arXiv:2603.05302v1 [cs.SD] 5 Mar 2026

SLICE：通过逐层注入条件嵌入实现语音增强

Seokhoon Moon [1], *Kyudan Jung* [2], *JaegulChoo* [2]**

[1] 韩国科学技术院 [2] KAIST AI

moonx010@kaist.ac.kr, kyudan@kaist.ac.kr, jchoo@kaist.ac.kr

摘要

真实场景录音通常会同时受到多种退化的影响，包括加性噪声、混响和非线性失真。基于扩散的增强方法在处理单一退化时表现良好，但难以应对复合退化。以往的噪声感知方法仅在输入层注入条件，这可能导致其性能不如无条件模型。为此，我们提出将退化条件注入时间步嵌入中。该条件由一个预训练编码器生成，该编码器配有用于识别噪声类型、混响和失真的多任务头；注入后，条件可传播至所有残差块，且无需进行架构改动。在仅改变注入方法的受控实验中，面对复合退化时，输入级条件的表现甚至不如完全不使用编码器，而逐层注入取得了最佳结果。该方法也能泛化至多种真实场景录音。

索引词：语音增强、扩散模型、多重退化、逐层条件、时间步嵌入

1. 引言

随着语音通信系统日益普及，在真实环境下实现稳健的语音增强仍是一项关键挑战 [1, 2, 3]。实际上，语音信号很少只受到单一来源的退化影响。在混响房间内进行的电话通话、用低质量设备录制的语音，或在嘈杂环境中进行的电话会议，都存在加性噪声、混响和非线性失真的组合 [4, 5]。这三类退化可视为主要信号损坏机制的代表：来自环境声源的加性干扰、由室内声学引起的卷积效应 [6], , 以及由录音设备或有损传输造成的非线性伪影；近期的通用增强挑战也证实了这些退化十分普遍 [7, 8, 9]。尽管这种划分并未涵盖所有情况，但可以说，它涵盖了实际语音通信场景中最常见的退化。因此，对于实际部署而言，使用单一模型处理此类复合退化至关重要。

基于扩散的生成模型已成为一种前景广阔的语音增强方法，凭借迭代去噪过程实现了强大的泛化能力 [1, 2, 10, 11]。SGMSE+ [1]在复数短时傅里叶变换域中，通过基于得分的随机微分方程对干净语音分布进行建模，在加性噪声去除方面取得了最先进的成果。为进一步提升泛化能力，噪声感知方法将外部噪声信息作为条件信号注入，以引导增强过程 [12, 13]。NASE [12] 采用预训练的 BEATs [14] 音频编码器，并结合噪声分类损失，将得到的嵌入添加到输入频谱图中。NADiffuSE [13] 同样提取噪声表征，并将其用作扩散步骤的全局条件。

然而，这些方法只处理加性噪声，却未处理混响和非线性失真，尽管它们在真实场景录音中十分常见 [3, 4, 5]。此外，这些方法依赖输入级条件注入，即在网络中的单个位置加入嵌入。由于 NCSN++ [15] 等分数网络包含数十个残差块，单次输入层扰动会在网络中逐步减弱，导致较深层无法获得条件信息。

为解决这些局限，我们将退化条件注入 NCSN++骨干网络的时间步嵌入中。由于每个残差块本就会使用时间步嵌入，这种做法无需任何架构改动，便可自然地将条件信号传播至整个网络。我们采用一个预训练的 WavLM [16] 编码器和三个专用头，通过多任务辅助损失 [17], 共同表征噪声类型、混响程度和失真强度，并生成单个条件向量，以整合噪声、混响和失真信息，从而调制分数网络的所有层。多任务设计使编码器能够分离共现退化，即使多种退化同时存在，分数网络也能获得有效的条件信息。通过仅改变注入方法的受控实验，我们发现，在复合退化情况下，输入级条件的表现甚至不如完全不使用编码器；相比之下，我们的逐层注入在所有指标上均取得最佳结果，并能泛化至各种真实场景录音。我们的贡献有两方面：我们揭示了浅层条件注入可能损害复合退化场景下的性能，并提出通过时间步嵌入注入实现逐层条件，作为多重退化语音增强的一种简单而有效的替代方案。

2. 所提方法

如图 1 所示，我们的方法在 SGMSE+框架的基础上扩展了两个组件：(a) 一个能够表征输入信号中存在的损坏类型的退化感知编码器；(b) 一种逐层条件机制，通过时间步嵌入将所得表征注入分数网络的每个残差块中。我们首先回顾 SGMSE+基线，然后详细介绍各个组件。

2.1. 预备知识：基于分数的语音增强

SGMSE+ [1] 使用基于分数的随机微分方程框架，对含噪观测条件下的干净语音分布进行建模。模型的前向过程逐渐破坏干净的

**表示通讯作者。

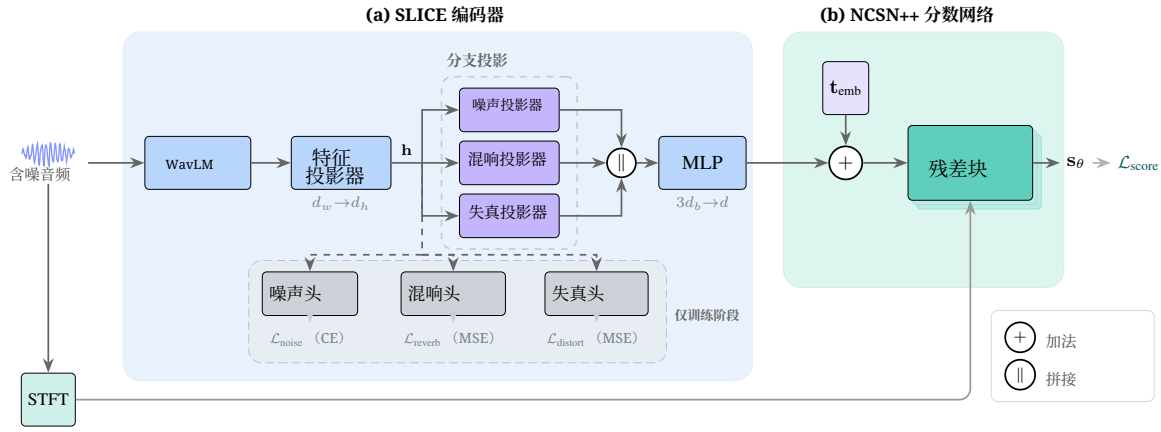

图 1：*SLICE* 概述。***(a)*** 编码器产生 $\mathbf{h} \in \mathbb{R}^{d_h}$；分支投影经拼接并映射到 $\mathbf{c}_{extra} \in \mathbb{R}^{d}$。***(b)*** $\mathbf{c}_{extra}$ 被加到$\mathbf{t}_{emb}$ 上，并在每个残差块中传播。橙色虚线：辅助头（仅训练阶段）。 d_w=768, d_h=256, d_b=128 d=512 ，。

信号 $\mathbf{x}$，使其趋近于含噪观测 $\mathbf{y}$：

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t \tag{1}$$

其中 γ 表示均值回复率， $g(t)$ 为扩散系数。分数网络 $\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}\ t) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ 通过去噪得分匹配进行训练，干净语音则借助预测-校正采样器求解反向随机微分方程来恢复。NCSN++ [15] 骨干网络计算时间步嵌入$\mathbf{e}_t$ = 多层感知机$_{\text{time}}(t) \in \mathbb{R}^d$，其中$d = 512$，并将其加到每个残差块的特征上，从而在网络的所有深度提供时间步感知。

2.2. 多重退化编码器

现在，我们介绍框架的第一个组件——退化感知编码器，如图 1(a) 所示。该模块提取紧凑的表征，以捕捉输入信号中每种退化类型的特征。

我们采用在大规模数据上预训练的 WavLM-Base [16], 语音编码器，为语音处理任务学习通用表征。给定退化波形 $\mathbf{y}_{\text{wav}}$，WavLM 会生成帧级特征 $\mathbf{F} = \text{WavLM}(\mathbf{y}_{\text{wav}}) \in \mathbb{R}^{T' \times d_w}$，其中 $d_w = 768$。在整个训练过程中，编码器参数保持冻结，以保留预训练表征。随后，卷积后处理网络通过沿时间维度进行均值池化，将这些语义丰富的特征压缩为共享表示$\mathbf{h} \in \mathbb{R}^{d_h}$，其中 $d_h = 256$，从而保留具有判别性的退化信息，并生成适用于下游条件输入的定长描述符。

以往具备噪声感知能力的方法发现，单模型在不同退化类型之间存在性能权衡 [12, 13], ，因为将所有退化类型混为一体的单一表征，可能无法分离每种失真的独特特征。受此观察启发，我们设计了三个专用头，分别刻画每种退化类型，使模型能够针对每种失真来源，分别推断相应的增强行为。遵循 DEMAND 的噪声分类体系 [18],，噪声头对 10 种噪声类别加上一个“无”类别进行 11 类分类。混响头则回归房间混响时间 T_{60}，失真头则估计非线性失真强度。这些头具有双重作用。首先，它们提供辅助多任务监督 [17]

这种监督促使共享表示为每种退化类型学习具有判别性的特征，从而缓解单个编码器必须联合处理所有退化类型时产生的跨退化权衡。其次，各头的预测结果能够在推理时对退化情况进行可解释的分析，使系统能够根据每个输入的具体失真特征调整其降噪行为。

2.3. 通过时间步嵌入注入实现逐层条件

我们的框架的第二个组件如图 1(b) 所示，它将退化表征注入分数网络。为此，我们采用 NCSN++ 骨干网络现有的时间步嵌入机制。共享表征$\mathbf{h}$ 被投影为三个 d_b 维的分支特定嵌入，其中 $d_b = 128$，然后拼接并映射至与时间步嵌入相同的维度：

$$\mathbf{c}_k = \mathbf{W}_k \mathbf{h}, \quad k \in \{\text{noise, reverb, distort}\} \tag{2}$$

$$\mathbf{c}_{\text{extra}} = \text{MLP}_{\text{cond}}([\mathbf{c}_{\text{noise}}; \mathbf{c}_{\text{reverb}}; \mathbf{c}_{\text{distort}}]) \tag{3}$$

其中 $[\cdot;\ \cdot]$ 表示拼接，$\text{MLP}_{\text{cond}}: \mathbb{R}^{3d_b} \to \mathbb{R}^d$。随后通过简单加法注入条件：

$$\tilde{\mathbf{e}}_t = \mathbf{e}_t + \mathbf{c}_{\text{extra}} \tag{4}$$

与 NASE [12],等仅在输入层进行注入的方法不同，这种方法会将信号传播至所有 ∼37 残差块，使分数网络的每一层都能感知退化。关键在于，这无需对骨干网络进行架构更改，因为只需在现有时间步嵌入上进行一次加法。

2.4. 训练目标

总损失由分数匹配目标和辅助多任务损失组成：

$$\mathcal{L} = \mathcal{L}_{\text{score}} + \lambda\,(\mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{reverb}} + \mathcal{L}_{\text{distort}}) \tag{5}$$

其中， $\mathcal{L}_{\text{noise}}$ 表示噪声分类的交叉熵损失， $\mathcal{L}_{\text{reverb}}$ 和 $\mathcal{L}_{\text{distort}}$ 表示回归的均方误差（MSE）。我们在所有实验中均设置 $\lambda = 0.3$。

表1： 仅噪声（*824* 个文件）和多重退化（*2472* 个文件）条件下的语音增强结果。所有基线方法均使用仅噪声数据训练。受控消融实验见表*2*。每种条件下的最佳结果以粗体标示。*SDR* 表示 *SI-SDR*（*dB*）。

方法	仅噪声				多重退化			
	感知语音质量评估	扩展短时客观可懂度	信号失真比	UTMOS	感知语音质量评估	扩展短时客观可懂度	信号失真比	UTMOS
MP-SENet [19]	**3.16**	**0.88**	**19.6**	3.90	2.39	0.62	−0.3	3.00
MetricGAN+ [20]	3.13	0.83	8.6	3.63	2.53	0.62	−5.5	2.82
SGMSE+ [1]	2.94	0.87	17.4	3.83	2.30	0.63	0.0	2.99
SGMSE+（去混响） [1]	1.98	0.78	10.9	3.16	1.99	0.66	−3.3	2.96
NASE [12]	2.75	0.85	17.3	3.83	2.22	0.64	−0.5	2.99
SLICE（我们的方法）	2.83	0.86	17.4	**3.93**	**2.60**	**0.80**	**3.7**	**3.71**

遵循无分类器引导（CFG），在训练期间，每个分支的嵌入都以 $p = 0.1$ 的概率独立置零。这使模型能够在推理时处理缺失的退化类型，因为它已经学会在任意条件分支子集处于激活状态时增强语音。

3. 实验

3.1. 实验设置

3.1.1. 数据集

我们使用 VoiceBank-DEMAND [22, 18] 作为基础数据集，其中包含 11,572 条训练语句，涵盖 10 种噪声类型，信噪比分别为 0、5、10 和 15 dB。为创建多重退化训练数据，我们通过以下组合增强每条干净语句：(i) 来自 DEMAND 的加性噪声；(ii) 使用 pyroomacoustics [23],生成的合成房间脉冲响应 ($T_{60} \in [0.3, 1.0]$ s) 所产生的混响；以及 (iii) 通过软削波 $f(x) = \tanh(\alpha x)/\tanh(\alpha)$ 实现的非线性失真，强度为 $\alpha \in [1.5, 5.0]$。增强后的数据集包含 34,716 条语句，涵盖六种退化类别。

我们在三种不同条件下评估该方法。**仅噪声**测试集包含来自 VoiceBank-DEMAND [24]的 824 条语句。**多重退化**测试集包含 2,472 个文件，涵盖六种退化类别。对于**真实环境**评估，我们使用 VOiCES [25] 数据集，其中包含 1,600 条房间录音；DAPS [26]数据集，其中包含 1,500 条真实设备录音；以及 URGENT [7] 数据集，其中包含 1,000 个经历不同程度退化的文件。

3.1.2. 基线方法

我们将模型与五种仅使用噪声数据训练的基线方法进行比较：SGMSE+ （去噪） [1], SGMSE+ （去混响） [1], MetricGAN+ [20]、MP-SENet [19],和 NASE [12]。为分离编码器和注入方法的影响，我们在相同的多重退化数据上从头训练 SGMSE+架构，不使用任何编码器；此外，还训练完整模型，但采用 NASE 风格的输入相加，而非逐层注入。这三个模型（无编码器、输入相加、逐层注入）使用完全相同的训练数据，从而实现受控比较。

3.1.3. 指标与实现

为符合超越侵入式指标的多维评估建议 [3, 27], ，我们报告感知语音质量评估（PESQ） [28], 、ES-TOI [29], 、SI-SDR [30], 以及神经网络平均意见分预测器 UT-MOS [31]。我们的模型采用 NCSN++ [15] 骨干网络，并结合 SGMSE+ [1] 中的常微分方程形式，同时在整个训练过程中保持 WavLM编码器冻结。模型训练 160 个 epoch，使用 Adam优化器，学习率为 10^{-4} ，指数移动平均衰减为 0.999。我们采用全局批量大小 32，并将其分配到八个 RTX 3090 GPU 上，每个设备分配四个样本。最后，推理阶段使用 30 个反向步数完成。

表2： 在多重退化测试集上评估的消融研究结果。“加法”指由 *[12]*提出的输入级相加方法。

变体	PESQ	扩展短时客观可懂度	SDR	UTMOS
SLICE	**2.60**	**0.80**	**3.7**	**3.71**
无辅助损失（λ=0）	2.58	0.77	2.2	3.63
输入相加	2.49	0.73	1.4	3.62
无编码器	2.60	0.77	2.3	3.70
零条件化	2.14	0.67	−1.6	3.04
均匀权重	2.60	0.80	3.7	3.71

3.2. 主要结果

多重退化（公平比较）。在使用相同数据训练的模型中，注入方法起着决定性作用。即使不使用任何编码器，SGMSE+基线模型也已能较好地处理复合退化，因而构成了一个有力的参照。加入 WavLM 编码器并采用输入级相加（NASE [12], 使用的注入方式）后，性能显著下降至扩展短时客观可懂度 0.73 和 SI-SDR 1.4 dB，甚至不如完全不使用编码器（$p < 10^{-7}$，对 2,472 个文件进行配对 t 检验）；如表 1 所示。相比之下，采用逐层注入的同一编码器则显著提升了性能，扩展短时客观可懂度达到 0.80（$p < 10^{-61}$），SI-SDR 达到 3.7 dB（$p < 10^{-15}$）。这表明，在复合退化场景中，有效利用退化信息的关键在于逐层注入，而非仅仅使用编码器。

在仅噪声数据上训练的基线模型，扩展短时客观可懂度介于0.62与0.66之间，SI-SDR为 ≤0，表现明显更差；而在仅噪声数据上训练的NASE仅达到扩展短时客观可懂度0.64，进一步证实多重退化数据与逐层注入都是必要的。

仅噪声。 在标准仅噪声基准上，专为加性噪声去除设计的模型取得了更高的感知语音质量评估 [9, 32], 采用并行幅度-相位去噪的MP-SENet，以及MetricGAN+ [20] 后者直接将感知语音质量评估作为其训练目标进行优化。值得注意的是，我们的模型在所有模型中取得了最高的UTMOS，甚至超过了MP-SENet [19], 表明尽管采用多重退化训练目标，其感知质量仍然更优。

3.3. 消融研究

表2在多重退化测试集上进一步进行了消融实验。除“无编码器”外，所有变体使用相同的训练数据和 WavLM 编码器，这与受控对比实验一致。

表3： 多重退化测试集中六种不同退化类别的详细性能细分。

Type	N	感知语音质量评估	扩展短时客观可懂度	信号失真比	UTMOS
仅噪声	824	2.82	0.86	17.5	3.93
噪声+混响	330	1.72	0.65	−17.2	3.31
噪声+失真	318	2.84	0.86	15.8	3.94
噪声+混响+失真	338	1.74	0.65	−18.4	3.34
仅混响	341	2.01	0.74	−17.1	3.42
仅失真	321	4.21	0.98	23.3	4.05

表1中的比较。

逐层注入与浅层注入。如表1所示，将我们的逐层注入替换为输入相加 [12],（即仅在输入端将条件向量加到 STFT 表征上），会导致所有指标下降，甚至比完全不使用编码器的效果还差。我们推测有两个原因：（1）在输入层，添加的嵌入会改变复数 STFT 表征，从而干扰网络已学习的频谱处理；（2）这种单次扰动会在 ∼37 残差块中逐渐减弱，导致更深层无法接收到条件信号。逐层注入则通过时间步嵌入调制每一层的仿射变换，避免了这两个问题，同时不改变输入频谱图。这与近期研究发现一致：深层语音增强网络的不同层发挥着不同作用，并且能够从专门的条件信息 [33]中获益。

多任务损失。 将辅助损失权重 λ 设为 0，会使扩展短时客观可懂度从 0.80 降至 0.77，并使 UTMOS 从 3.71 降至 3.63。按退化类型分析发现，这种影响主要集中在混响条件下：对于噪声+混响+失真，不使用辅助损失时，扩展短时客观可懂度从 0.647 降至 0.517，SI-SDR 从 −18.4 dB 降至 −23.8 dB。这证实了多任务监督能够增强编码器的混响辨别能力。

零条件化。 在推理时将 $\mathbf{c}_{\text{extra}}$ 设为 $\mathbf{0}$ 会导致性能大幅下降：感知语音质量评估从 2.60 降至 2.14，UTMOS 从 3.71 降至 3.04，这证实了模型高度依赖条件信号。

自适应加权与均匀加权。 有趣的是，根据编码器的预测置信度动态调整各分支的权重，并未比直接采用均匀权重带来性能提升。这一结果有力地表明，我们的逐层注入机制能够自然地使分数网络判断哪些分支最为重要。因此，推理时添加显式加权并无必要，因为网络已经能够在内部完成这一过程。

3.4 分析

编码器头准确率。 多任务编码器展现出表征各种音频退化的能力。具体而言，其噪声二分类检测准确率达到 96.7%；在混响方面，相关性高达 0.981，且平均绝对误差仅为 0.033。此外，其失真强度相关性达到 0.845，检测准确率也稳健地达到 86.2%。这些出色的指标凸显了辅助损失如何塑造出校准良好的编码器，使其能够准确识别每种退化类型。

按退化类型细分。如表3所示，该模型在不同退化类型下的表现各不相同。它几乎完美地处理了失真，PESQ 达到 4.21，UTMOS 达到 4.05。相比之下，混响会导致 SI-SDR 明显下降。然而，

表4： 在真实场景数据集上的评估。*SGMSE+*（预训练）是公开可用的降噪模型 *[1]*；*SGMSE+* （无编码器）是在与 *SLICE* 相同的多重退化数据上从头训练的，但不使用编码器。由于信号失真比对混响录音不可靠，因此省略了 *SI-SDR*。

数据集	模型	PESQ	扩展短时客观可懂度	UTMOS
VOiCES	SGMSE+ （预训练）	1.20	0.34	1.98
	SGMSE+（无编码器）	**1.59**	0.65	**2.83**
	SLICE	1.56	**0.67**	2.77
DAPS	SGMSE+ （预训练）	2.10	**0.69**	2.96
	SGMSE+（无编码器）	2.04	0.65	2.96
	SLICE	**2.20**	0.60	**3.32**
URGENT	SGMSE+ （预训练）	**1.68**	**0.70**	2.97
	SGMSE+（无编码器）	1.66	0.67	3.01
	SLICE	1.67	0.68	**3.10**

即使在这些具有挑战性的混响条件下，UTMOS 仍高于 3.3，这表明尽管 SI-SDR 较低，感知质量仍得到了较好的保留。这是一项有意义的提升，尤其是考虑到单一退化基线方法在混响环境中的表现要差得多，其 ESTOI 得分仅徘徊在 0.20 至 0.30 之间。

真实场景泛化。 表4评估了三种模型对真实场景中未见录音的泛化能力：SLICE、从头开始在多重退化数据上训练的SGMSE+ （无编码器），以及公开提供、在仅噪声数据上预训练的SGMSE+ 。在VOiCES数据集上，两种多重退化模型的表现都远胜于预训练的SGMSE+ ，这证实了多样化训练数据是实现真实场景数据泛化的主要驱动因素。此外，在全部三个数据集上，SLICE和SGMSE+ （无编码器）的感知语音质量评估和扩展短时客观可懂度得分相当，这表明与受控测试集相比，编码器在域外真实场景录音上的优势并不那么显著。至于感知质量，UTMOS结果呈现出更细致的情况：所提出的模型在DAPS和URGENT上得分最高，而SGMSE+ （无编码器）在VOiCES上的表现略好。总体而言，在感知质量方面，两种多重退化模型都持续且显著地优于仅噪声数据预训练基线。</p>

4. 结论

通过隔离注入方法的受控实验，我们证明条件注入的深度对基于分数的语音增强中退化感知条件的有效性起着重要作用。浅层输入相加是先前工作中的常见方法，在复合退化上的表现可能不如无条件模型。相比之下，将相同的条件注入时间步嵌入可产生更稳健的结果。此外，采用多任务辅助损失的WavLM编码器能够产生校准良好的退化表征，使单一模型能够同时处理噪声、混响和失真。这些结果表明，仅仅存在条件并不能保证性能提升。我们认为，注入方法与条件特征本身同样重要，这一发现对语音增强之外的条件式基于分数的模型也可能具有启示意义。

5. 生成式人工智能使用披露

生成式人工智能工具被用于代码开发辅助和稿件编辑。所有实验设计、分析及科学贡献均为作者的工作。

6. 参考文献

[1] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, 和 T. Gerkmann, “使用基于扩散的生成模型进行语音增强和去混响”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 31 卷，第 2351–2364 页，2023 年。

[2] Y. Yu, and Y. Tsao, “用于语音增强的条件扩散概率模型”，载于*ICASSP*会议论文集。IEEE，2022年，第7402–7406页。

[3] W. Zhang, K. Saijo, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Ku- mar, M. Sach, W. Wang, Y. Fu, S. Watanabe, T. Fingscheidt, and Y. Qian, “从 URGENT 2024 语音增强挑战赛中汲取的经验”，载于*Interspeech*会议论文集，2025年。

[4] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN：基于对抗网络中语音深度特征的高保真去噪与去混响”，载于*Interspeech*会议论文集，2020，第4506–4510页。

[5] H. Liu, Q. Chen, Y. Liu, Z. Yuan, D. Gu, W. Wang 和 M. D. Plumbley，《VoiceFixer：一种用于高保真语音修复的统一框架》，载于*Interspeech* 论文集，2022年，第 4232–4236 页。

[6] H.-T. Chiang 和 J. H. Hansen，《一种面向人工耳蜗使用者语音去混响的可变形卷积生成对抗网络方法》，载于*Interspeech* 论文集，2025，第 833–837 页。[7] K. Saijo、W. Zhang、S. Cornell、R. Scheibler、C. Li、Z. Ni、A. Ku- mar、M. Sach、Y. Fu、W. Wang、T. Fingscheidt 和 S. Watanabe，《Interspeech 2025 URGENT语音增强挑战赛》，载于*Interspeech* 论文集，2025，第 858–862 页。

[8] X. Rong, D. Wang, Q. Hu, Y. Wang, Y. Hu, and J. Lu, “TS-URGENet：一种三阶段通用鲁棒可泛化语音增强网络，”发表于*Interspeech*论文集，2025，第863–867页。

[9] Z. Su, Y. Yu, 与 T. Yei, “超越去噪的扩展：URGENT 挑战赛 2025 中提交的系统与发现”，载于 *Interspeech* 会议论文集，2025，第 873–877 页。

[10] B. Lay、R. Makarov 与 T. Gerkmann，“扩散缓冲器：具有亚秒级延迟的在线基于扩散的语音增强”，载于 *Interspeech* 会议论文集，2025，第 793–797 页。

[11] S. Han, 和 J. Serdyuk, “面向实时语音增强的对抗一致性训练”，载于 *Interspeech* 论文集，2025 年，第 2380–2384 页。

[12] Y. Hu, C. Chen, R. Li, Q. Zhu, 和 E. S. Chng，“噪声感知语音增强使用扩散概率模型”，发表于 *Interspeech* 论文集，2023。

[13] W. Wang，D. Yang，Q. Ye，B. Cao 和 Y. Zou，“NADiffuSE：用于语音增强的噪声感知扩散模型”，发表于 *IEEE ASRU* 论文集，2023。

[14] S. Chen，C. Yu，Z. Ma 等，“BEATs：使用声学分词器进行音频预训练”，发表于国际机器学习大会论文集，2023，第 5178–5193 页。

[15] Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, 和 B. Poole，“通过随机微分方程进行基于分数的生成建模”，*arXiv* 预印本 *arXiv:2011 13456*，2020。。

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, 和 F. Wei，“WavLM：面向全栈语音处理的大规模自监督预训练，”*IEEE*信号处理精选主题期刊，第16卷，第6期，第1505–1518页，2022年。

[17] R. Caruana，“多任务学习”，*Machine Learning*，第28卷第1期，第41–75页，1997。

[18] J. Thiemann、N. Ito 和 E. Vincent，“多样化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”，载于声学会议论文集，第19卷。美国声学学会，2013年。[19] Y.-X. Lu、Y. Ai、Z.-H. Du、J.-W. Zhong 和 Z.-H. Ling，“MP-SENet：一种并行去噪幅度谱和相位谱的语音增强模型”，载于*Interspeech*论文集，2023，第3834–3838页。[20] S.-W. Fu、C. Yu、T.-A. Hsieh、P. Plantinga、M. Ravanelli、X. Lu 和 Y. Tsao，“MetricGAN+：一种用于语音增强的改进版MetricGAN”，载于*Interspeech*论文集，2021，第2167–2171页。[21] J. Ho 和 T. Salimans，“无分类器扩散引导”，*arXiv*预印本 *arXiv:2207.12598*，2022年。[22] C. Valentini-Botinhao、X. Wang、S. Takaki 和 J. Yamagishi，“探究适用于抗噪文本转语音的基于RNN的语音增强方法”，载于第*9*届*ISCA*语音合成研讨会，2016，第146–152页。

[23] R. Scheibler, E. Bezzam, 和 I. Dokmanić, “Pyroomacoustics：用于音频房间仿真和阵列处理算法的Python软件包，”载于*ICASSP*会议论文集，2018，第351–355页。

[24] C. Valentini-Botinhao, X. Wang, S. Takaki, 和 J. Yamagishi，“探讨基于RNN的语音增强方法用于噪声稳健的文本转语音”，载于 第九届 *ISCA* 语音合成研讨会（*SSW 9*），2016 年，第 146–152 页。

[25] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout,P. Gamble, J. Hetherly, C. Stephenson,and K. Ni，“复杂环境中被遮蔽的语音语料库（VOICES）”，2018。

[26] G. J. Mysore，“DAPS（设备与生成语音）数据集”，2014年5月。 [在线]。获取地址：https://doi.org/10.5281/zenodo. 4660670

[27] D. de Oliveira, J. Richter, J.-M. Lemercier, S. Welker 和 T. Gerkmann，“基于干净语音训练的扩散模型用于非侵入式语音质量评估”，载于*Interspeech* 会议论文集，2025年，第 2330–2334 页。

[28] A. W. Rix、J. G. Beerends、M. P. Hollier 和 A. P. Hekstra，“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* 会议论文集，第 2 卷。IEEE，2001 年，第 749–752 页。[29] J. Jensen 和 C. H. Taal，“一种预测调制噪声掩蔽器掩蔽下语音可懂度的算法”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 24 卷，第 11 期，第 2009–2022 页，2016 年。[30] J. Le Roux、S. Wisdom、H. Erdogan 和 J. R. Hershey，“信号失真比——半生不熟还是恰到好处？”，载于*ICASSP* 会议论文集。IEEE，2019 年，第 626–630 页。[31] T. Saeki、D. Xin、W. Nakata、T. Koriyama、S. Takamichi 和 H. Saruwatari，“UTMOS：UTokyo-SaruLab 参加 2022年语音质量挑战赛的系统”，*arXiv* 预印本 *arXiv:2204.02152*，2022年。[32] R. Chao、R. Nasretdinov、Y.-C. F. Wang、A. Jukic、S.-W. Fu 和 Y. Tsao，“结合回归与生成式 Mamba 的通用语音增强”，载于*Interspeech* 会议论文集，2025年，第 888–892 页。[33] V. Parvathala 和 K. S. R. Murty，“用于语音增强的动态层门控”，载于*Interspeech* 会议论文集，2025年，第 2365–2369 页。