掩码自编码器作为通用语音增强器

Rajalaxmi Rajagopalan* RitwikGiri Zhiqiang Tang KyuHan伊利诺伊大学厄巴纳-香槟分校 *AWS AI Labs AWS AI Labs AWS AI Labs*

Gerald Friedland*AWS AI Labs*

arXiv:2602.02413v1 [cs.SD] 2 Feb 2026

摘要——监督式语音增强方法已取得显著成效。然而，在实际场景中，干净语音数据往往匮乏，因此人们需要基于自监督学习（SSL）的语音增强方法，这类方法不仅能达到可比的增强性能，还可应用于其他语音相关的下游应用。在本研究中，我们开发了一种基于掩码自编码器的通用语音增强器，该方法不受影响语音的失真类型限制，能够同时处理多种失真，并以自监督方式进行训练。增强堆栈会向含噪输入数据添加更多失真。在预训练期间，掩码自编码器模型学习去除添加的失真，同时重建声谱图中的掩码区域。随后，微调模型利用预训练嵌入，并通过少量配对数据针对特定下游任务进行训练。我们评估了预训练特征在降噪和去混响下游任务中的表现。我们探索了预训练增强堆栈中的不同增强方式（如单说话人或多说话人），以及不同含噪输入特征表示（如 $log1p$压缩）对预训练嵌入和下游微调增强性能的影响。结果表明，所提方法不仅优于基线，在域内和域外评估数据集上也都达到了最先进的性能。

索引词—**语音增强、自监督学习、零样本评估**

I. 引言

语音增强旨在从受环境噪声、干扰语音、失真等影响的含噪录音中重建干净目标语音。深度学习方法 [1]–[11] 已成功超越传统方法，尤其是在降噪方面。近年来，研究范式已从针对每种语音任务分别构建独立模型，转向使用预训练模型学习通用且实用的语音表征，以提升各种下游任务的表现。当预训练模型采用自监督方式训练时，例如 HuBERT[12] 和 WavLM [13]，预训练-微调模型框架尤其具有吸引力。下游任务主要分为两类：音频分类任务，如说话人识别、语音识别等；以及音频增强任务，如降噪、去混响等。增强类下游任务从通用预训练模型中获益较少。例如，将预训练模型用于关键词检测、说话人识别等分类任务时 [14], [15] 表现出色，但用于音频增强时却并非如此。人们越来越关注如何改进预训练模型，以提升其在增强类下游任务中的表现 [16]–[18]。因此，本研究开发了一种基于掩码自编码器（MAE）的模型，并以自监督方式进行预训练。该模型能够（1）同时应对广泛的失真类型，（2）无需使用大规模干净语音语料进行训练，以及（3）在增强类下游任务中表现稳健。

II. 相关工作

大多数语音增强方法依赖全监督学习，而这需要高质量的干净语音参考信号，但此类信号难以获取。因此，能够放宽这一限制的自监督方法颇具吸引力。HuBERT [12]和 WavLM [13] 等自监督学习方法取得了巨大成功，它们开发了用于说话人识别、分离和说话人分离等各类下游音频任务的预训练模型。目前一项活跃的研究方向，是探索此类自监督预训练模型用于通用语音增强的可行性。

WavLM 和 HuBERT 采用基于分类/预测的目标函数，这与增强任务的连续性不匹配；正如 [16]等研究所强调的，该任务需要估计连续的增强/恢复语音。结果表明，将预训练模型扩展至增强任务的效果有限，无法超越最先进的判别模型（如相位感知 U-Net 语音增强 [11] ）和最先进的扩散模型 [8], [19]。由于我们的主要目标是开发用于语音增强的基础模型，因此我们改用类似掩码自编码器 [20]的回归损失：将含噪短时傅里叶变换输入中的部分区域遮蔽，再由基于视觉 Transformer（ViT）的模型学习重建这些区域，从而学习鲁棒的语音表征。 [17]的作者展示了掩码自编码器预训练对降噪微调的益处。然而，他们在预训练期间没有考虑背景干扰语音、削波、编解码器伪影等其他失真，因此，他们的预训练模型和下游微调模型都无法应对其他失真。

相比之下，我们提出了一种通用自监督语音增强框架，它以掩码自动编码器 [20] 为基础，通过添加一个增强堆栈，加入更多失真，例如基于距离的多说话人语音混合，

*本工作在 AWS AI Labs 实习期间完成。

混响、削波、编解码器伪影等，作用于通用含噪语音输入；预训练模型则学习重建被掩码的部分并去除添加的失真。这一双重目标确保预训练模型是**自监督**且**通用的**。例如，与以往的工作不同， [21],多说话人增强预训练将学习能够去除其他不需要的语音类噪声的特征，例如背景 babble 噪声、背景电视播放的清晰语音，或远离目标说话人的背景干扰说话人。因此，这种预训练方案可提供鲁棒的嵌入，之后只需少量带标签数据进行微调，便可用于不同的下游任务，如降噪、去混响、源分离、带宽扩展等。

III. 所提方法

对于语音片段 $s \in \mathbf{R}^L$，语音失真过程被建模为函数 $d(\cdot)$，退化语音 $x \in \mathbf{R}^L$,$x = d(s)$。语音增强是一个逆问题，旨在从 x 中恢复高质量语音 $\hat{s} = f(x)$，其中 $f(\cdot)$ 是增强函数；即 $d(\cdot)$ 的逆近似。影响 x 的各种失真构成一个复合函数，

$$d(x) = d_1 \circ d_2 \circ \dots d_Q(x) \in \mathbf{D} \tag{1}$$

其中 $\circ$ 表示函数复合，$\mathbf{D} = d_v(\cdot)_{v=1}^{V}$ 是失真类型的集合，$Q \leq V$ 是 $d(\cdot)$ 中失真的数量。在我们的 SSL 通用增强框架中，我们首先以自监督方式在大量噪声语音上对通用模型进行预训练，以学习语音表征；随后，利用这些表征，由在少量成对干净-噪声语音上训练的特定微调模型完成下游任务。

A. 预训练

图 1 展示了预训练的两个主要模块：

(1) 增强堆栈：通过在时域波形级别（背景噪声、混响、编解码器伪影等）和短时傅里叶变换频谱图级别（掩蔽）添加多种失真，对一般含噪语音进一步进行失真处理。(2) 掩码自编码器（掩码自编码器）：一个普通的 ViT-AE [20]模型，其编码器和解码器以增强后的含噪 STFT 幅度为输入，并重建添加增强处理之前的语音。

这是一种自监督学习方法，无需干净的参考信号；通过注入失真，增强处理前的含噪波形便可作为参考信号。

1) 增强堆栈：

波形级增强：

加性噪声：向语音中添加噪声 $n \in \mathbf{R}^L$ s：

$$d(s) = s + n \tag{4}$$

混响：通过将语音信号与房间脉冲响应（RIR）滤波器 r进行卷积来模拟：

$$d(s) = s * r \tag{5}$$

削波：当单个振幅超过范围 [-γ 时， +γ] 会被削波：

$$d(s) = \max(\min(s, \gamma), -\gamma), \quad \gamma \in [0, 1] \tag{6}$$

多说话人混合：将不同说话人的干扰语音添加到目标说话人 s 的语音中。算法1展示了

算法1基于距离的混合生成：2 位说话人

1：目标说话人： s_1，干扰说话人： $s_2$2：为确保所有说话人都在同一个房间内：选取房间脉冲响应 r_0。3：首先， s_1 不能离麦克风太远；4：其次， s_1 必须比 s_2离麦克风更近；5：计算 r_0的直达声与混响声比（DRR）；

$$\text{DRR(dB)} = 10 * log\left(\frac{P_D}{P_R}\right) \tag{2}$$

其中， P_D, P_R 分别为直达路径功率和混响功率。6：**如果** DRR $\geq$ 阈值**则**

7：将 r_0 应用于 s_1， s_1 与麦克风的距离适当8：衰减直达路径 $+$ 的早期反射 r_0： r_0'9：将 r_0' 应用于 s_2， s_2 与麦克风的距离比 $s_1$10：**否则**

11：将 r_0 应用于 s_2， s_2 距离麦克风较远

12: 使用衰减函数对 r_0的后期混响部分进行衰减

$$A(t) = \begin{cases} 1 & t < T_0 \\ \frac{1+\alpha}{2} + \frac{1-\alpha}{2}\cos\frac{\pi(t-T_0)}{(T_1-T_0)} & T_0 < t < T_1 \\ \alpha & t > T_1 \end{cases} \tag{3}$$

T_0 和 T_1 是经衰减的房间脉冲响应的混响部分。13: 将新的 r_0'' 应用于 s_1， s_1 现在比 $s_2$14: **结束 if**

说明对每个说话人应用修改后的房间脉冲响应（RIR）如何生成距离分离的说话人混合。与 WavLM 不同，WavLM 通过确保目标话语比推断出的说话人更长来生成不现实的多说话人混合（这种情况在实践中并不常见）；而在实际场景中，距离分离的说话人混合十分常见，例如目标语音受到背景 babble 噪声、不属于对话的背景干扰说话人、后台播放语音的电视等影响。

$$d(s) = \mathbf{A}\left[s * r_0 \quad s_1 * r_1 \quad \dots \quad s_M * r_M\right]^T \tag{7}$$

其中，$\mathbf{A}$ 是 mix ing matrix，r_i， $i \in \{0, 1, \dots, M\}$ 是每个说话人的房间脉冲响应（RIR）。这种基于距离的线索使模型能够绕过说话人之间的排列问题，且无需使用置换不变训练（PIT）。因此，未经任何微调的预训练模型也可以视为一种免注册的目标语音提取替代方案，其中假定目标语音是距离麦克风最近且 DRR 最高的语音。

频谱级增强：

时间掩蔽：随机掩蔽 x%的时间帧，以模拟丢包隐藏 [20] 失真，即音频片段的某些部分丢失。

频率掩蔽：掩蔽声谱图中 x%的高频单元。这模拟了音频录音中因采样率较低或录音设备缺陷而导致的带宽受限失真。

随机时频掩蔽：声谱图中 x% 的时频单元被随机掩蔽，从而实现自监督。当声谱图的一部分以概率 p 被掩蔽时，会缩短输入序列长度，并促使模型从有限的未掩蔽片段中学习全局上下文化表示。

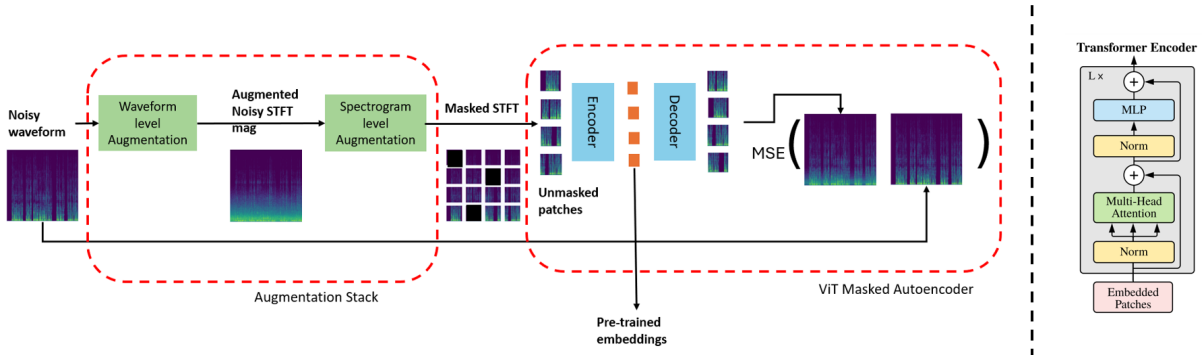


图 1. 预训练框架：(a) 增强堆栈 (b) ViT掩码自编码器 [22]

图 2 展示了应用于输入 STFT 幅度的点积掩码。

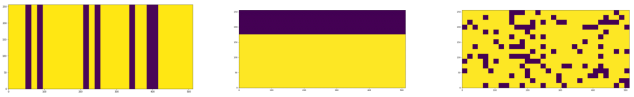


图 2. 频谱图增强：(a) 时间 (b) 频率 (c) 随机时频*2)* 掩码自编码器：音频掩码自编码器使用标准 Transformer 堆栈作为编码器和解码器。编码后的图块会填充可训练的掩码标记，然后传递给解码器。解码器处理嵌入和掩码标记，以重建输入。借助解码器，编码器只需在未掩蔽片段上进行训练，从而降低计算成本。基于图像的掩码自编码器中的全局自注意力不适用于音频频谱图，因为其时频上下文主要是局部的。因此，音频掩码自编码器通过移位窗口，利用 Swin Transformer 解码器块采用局部自注意力机制 [23]。

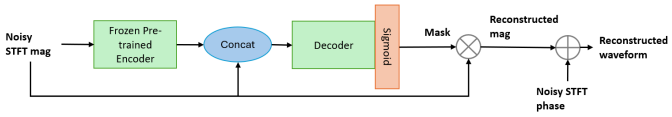


图 3. 微调：增强

B. 微调

现在可以将预训练模型针对特定的增强相关下游任务进行微调。微调模型可使用少量配对数据。图3表明，微调模型的输入是预训练模型编码器嵌入与含噪输入短时傅里叶变换的拼接。输入数据为受到背景噪声、背景语音和多人嘈杂语音以及混响干扰的单说话人语音，期望输出为干净的无混响语音。模型输出为一个掩码 $\in [0,1]$，用于计算每个时频单元中干净语音目标的含量。将该时频掩码（请勿与预训练混淆）与含噪输入短时傅里叶变换相乘，即可得到干净重建。噪声相位用于生成原始重建波形。

多说话人预训练增强也支持源分离这一项下游任务。声谱图增强还可用于其他下游任务，例如（1）带宽扩展：通过预训练模型重建频率掩码增强所掩蔽的高频部分，可实现带宽扩展。（2）数据包丢包隐藏（*PLC*）：通过网络电话传输时丢失或损坏的语音帧，可通过时间掩蔽增强预训练来恢复，该方法能够重建被掩蔽的时间帧[20]。我们将源分离、带宽扩展和丢包隐藏的详细分析留待未来工作。

IV. 结果

A. 实验设置

1) 掩码自编码器预训练：在 LibriTTS[29] 上使用全部三个训练子集（**960 小时**）进行预训练。ViT-AE 预训练模型的超参数取自 [14]。输入音频数据会裁剪或填充为 4 秒片段，采样率为 16千赫。我们使用 32 毫秒的短时傅里叶变换窗和 8 毫秒的步长。ViT-AE 使用 AdamW 优化器进行 60 个训练轮次的预训练，批量大小为 256，权重衰减为$1e^{-4}$，学习率（LR）为 $1e^{-4}$，在预热阶段（5 个训练轮次）线性增至 $1e^{-4}$，随后通过余弦退火调度器 [20] 降至 $1e^{-6}$。在增强堆栈中，每种增强均以概率 p 应用。首先，将输入语音缩放至随机响度级别 $[-30, 10]$ dB；随后，通过从 LibriTTS 中选取次要干扰说话人，并从 Open SLR28 [30] 中选取一个房间脉冲响应，进行多说话人混合生成（算法1）。接着，选择一种音频编解码器，加入编解码器伪影。通过选取一个级别 $[0, 1]$ 进行随机削波；最后，从 DNS噪声 [31] 中选取噪声，并以随机信噪比 $[-30, 0]$ dB 添加背景噪声。对于频谱图级增强，以概率 $p = [0.1, 0.1, 0.8]$ 为每个输入片段选择以下三种掩码之一：时间（屏蔽 20% 的时间帧）、频率（随机屏蔽至多 50% 的高频单元）和随机时频（屏蔽 75% 的时频单元）。

2) 微调：冻结预训练的 ViT 编码器，并按照 [14] 进行时间对齐后，将编码器的补丁嵌入与输入的含噪 STFT 幅度拼接。与预训练不同，微调解码器采用全局注意力，批量大小为 256，学习率在预热阶段（5 个训练轮次）升至 $2e^{-4}$，并通过余弦退火在 95 个训练轮次内降至 $1e^{-6}$。模型的输出是一个时频掩码，经 Sigmoid 函数限制在 [0, 1]。我们在 Valentini 数据集 [27]（**29 小时**）上微调预训练的 ViT-AE 模型；测试集包含 824 个噪声-干净配对片段。在 DAPS [28] 数据集的 M10 和 F10 说话人上进行零样本评估，该数据集

预训练增强类型	模型	微调	CSIG	CBAK	COVL	PESQ	SSNR	NISQA
-	含噪	-	3.34	2.44	2.63	1.97	1.68	3.05
-	DCUNet [11]	-	4.24	**4.00**	3.69	3.13	-	-
-	UNIVERSE [19]	-	4.03	3.11	3.46	2.9	-	-
-	SGMSE+ [8]	-	4.25	3.4	3.61	2.94	-	**4.56**
-	VoiceFixer [24]	-	-	-	-	2.43	-	-
-	PANN 搭配 UNet [18]	-	2.43	2.96	2.3	2.28	-	-
-	Huang等人 [25]	-	-	-	-	2.68	-	-
-	Boosting SSL [26]	Y	4.4	3.52	3.74	3.05	-	-
随机时频掩码	**用于修复的掩码自编码器**[1] [17]	Y	4.28	2.95	3.51	3.25	11.34	4.28
单说话人	Ours	Y	4.31	3.02	3.68	3.33	11.35	4.13
多说话人	Ours	N	4.22	2.75	3.36	3.19	11.35	4.14
多说话人	Ours	Y	4.28	2.99	3.69	3.32	11.35	4.12
多说话人+log1p	Ours	N	4.3	2.75	3.4	3.23	11.35	4.14
多说话人+log1p	Ours	Y	**4.41**	3.21	**3.78**	**3.4**	**11.49**	4.22

表 I 域内增强结果：VALENTINI [27]

预训练增强类型	模型	微调	CSIG	CBAK	COVL	PESQ	SSNR	NISQA
-	含噪	-	2	1.52	1.6	1.47	1.42	2.79
-	UNIVERSE [19]	-	2.19	-	1.76	1.51	-	-
-	SGMSE+ [8]	-	**2.81**	-	**2.33**	1.97	-	**4.44**
随机时频掩码	**用于修复的掩码自编码器**[1] [17]	Y	2.49	-	2.05	1.81	-	2.95
多说话人	Ours	N	2.34	1.75	1.99	1.78	7.24	2.71
多说话人	Ours	Y	2.57	2.3	2.21	2.09	8.51	3.43
多说话人+log1p	Ours	N	2.45	1.8	2.1	1.8	7.24	2.74
多说话人+log1p	Ours	Y	2.64	**2.51**	2.32	**2.17**	**9.1**	3.5

表 II 域外增强结果：DAPS [28]

[1] 表 I 和表 II 中的这些数字，是通过复现 [21] 得到的；复现时仅在预训练阶段使用随机时频掩蔽，并采用短时傅里叶变换。

通过在含噪且混响的环境中播放预先录制的干净语音生成。将用于音频恢复的音频掩码自编码器[17]用作比较基线。

B. 评估

评估时，我们在 Valentini 和 DAPS 数据集上使用标准侵入式指标（感知语音质量评估、CSIG、CBAK、COVL、段信噪比 [32]）以及非侵入式指标 NISQA [33] 。表 I 中，我们复现了基线：音频修复的掩码自编码器（表 3 第 4 [17] 行），仅使用声谱图级别的 75% 随机时频掩蔽进行预训练增强。数值略有差异，是因为基线在预训练中使用对数梅尔频谱，而本文使用短时傅里叶变换。我们发现，除多说话人混合外，应用所有增强（波形、声谱图）后，模型在域内 Valentini 数据上的微调性能有所提升。加入多说话人增强后，性能与单说话人增强相当，因为它有助于去除未扩散的背景多人嘈杂语音、背景干扰语音等。因此，完整的增强预训练能够学习到适用于广泛下游任务的可靠特征。我们通过仅在预训练模型上直接评估来验证微调的必要性，并观察到性能下降。借鉴 Boosting SSL [26], 的思路，我们用其 $log1p$ 压缩特征替代线性短时傅里叶变换。视觉变换器掩码自编码器的输入是含噪短时傅里叶变换幅度的 $log1p$ 压缩版本，均方误差损失（图 1）则针对 $log1p$ 特征计算。表 I 显示，使用 $log1p$ 压缩可提升仅预训练模型和预训练+微调模型的各项数值。我们的模型在 4 项性能指标中的 3 项上优于基于分类的自监督学习预训练 [25] 和最先进方法（Boosting SSL [26], 该方法在微调时使用 WavLM[13], HuBERT [12] 预训练的 $log1p$ 压缩特征）。我们的表现优于监督式预训练方法（如使用 UNet [18] 的 PANNs）则凸显了自监督学习通过增强（数据增强）提高训练数据可用性和多样性的优势。采用多说话人增强预训练进行微调（线性和 $log1p$）的性能在除 CBAK 外的所有指标上均优于最先进的判别模型 DCUNet [11]。对于 UNIVERSE [19] 和 SGMSE+ [8] 等基于扩散的模型，我们在侵入式指标上的表现更好，但在 NISQA 等非侵入式指标上的表现较差。

表 II 显示，与基线 [17] （表 5 第 13 行）相比，增强预训练在域外泛化（DAPS [28]）方面表现出色；使用 $log1p$ 压缩特征后，性能进一步提升。因此，通过增强学习失真去除的自监督学习预训练模型，能够为域内和域外数据的下游增强任务提供鲁棒泛化特征。

V. 局限性与结论

我们设计了一种通用的自监督预训练模型，可为各种语音增强下游任务提供稳健的语音特征。我们表明，在掩码自编码器预训练中加入失真去除目标，既能提升增强性能，也能学习适用于其他下游应用的特征。因此，该预训练模型可用于通用语音增强，去除污染语音的各类失真，并作为 WavLM 的一种增强方案。我们考虑通过多说话人增强来去除背景语音的增强任务，同时希望未来探索基于距离的混合语音如何助力下游更困难的混合语音分离（不使用距离线索），以及如何利用时间与频率掩蔽的频谱图增强实现零样本带宽扩展和丢包隐藏。

参考文献

[1] X. Hao、X. Su、R. Horaud 和 X. Li，“Fullsubnet：一种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声学、语音与信号处理国际会议（ICASSP）。IEEE，2021，第6633–6637页。[2] A. Defossez、G. Synnaeve 和 Y. Adi，“波形域实时语音增强”，arXiv预印本 arXiv:2006.12847，2020。[3] Y. Zhao、Z.-Q. Wang 和 D. Wang，“用于含噪混响语音增强的两阶段深度学习”，IEEE/ACM音频、语音与语言处理汇刊，第27卷，第1期，第53–62页，2018。[4] H. Zhao、S. Zarar、I. Tashev 和 C.-H. Lee，“用于语音增强的卷积循环神经网络”，载于 2018年 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2018，第2401–2405页。[5] Y. Koizumi、K. Yatabe、M. Delcroix、Y. Masuyama 和 D. Takeuchi，“利用自适应和多头自注意力进行语音增强”，载于 ICASSP 2020-2020 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2020，第181–185页。[6] U. Isik、R. Giri、N. Phansalkar、J.-M. Valin、K. Helwani 和 A. Krishnaswamy，“Poconet：利用频率位置嵌入、半监督对话数据和偏置损失改进语音增强”，arXiv预印本 arXiv:2008.04470，2020。[7] A. Pandey 和 D. Wang，“用于时域语音增强的带自注意力密集卷积神经网络”，IEEE/ACM音频、语音与语言处理汇刊，第29卷，第1270–1279页，2021。[8] J. Richter、S. Welker、J.-M. Lemercier、B. Lay 和 T. Gerkmann，“利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。[9] S.-W. Fu、C.-F. Liao、Y. Tsao 和 S.-D. Lin，“Metricgan：基于生成对抗网络的语音增强黑盒度量分数优化”，载于国际机器学习会议。PMLR，2019，第2031–2041页。[10] S. Pascual、A. Bonafonte 和 J. Serra，“Segan：语音增强生成对抗网络”，arXiv预印本 arXiv:1703.09452，2017。[11] H.-S. Choi、J.-H. Kim、J. Huh、A. Kim、J.-W. Ha 和 K. Lee，“利用深度复数U-Net进行相位感知语音增强”，载于国际学习表征会议，2018。[12] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习”，IEEE/ACM音频、语音与语言处理汇刊，第29卷，第3451–3460页，2021。[13] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao 等，“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。[14] D. Niizumi、D. Takeuchi、Y. Ohishi、N. Harada 和 K. Kashino，“利用掩码自编码器进行掩码频谱图建模，以学习通用音频表征”，载于 HEAR：音频表征整体评估。PMLR，2022，第1–24页。[15] D. Chong、H. Wang、P. Zhou 和 Q. Zeng，“用于自监督音频预训练的掩码频谱图预测”，载于 ICASSP 2023-2023 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2023，第1–5页。[16] H. Song、S. Chen、Z. Chen、Y. Wu、T. Yoshioka、M. Tang、J. W. Shin 和 S. Liu，“探索 WavLM 在语音增强中的应用”，载于2022年 IEEE口语语言技术研讨会（SLT）。IEEE，2023，第451–457页。[17] Z. Zhong、H. Shi、M. Hirano、K. Shimada、K. Tateishi、T. Shibuya、S. Takahashi 和 Y. Mitsufuji，“将音频掩码自编码器扩展至音频修复”，载于2023年 IEEE信号处理在音频与声学中的应用研讨会（WASPAA）。IEEE，2023，第1–5页。[18] Q. Kong、H. Liu、X. Du、L. Chen、R. Xia 和 Y. Wang，“利用来自 AudioSet 的弱标注数据进行语音增强”，arXiv预印本 arXiv:2102.09971，2021。[19] J. Serr`a、S. Pascual、J. Pons、R. O. Araz 和 D. Scaini，“利用基于分数的扩散实现通用语音增强”，arXiv预印本 arXiv:2206.03065，2022。[20] P.-Y. Huang、H. Xu、J. Li、A. Baevski、M. Auli、W. Galuba、F. Metze 和 C. Feichtenhofer，“能够聆听的掩码自编码器”，神经信息处理系统进展，第35卷，第28 708–28 720页，2022。

[21] J. Zhang、S. Jayasuriya 和 V. Berisha，“通过改进的扩散模型修复退化语音”，*arXiv* 预印本 *arXiv:2104.11347*，2021 年。[22] D. Alexey，“一张图像胜过 16×16 个单词：用于大规模图像识别的 Transformer”，*arXiv* 预印本 *arXiv: 2010.11929*，2020 年。[23] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo，“Swin Transformer：使用移位窗口的分层视觉 Transformer”，载于*IEEE/CVF* 计算机视觉国际会议论文集，2021 年，第 10 012–10 022 页。[24] H. Liu、Q. Kong、Q. Tian、Y. Zhao、D. Wang、C. Huang 和 Y. Wang，“Voicefixer：借助神经声码器实现通用语音修复”，*arXiv* 预印本 *arXiv:2109.13731*，2021 年。[25] Z. Huang、S. Watanabe、S.-w. Yang、P. Garc´ıa 和 S. Khudanpur，“探究自监督学习在语音增强与分离中的应用”，载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2022 年，第 6837–6841 页。[26] K.-H. Hung、S.-w. Fu、H.-H. Tseng、H.-T. Chiang、Y. Tsao 和 C.-W. Lin，“增强用于语音增强的自监督嵌入”，*arXiv* 预印本 *arXiv:2204.03339*，2022 年。[27] C. V. Botinhao、X. Wang、S. Takaki 和 J. Yamagishi，“使用深度循环神经网络为抗噪文本转语音合成系统进行语音增强”，载于*Interspeech 2016*，2016 年，第 352–356 页。[28] G. J. Mysore，“我们能否将现实环境中由常见消费电子设备录制的语音自动转换为专业制作水准的语音？——数据集、见解与挑战”，*IEEE* 信号处理快报，第 22 卷，第 8 期，第 1006–1010 页，2014 年。[29] H. Zen、V. Dang、R. Clark、Y. Zhang、R. J. Weiss、Y. Jia、Z. Chen 和 Y. Wu，“LibriTTS：一个由 LibriSpeech 派生的文本转语音语料库”，*arXiv* 预印本 *arXiv:1904.02882*，2019 年。[30] T. Ko、V. Peddinti、D. Povey、M. L. Seltzer 和 S. Khudanpur，“面向鲁棒语音识别的混响语音数据增强研究”，载于*2017 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2017 年，第 5220–5224 页。[31] Y. Chen、M. Antonakakis、R. Perdisci、Y. Nadji、D. Dagon 和 W. Lee，“DNS噪声：衡量一次性域名在现代 DNS 流量中的普遍程度”，载于*2014* 年第 *44* 届 *IEEE/IFIP* 可靠系统与网络国际会议。IEEE，2014 年，第 598–609 页。[32] “https://github.com/schmiph2/pysepm。”[33] G. Mittag、B. Naderi、A. Chehadi 和 S. M¨oller，“NISQA：一种结合 CNN 与自注意力的深度模型，利用众包数据集进行多维语音质量预测”，*arXiv* 预印本 *arXiv:2104.09494*，2021 年。