}^{(m)}$ 自适应，其来自同一场景，并按因子 α 进行缩放，该因子对应于随机采样的信噪比，再将其与$\hat{x}$重新混合，从而得到自适应输入：

$$\tilde{y} = \hat{x} + \alpha n. \tag{4}$$

适配后的模型 f_{θ_m} 随后产生增强输出：

$$\tilde{x} = f_{\theta_m}(\tilde{y}). \tag{5}$$

通过最小化适应后的输出与伪目标之间的任务特定损失$\mathcal{L}(\tilde{x},\ \hat{x})$，更新参数 θ_m：

$$\theta_m = \arg\min_\theta \mathbb{E}_{y\sim\mathcal{D}_{\text{adapt}}^{(m)}}\Big[\mathcal{L}\big(f_\theta(\tilde{y}), \hat{x}\big)\Big]. \tag{6}$$

算法 1自监督低秩自适应

要求： f_{θ_0}：冻结基础模型； $\phi_m = \{A_m, B_m\}$：适配器场景 m的参数

对于当前场景中的每个含噪语音 y **执行**
  生成伪目标：$\hat{x} = f_{\theta_0}(y)$
  根据目标采样一个噪声片段 n 和缩放系数 α SNR
  创建自适应输入：$\tilde{y} = \hat{x} + \alpha n$
  前向传播适配后的模型：$\tilde{x} = f_{\theta_0,\ \phi_m}(\tilde{y})$
  通过最小化 $\mathcal{L}(\tilde{x},\ \hat{x})$来更新 ϕ_m
结束循环
在场景 m中进行评估时，通过以下方式将 ϕ_m 合并到基础模型中 公式 7

当场景发生变化时，使用新的数据集 $\mathcal{D}^{(m+1)}$ 进行适应，并重复相同的过程，模型参数将从 θ_m 适应到 θ_{m+1}。通过这种方式，系统能够在不同场景间逐步适应，从而跟踪一系列变化条件。

3.2. 低秩适配器

一种直接的自适应策略是对所有参数进行微调θ。虽然这可能会在单一场景中带来改进，但容易出现过拟合，并且往往会导致基础模型 f_{θ_0}所编码的知识发生灾难性遗忘。基础模型通常是在一个全面的源数据集上训练的，该数据集涵盖了不同的说话人、噪声类型和信噪比条件。

为克服这些挑战，我们采用低秩适应(**低秩适应**) [22], ，将自适应限制在低维子空间内，同时保持预训练骨干网络冻结。对于预训练权重矩阵 $W_0 \in \mathbb{R}^{d\times k}$ 在 θ_0中的场景专属参数化定义如下

$$W_m = W_0 + \beta B_m A_m, \qquad B_m \in \mathbb{R}^{d\times r}, \quad A_m \in \mathbb{R}^{r\times k}, \tag{7}$$

其中，m 表示当前场景，$r \ll \min(d,\ k)$为适配秩，β 为缩放因子。在场景 m内进行自适应时，仅更新 A_m 和 B_m，而基础权重 W_0 保持不变。在推理时，通过将残差更新 $B_m A_m$ 与基础权重合并，得到适配后的模型。自适应过程总结于算法1。

这种自适应方法天然适用于场景变化过程中的连续自适应。特定场景下语音增强模型的性能 m 通过其自身的一对轻量级适配器得到提升$(A_m,\ B_m)$同时保留 W_0中编码的通用知识。当声学条件从场景 m 转变为场景$m+1$时，系统只需从$(A_m,\ B_m)$ 切换到一对新的$(A_{m+1},\ B_{m+1})$，而无需修改骨干网络。这样，自适应就局限于场景特定参数，从而确保效率并减轻灾难性遗忘。

4. 实验设置

4.1. 网络

我们使用两种具有代表性的语音增强架构评估所提出的方法：一种是基于门控循环单元(**GRU**)的网络，另一种是基于双路径循环神经网络的网络 [23]。

该门控循环单元网络由一个输入全连接(**FC**) 层组成，该层包含 128 个单元，后接两个门控循环单元层，每层各有 128 个隐藏状态，以及一个包含 128 个单元的输出 FC 层。输入

表1：比较所提出的适应框架与 RemixIT 框架在 GRU 和双路径循环神经网络模型上的表现，分别在独立（✗）和序列（✓）场景设置下进行。**粗体**数值表示每种模型配置中的最佳表现，而蓝色数值表示相较于预训练基线的性能下降。

模型	序列场景?	自适应框架	可适应参数 #	可适应参数 %	信噪比 ∈ [-8, 0] 分贝			信噪比 ∈ [0, 5] 分贝			信噪比 ∈ [5, 10] 分贝		
					感知语音质量评估	短时客观可懂度	尺度不变信号失真比	感知语音质量评估	短时客观可懂度	尺度不变信号失真比	感知语音质量评估	短时客观可懂度	尺度不变信号失真比
带噪语音					1.08	67.88	-3.89	1.16	81.48	2.53	1.35	89.25	7.49
GRU*		预训练			1.16	71.01	3.86	1.35	84.85	6.82	1.57	90.99	9.58
	✗	RemixIT	230,144	100	1.19	71.64	4.64	1.39	84.96	8.27	1.62	91.07	11.50
	✗	Ours	**512**	**0.22**	1.22	71.96	4.71	1.44	85.34	8.36	1.67	91.38	11.43
	✓	RemixIT	230,144	100	1.18	70.34	4.63	1.34	83.13	8.42	1.51	88.63	11.03
	✓	Ours	**512**	**0.22**	**1.23**	**72.65**	**4.84**	**1.47**	**85.84**	**8.65**	**1.72**	**91.64**	**11.89**
双路径循环神经网络*		预训练			1.21	74.93	5.15	1.46	87.82	8.70	1.73	93.05	11.91
	✗	RemixIT	89,258	100	1.24	73.74	5.58	1.49	87.48	9.88	1.79	92.91	13.45
	✗	Ours	**708**	**0.79**	1.26	**75.52**	5.52	1.54	**87.93**	9.46	1.84	93.19	12.94
	✓	RemixIT	89,258	100	**1.27**	74.51	5.82	1.44	87.68	**10.11**	1.66	92.78	13.60
	✓	Ours	**708**	**0.79**	**1.27**	75.18	**5.85**	**1.54**	**87.93**	**10.11**	**1.84**	**93.21**	**13.76**

*：门控循环单元模型：23 万个参数，17M MAC/s；双路径循环神经网络模型：89 千个参数，1503M MAC/s。

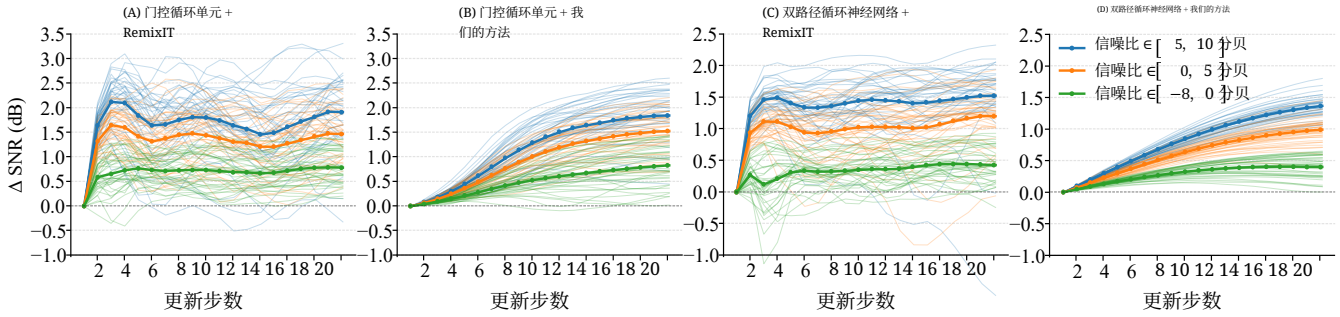

图1：在三个信噪比范围内，自适应门控循环单元和双路径循环神经网络骨干网络的每次更新信噪比提升（Δ信噪比，单位为分贝）。浅色曲线表示各个声学场景。使用 RemixIT 进行自适应可快速取得初期提升，但轨迹不稳定；而我们的方法则能随着步骤推进，带来稳定且持续的改进。

特征被压缩为等效矩形带宽(**ERB**域幅度频谱图，使用 128 个 ERB滤波器和 0.3 的压缩比提取。分析使用 512 的帧长和 256 的帧移。网络预测一个 128 维的ERB 掩码，其值通过 [0,1] Sigmoid激活函数限制在该范围内。最终网络包含 230.14 k 个参数，每秒的计算成本为 16.80 M乘加运算(**MAC**)。在适应过程中，低秩适应应用于输入和输出FC 层，秩为 1，缩放因子为 64。

双路径循环神经网络直接在时频域中运行，其输入是短时傅里叶变换（**STFT**）频谱的实部和虚部拼接结果，帧长为 320，跳跃长度为 160。模型估计一个复数掩码，其值被限制为 [1, , 并使用双曲正切激活函数。该架构首先通过输入卷积层将二维输入（实部和虚部）映射为 32 个通道。网络包含四个双路径模块。每个模块包括：1）帧间模块，由一个包含 32 个单元的单向门控循环单元和一个包含 32 个单元的全连接层构成；2）帧内模块，由一个每个方向包含 32 个单元的双向门控循环单元以及一个包含 32 个单元的全连接层构成。输出由一个卷积层生成，该卷积层将 32 通道特征映射回 2 个通道（实部和虚部）。双路径循环神经网络模型的参数量为 89.25k，计算成本为每秒 1503.32 M MAC 。由于双路径循环神经网络中的输入和输出卷积层均使用大小为 1 的卷积核，因此采用全连接层实现。在自适应过程中，低秩适应应用于所有全连接层，秩为 1，缩放因子为 8。

4.2. 数据集

基于门控循环单元和双路径循环神经网络的模型使用深度噪声抑制（DNS）挑战数据集进行训练 [24]。该数据集包含来自6,000多名说话人的760.5小时干净语音，以及涵盖150多个声学类别、超过65,000条片段的噪声集。所有语音和噪声信号均重采样至16 kHz。带噪混合即时生成，以提高数据多样性，其中信噪比从 [-5, 20] 分贝范围内随机采样。

WSJ0语音语料库 [25]和 WHAM!噪声数据集 [26]用于模拟声学场景，以进行模型自适应和性能评估。噪声样本取自 WHAM!评估子集，包含在咖啡店、餐馆、酒吧、办公楼、公园等非平稳环境中录制的音频。为确保同一声学场景在自适应和评估阶段使用一致的噪声场景，我们选取在同一天、同一地点录制的噪声片段

表2：不同秩和缩放因子下适配后的门控循环单元模型的性能。结果为所有测试场景的平均值。

（秩，缩放系数）	可适应参数	PESQ	STOI	尺度不变信号失真比
含噪	-	1.20	79.54	2.04
预训练	-	1.36	82.28	6.75
(16, 1)	8,192	1.42	82.59	7.85
(32, 1)	16,384	**1.43**	82.88	7.91
(64, 1)	32,768	**1.43**	**82.95**	**8.03**
(1, 32)	512	1.42	82.81	8.04
(1, 64)	512	**1.44**	**82.89**	**8.17**
(1, 128)	512	1.41	82.65	8.14

用于模拟每组自适应和评估配对中的带噪语音，但两个阶段使用不同的片段以避免重叠。按照这一流程，我们构建了37种不同的噪声场景。对于每种噪声场景，分别生成三种不同信噪比范围的混合语音： [-8, 0] 分贝、 [0, 5] 分贝和 [5, 10] 分贝，分别代表噪声极大、中等和相对安静的条件。由此共得到 $37 \times 3 = 111$ 个模拟测试声学场景。在每个场景中，从WSJ0说话人相关数据集中随机选取两到五名说话人。使用训练子集中的语音样本进行自适应，而评估子集中的语音样本则专门用于测试。

在测试集中，每个场景包含 20 个样本，因此混合后，$111 \times 20 = 2220$ 可获得干净-噪声配对，用于评估模型性能。对于自适应阶段，输入混合信号按照第 3 节提出的流程生成，信噪比 范围为 [-5, 5]分贝。此外，混合前，语音和噪声信号都会随机裁剪为 2 秒片段。与训练过程类似，自适应和评估中使用的所有音频文件均以 16 kHz 采样。

4.3. 训练与适配详情

两个基础模型均采用监督方式训练，通过最小化估计频谱图与目标频谱图之间的均方误差(**MSE**) 进行训练。采用 Adam优化器，初始学习率为 1e-3。如果平均训练损失连续两个训练轮次未下降，则将学习率降低 10 倍。此阶段的批量大小设为 8，两个模型均训练 100 个训练轮次。自适应过程中，我们使用倒数信噪比作为损失函数。使用 Adam 优化器优化可训练参数，基于GRU的模型和基于DPRNN的模型的固定学习率分别为 1e-3 和 5e-4。批量大小设为 24，每个适配会话最多进行 20 次更新，对应最多使用 $24 \times 20 = 240$ 秒的音频信号，即 480 秒。

5. 结果

5.1. 与最先进的适配框架比较

我们将所提出的轻量级自适应框架与最先进的 RemixIT 方法进行了比较 [17], , 该方法需要更新基础模型的所有参数。我们使用感知语音质量评估 [27], 、短时客观可懂度 [28], 和尺度不变信号失真比 [29] , 在三个信噪比范围内进行了评估。详细结果总结于表1中。

我们首先评估了孤立场景下的自适应，其中每个测试场景都独立处理，且所有模型均以 DNS 预训练骨干网络为初始化。如表1所示，两种 RemixIT 和我们的方法在所有信噪比范围内的表现均优于预训练模型。值得注意的是，RemixIT 需要更新全部参数（GRU 为 230k，双路径循环神经网络为 89k），而我们的方法只需更新不到 1% 的参数（GRU 为 512，双路径循环神经网络为 708），便取得了相当或更大的性能提升。例如，在信噪比为 [0, 5] 分贝时，采用 GRU 骨干网络的我们的方法达到 1.44 的感知语音质量评估、85.34 的短时客观可懂度和 8.36 分贝的尺度不变信号失真比，优于 RemixIT（1.39 的感知语音质量评估、84.96 的短时客观可懂度和 8.27 分贝的尺度不变信号失真比）。这些结果表明，参数高效适应能够达到与全量微调相当甚至更好的性能。此外，要取得这样的结果，RemixIT 需要存储一份独立的预训练模型副本，以生成用于自监督适应的伪目标；而我们的方法只需存储轻量级适配器参数。

图1进一步展示了 Δ信噪比（单位为分贝）下，在三个信噪比范围内，采用 RemixIT 和我们提出的框架对门控循环单元和双路径循环神经网络骨干网络进行自适应后模型相较于预训练基线的表现。采用 RemixIT 时，平均 Δ信噪比在初始阶段迅速上升，但随后出现波动，表明自适应过程不稳定。相比之下，我们的框架带来了单调上升的曲线，体现出稳定且持续的性能提升。

随后，我们随机打乱全部111个测试环境，构建连续场景，并要求模型在这些环境中持续自适应。在此设置下，每个场景的初始化参数均取自前一个场景的适配后模型，以模拟长期部署。在这一设置中，我们框架的优势更加明显。虽然 RemixIT 因累积参数漂移而性能下降，但我们的方法在所有指标上均保持或超过预训练基线。使用门控循环单元时，我们的框架在 [5, 10] 分贝下取得了 1.72 的感知语音质量评估分数、91.64 的短时客观可懂度分数和 11.89 分贝的尺度不变信号失真比，明显优于 RemixIT（1.51 的感知语音质量评估分数、88.63 的短时客观可懂度分数和 11.03 分贝的尺度不变信号失真比）。双路径循环神经网络也呈现出类似趋势：我们的模型仅自适应了 0.79% 的参数，表现仍始终优于 RemixIT。

5.2 秩和缩放因子的影响

表2报告了使用 门控循环单元 网络作为骨干网络时，不同秩和缩放因子配置下自适应模型的结果。结果为所有测试环境下的平均值。与噪声基线和预训练基线相比，所有基于低秩适应的自适应方法在感知语音质量评估、短时客观可懂度和 尺度不变信号失真比方面均取得了稳定提升。在缩放因子固定为1的情况下改变秩，性能随着秩的增大而稳步提升，并在秩为64时达到最佳结果。然而，这也大幅增加了可训练参数的数量。相比之下，在秩固定为1的情况下改变缩放因子，结果表明较大的缩放因子能够显著提高参数效率：配置（1, 64）仅需512个可训练参数，便取得了总体最佳结果。

6. 结论

我们对真实声学环境下部署后的语音增强模型自适应问题进行了形式化描述。这类环境具有动态场景变化的特点，因此需要轻量级解决方案。为应对这一挑战，我们提出了一种自监督低秩自适应框架，只需更新少于语音增强模型1%的参数。在111种噪声环境下进行评估后，该框架始终能提升感知语音质量评估、短时客观可懂度和尺度不变信号失真比，同时在每个自适应步骤中保持稳定的改进。这些结果彰显了我们的方法在真实声学条件下实现语音增强模型有效且高效自适应的实用性。

7. 参考文献

[1] Dorothea Wendt、Renskje K Hietkamp 和 Thomas Lunner，“噪声及噪声抑制对处理努力的影响：一项瞳孔测量研究”，*Earand Hearing*，第38卷，第6期，第690–700页，2017年。[2] DeLiang Wang，“深度学习重塑助听器”，*IEEESpectrum*，第54卷，第3期，第32–37页，2017年。[3] Longbiao Cheng、Ashutosh Pandey、Buye Xu、Tobi Delbruck、Vamsi Krishna Ithapu 和 Shih-Chii Liu，“利用SlowFast框架调制状态空间模型，实现计算高效、超低延迟的语音增强”，发表于*ICASSP 2025*。[4] Longbiao Cheng、Ashutosh Pandey、Buye Xu、Tobi Delbruck 和 Shih-Chii Liu，“用于计算高效语音增强的动态门控循环神经网络”，发表于*2024*年国际语音通信协会会议论文集。[5] 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年。[6] Douglas O’S haughnessy，“语音增强——现代方法综述”，*IEEE*人机系统汇刊，第54卷，第1期，第110–120页，2024年。[7] Sebastian Braun 和 Ivan Tashev，“深度噪声抑制中的数据增强与损失归一化”，发表于*2020*年国际语音与计算机会议。[8] Ke Chen、Jiaqi Su、Taylor Berg-Kirkpatrick、S hlomo Dubnov 和 Zeyu Jin，“提升真实场景中语音分离的泛化能力：仿真、优化与评估策略”，发表于*Interspeech 2024*论文集。[9] Zili Huang、Shinji Watanabe、Shu-wen Yang、Paola Garc´ıa 和 Sanjeev Khudanpur，“探究自监督学习在语音增强与分离中的应用”，发表于*ICASSP 2022*。[10] Kuo-Hsuan Hung、Szu-wei Fu、Huan-Hsin Tseng、Hsin-Tien Chiang、Yu Tsao 和 Chii-Wann Lin，“增强用于语音增强的自监督嵌入”，发表于*Interspeech 2022*论文集。[11] Ching-Hua Lee、Chouchang Yang、Rakshith Sharma Srinivasa、Yashas Malur Saidutta、Jaejin Cho、Yilin Shen 和 Hongxia Jin，“利用自监督语音表征实现语音增强中的领域自适应”，发表于*ICASSP 2024*。[12] Philippe Gonzalez、Zheng-Hua Tan、Jan Østergaard、Jesper Jensen、Tommy Sonne Alstrøm 和 Tobias May，“探究用于语音增强的扩散模型设计空间”，*IEEE/ACM*音频、语音与语言处理汇刊，2024年。[13] Julius Richter、Simon Welker、Jean-Marie Lemercier、Bunlong Lay、Tal Peer 和 Timo Gerkmann，“用于泛化语音增强的因果扩散模型”，*IEEE*开放信号处理期刊，第5卷，第780–789页，2024年。[14] Julius Richter 和 Timo Gerkmann，“基于扩散的语音增强：性能与泛化能力展示”，发表于*Audio Imagination*：*NeurIPS 2024*研讨会。[15] Wenpeng Hu、Zhou Lin、Bing Liu、Chongyang Tao、Zheng-wei Tao Tao、Dongyan Zhao、Jinwen Ma 和 Rui Yan，“通过模型自适应克服持续学习中的灾难性遗忘”，发表于*ICLR*，2019年。[16] Joachim Ott与Shih-Chii Liu，“受生物启发的人体运动序列持续学习”，见*ICASSP 2023*。[17] Efthymios Tzinis、Yossi Adi、Vamsi K Ithapu、Buye Xu、Paris Smaragdis与Anurag Kumar，“RemixIT：通过自举重混实现语音增强模型的持续自训练”，《*IEEE*信号处理精选主题期刊》，第16卷，第6期，第1329–1341页，2022年。[18] Huajian Fang与Timo Gerkmann，“基于不确定性的重混用于深度语音增强中的无监督域适应”，见*IWAENC2024*。[19] Cong Han、Kevin Wilson、Scott Wisdom与John R Hershey，“无监督多通道分离与自适应”，见*ICASSP2024*。[20] Yu Liao、Haixin Guan、Shuang Wei与Yanhua Long，“利用域外噪声实现语音增强中的无监督域适应”，见*ICASSP 2025*。[21] Avishkar Behera、Riya Ann Easow、Venkatesh Parvathala与K Murty，“面向语音增强的测试时训练”，见*Interspeech 2025*会议录。[22] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen等，“低秩适应：大语言模型的低秩适应”，*ICLR2022*。[23]Yi Luo、Zhuo Chen与Takuya Yoshioka，“双路径循环神经网络：面向时域单通道语音分离的高效长序列建模”，见*ICASSP2020*。[24] Chandan KA Reddy、Vishak Gopal、Ross Cutler、Ebrahim Beyrami、Roger Cheng、Harishchandra Dubey、Sergiy Matusevych、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赛结果”，见*Interspeech 2020*会议录。[25] Douglas B. Paul与Janet M. Baker，“基于《华尔街日报》的连续语音识别语料库设计”，见语音与自然语言：*1992*年*2*月*23*日至*26*日在纽约哈里曼举行的研讨会论文集。[26]Gordon Wichern、Joe Antognini、Michael Flynn、Licheng Richard Zhu、Emmett McQuinn、Dwight Crow、Ethan Manilow与Jonathan Le Roux，“WHAM!：将语音分离扩展至噪声环境”，见*Interspeech 2019*会议录。[27] Antony W Rix、John G Beerends、Michael P Hollier与Andries P Hekstra，“感知语音质量评估（感知语音质量评估）——一种用于电话网络与编解码器语音质量评估的新方法”，见*ICASSP2001*。[28] Cees H Taal、Richard C Hendriks、Richard Heusdens与Jesper Jensen，“时频加权带噪语音的可懂度预测算法”，《*IEEE*音频、语音与语言处理汇刊》，第19卷，第7期，第2125–2136页，2011年。[29] Jonathan Le Roux、Scott Wisdom、Hakan Erdogan与John R Hershey，“信号失真比——半成品还是已完成？”，见*I CASSP 2019*。