基于多粒度向量量化的语音增强 n

Xiao-Ying Zhao[1],*Qiu-Shi Zhu*[2],*Jie Zhang*[2]

[1]中国科学技术大学（USTC），中国合肥[2]NERC-SLIP，中国科学技术大学（USTC），中国合肥xyzhao1123@mail.ustc.edu.cn, qszhu@mail.ustc.edu.cn, jzhang6@ustc.edu.cn

arXiv:2302.08342v1 [eess.AS] 16 Feb 2023

摘要

随着深度学习的发展，过去十年间，基于神经网络的语音增强（SE）取得了快速进展。与此同时，自监督预训练模型和向量量化（VQ）在许多语音相关任务中表现出色，但在语音增强领域的探索相对较少。我们先前的工作表明，使用矢量量化模块对噪声语音表示进行离散化有助于语音去噪。因此，本研究探讨了在不同层使用VQ以及采用不同数量码本的影响。不同的VQ模块确实能够提取多粒度语音特征。通过注意力机制，将预训练模型提取的上下文特征与编码器提取的局部特征融合，从而保留全局和局部信息，以重构增强语音。在Valentini数据集上的实验结果表明，所提出的模型能够提升语音增强性能，同时也揭示了预训练模型选择所带来的影响。

索引词：语音增强、矢量量化、表征学习、自监督预训练。

1. 引言

语音增强（SE）旨在减轻噪声语音录音中包含的背景噪声和混响，通常被广泛配置为语音识别 [1, 2]，辅助听觉 [3]，鲁棒说话人识别 [4]等系统的前端。传统的基于统计的语音增强方法通常依赖平稳性假设，然而这一假设在实践中很难成立。近年来，深度神经网络（DNN）的发展为实现语音增强提供了新的方向，其性能甚至优于传统方法，尤其是在非平稳场景中。总体而言，基于DNN的语音增强模型可以采用生成式 [5, 6] 或判别式 [7, 8]方式实现。生成模型旨在学习语音分布：将含噪语音输入模型，以生成估计的语音信号，同时对增强语音和干净语音的分布施加一致性约束。后者则通过训练判别模型，直接最小化增强信号与干净信号之间的差异，从而消除环境噪声对频谱的影响。判别模型可以在频域 [9, 10] 或时域 [7, 11], 中实现，其中，频域语音增强通过估计掩蔽矩阵来完成，而时域语音增强通常遵循编码器-解码器架构实现。

离散表示学习已在许多领域展现出成效，尤其适用于低资源任务，其中常使用矢量量化（VQ）。例如，在 [12,13] VQ 被用于学习离散潜在表示，以生成高质量图像。在 [14, 15], 中，VQ 通过学习离散表示用于语音合成，从而生成更自然的语音。在 [16], 中，可扩展矢量量化被用于编码多尺度特征，以实现语音编码。考虑到学习离散表示能够产生有意义的离散单元并减少噪声成分， [17] 表明，VQ 还可通过从含噪离散表示中重建语音用于语音增强。然而，研究也表明，采用单一固定码本的 VQ 无法充分学习不同粒度的聚类单元，这限制了其在语音增强中的适用性。

在语音领域，自监督预训练模型 [18–23]，（例如 wav2vec2.0 [18]，、HuBERT [20], 和 data2vec [21]，）利用大量无标注数据迅速发展，并已证明有助于各种下游任务。它们的区别在于，wav2vec2.0、HuBERT 和 data2vec 分别执行对比学习、掩码预测和回归任务。采用自监督预训练模型也能提升语音增强性能。在 [24], , 研究者将十三种预训练模型应用于语音增强任务，并将其用作特征提取器来估计频谱掩码，从而重建干净语音波形。这表明，自监督预训练模型提取的高层特征适用于语音增强，也可以与传统声学特征 [25] 结合。这些预训练特征都用于时频掩码的估计，但尚未在基于 U-Net 的语音增强模型中得到验证。

本文基于 [17]，（其使用单个带有固定码本的向量量化模块），探索如何更好地将向量量化与预训练模型结合，应用于语音增强任务。鉴于语音增强模型的不同层学习不同粒度的信息，我们首先提出一种多粒度矢量量化技术，通过在不同层使用不同的向量量化来建模不同层级的信息。具体而言，不同的向量量化模块具有不同数量的码本和可学习码向量。与 [17], 相比，结果表明，使用多粒度向量量化能够提升性能，同时仅小幅增加模型参数。其次，为了使模型更好地利用全局和局部信息，我们将预训练模型提取的上下文特征与语音增强模型编码器提取的局部特征进行融合。上下文特征和局部特征均输入 U-Net 框架。这不同于 [24, 25], 中的方案，后者仅将预训练模型提取的特征用于估计时频掩码；相比之下，本文方法所采用的特征更加丰富。此外，结果还表明，不同预训练模型对性能的影响各不相同，其中 data2vec 模型在参数量几乎相同的情况下表现最佳。

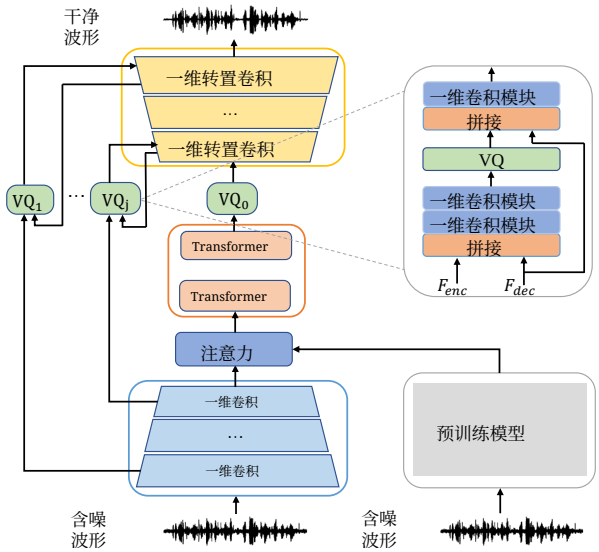

图1：所提出的语音增强模型示意图，其中包含多粒度*VQ*模块和自监督预训练模型。所提出的多粒度*VQ*的特征融合机制如右上角所示。

Valentini 数据集上的实验结果表明，在不同层使用多粒度向量量化有利于语音增强。本文其余部分安排如下：第 2 节介绍所提出的语音增强方法；第 3 节和第 4 节分别展示实验设置和结果；最后，第 5 节总结本文工作。

2. 方法

为便于读者理解，本节首先简要介绍作为所提方法基础的 [17], 语音增强模型。该语音增强模型遵循U-Net结构，包含基于卷积的编码器和解码器，以及基于Transformer 编码器的瓶颈层。这与DEMUCS网络类似。向量量化模块用于将瓶颈层输出的表征离散化，随后将其输入解码器以重建干净语音波形。第4节表明，向量量化模块和预训练模型均有助于提升性能。然而，我们发现，单一向量量化无法学习不同粒度的语音表征。更多细节请参见 [17]。

2.1. 模型结构

我们提出的模型结构如图1所示，主要包含编码器、瓶颈层、多粒度VQ和解码器。编码器和解码器分别包含 D层一维（1D）卷积和 D层1D转置卷积，瓶颈层由N层Transformer编码器组成。在第 i个编码器层与第 i个解码器层之间采用了跳跃连接。编码器采用卷积模块，能够高效提取局部特征。为了更好地利用预训练模型建模长时间上下文信息，我们将预训练模型作为特征编码器。局部特征和全局特征被输入注意力网络进行融合。融合后的特征被送入瓶颈层，其输出在经过量化器 D离散化后被输入解码器。为了融合不同粒度的离散单元，我们采用了不同的向量量化器。具体而言，第 i个编码器层的输出与第 i个解码器层的输出融合为 VQ_i ，以获得第 i个离散表示，随后再与第 i个解码器层的输出进行融合。

2.2. 多粒度VQ

所提出的多粒度VQ机制如图1右上角所示，主要包含融合VQ模块。首先，我们将第 i个编码器层的输出与第 i个解码器层的输出拼接。将拼接后的特征输入两层一维卷积，得到融合特征，再将其输入VQ模块以获得离散特征。随后，将离散特征与解码器输出拼接，并输入一维卷积模块，以构建最终输出。VQ模块基于乘积量化构建，并针对不同层考虑不同数量的码本和码字。给定G 个码本，每个码本包含 V个可学习的 d维码字，我们首先将输入特征映射为 $\boldsymbol{l} \in \mathbb{R}^{G\times V}$ 个logit，然后通过Gumbel-Softmax [26] 操作，以可微方式选取离散向量。从第 g个码本中选择第 v个码字的概率为

$$\overline{p}_{g,v} = \frac{\exp(\bar{l}_{g,v} + n_v)/\tau}{\sum_{k=1}^{V}\exp(\bar{l}_{g,k} + n_k)/\tau}, \tag{1}$$

其中 $n_v = -\log(-\log(u))$，其中 u 在[0, 1] 上均匀分布，且 τ是一个非负温度系数。在前向阶段，通过 $argmax$ 操作选择代码向量；在反向传播阶段，则使用 Gumbel-Softmax 输出的真实梯度。由于通过最大化 softmax 分布 l 可以使更多码字用于量化，我们考虑多样性损失函数 L_d，其定义为

$$L_d = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V}\overline{p}_{g,v}\log\overline{p}_{g,v}. \tag{2}$$

2.3. 总损失函数

$$L_{\text{total}} = L_{\text{se}} + \sum_{i=0}^{D}\lambda_i L_{d_i}, \tag{3}$$

其中 L_{d_i} 是第 i 个向量量化的码本多样性损失，λ_i 是权重参数，而 L_{se} 是语音增强损失函数，其定义为 [7, 17]

$$L_{\text{se}} = \frac{1}{T}||\boldsymbol{y} - \hat{\boldsymbol{y}}||_1 + \sum_{i=1}^{M}L_{\text{stft}}^{(i)}(\boldsymbol{y}, \hat{\boldsymbol{y}}), \tag{4}$$

其中

$$L_{\text{stft}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\text{sc}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + L_{\text{mag}}(\boldsymbol{y}, \hat{\boldsymbol{y}}), \tag{5}$$

$$L_{\text{sc}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{|||\text{STFT}(\boldsymbol{y})| - |\text{STFT}(\hat{\boldsymbol{y}})|||_F}{||\text{STFT}(\boldsymbol{y})||_F}, \tag{6}$$

$$L_{\text{mag}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{T}||\log|\text{STFT}(\boldsymbol{y})| - \log|\text{STFT}(\hat{\boldsymbol{y}})|||_1, \tag{7}$$

其中，T 表示语音长度，$\boldsymbol{y}$ 和 $\hat{\boldsymbol{y}}$ 分别表示干净语音信号和增强语音信号，M 表示多尺度 STFT 损失函数的数量，$||\cdot||_F$ 表示 Frobenius 范数

表1：性能　在以下数据集上的客观评价指标比较　　*Valentini*测试集。

模型	领域	感知语音质量评估（宽带）	短时客观可懂度 (%)	预测 CSIG	预测 背景噪声分量	预测 总体质量分量
含噪	-	1.97	92.1	3.35	2.44	2.63
SEGAN [27]	波形	2.16	-	3.48	2.94	2.80
Wave U-Net [28]	波形	2.40	-	3.52	3.24	2.96
SEGAN-D [29]	波形	2.39	-	3.46	3.11	3.50
MMSE-GAN [30]	时频	2.53	93.0	3.80	3.12	3.14
MetricGAN [31]	波形	2.86	-	3.99	3.18	3.42
DeepMMSE [32]	波形	2.95	94.0	4.28	3.46	3.64
DEMUCS [7]	波形	3.07	95.0	4.31	3.40	3.63
CleanUNet [8]	波形	3.09	95.8	4.38	3.47	3.69
我们之前的 [17]	波形	3.13	96.1	4.46	3.56	3.82
Ours	波形	3.18	96.7	4.53	3.64	3.85

以及 $||\cdot||_1$ L_1 范数。多分辨率损失 $L^{(i)}$ 短时傅里叶变换 使用短时傅里叶变换损失，其中快速傅里叶变换频点的数量范围为 {512,1024, 2048}，跳跃长度范围为 {50, 120, 240}，窗口长度范围为 {240, 600, 1200}。

3. 实验设置

3.1. 数据描述

为评估所提方法的有效性，我们采用 Valentini [33] 数据集。该数据集包含 28.4 小时的干净-含噪语音对，其中训练集包含 34647 个单声道音频样本，测试集包含 824 个单声道音频样本，采样率为 48 kHz。音频时长范围为 1 秒至 15 秒（平均时长约为 3 秒）。这些数据采集自 84 名说话人，训练集中的干净语音添加了 10 种不同类型的噪声，信噪比（SNR）分别为 0、5、10 和 15 dB；测试集中的语音添加了 5 5、12 4 2 5、7 5 17 5 种噪声类型，信噪比为（. . . 和 . dB），且所有说话人均以英语为母语。原始语音波形经过降采样至 16kHz 后，按照 [7]中的方法应用重混音和频带掩码技术进行数据增强。

所提出方法的性能通过客观指标进行评估，包括：1）语音质量感知评估（PESQ） [34], 2；2）短时客观可懂度（STOI） [35], 3；3）语音信号失真（SIG）的平均意见分（MOS）预测 [36], 4；4）背景噪声侵扰性（BAK）的MOS预测 [36], 5；5）整体质量（OVRL）的MOS预测 [36]。

3.2. 模型配置

我们的网络基于 DEMUCS 模型 [7],构建，该模型通过将 LSTM 替换为 Transformer，并使用注意力模块作为特征融合模块进行了修改。编码器和解码器均包含 $D = 5$ 层，维度为 512。我们在瓶颈层使用了 2 个 Transformer 编码器，每个编码器的维度为 512，前馈神经网络的维度为 2048，自注意力模块包含 12 个头。编码器中的一维卷积使用大小为 8 的卷积核，步长为 2。对于预训练模型，我们在实验中使用了公开可用的 wav2vec2.0 base[1], HuBERT base[2] 和 data2vec base[3] 模型。这三个模型均通过使用相同的无标注数据进行自监督预训练获得，并且具有相近的参数量。这些模型参数不用于训练，我们仅使用它们提取特征。对于注意力融合模块，我们采用了维度为 512 的交叉注意力层。对于 VQ_0 和 VQ_i(
$i = 1$，2、3、4、5)模块，我们分别采用了包含 $G = 2$ 个码本以及包含 $V = 320$个码本的结构，其中每个码本分别包含 $G = 1$个具有 $V = \{320, 640, 960\ 2560, 5120\}$ 可学习的 128 维向量。式 (3) 中的 λ_i 设为 0.01，这是由于多样性损失的取值范围所致，因为码本多样性损失函数 L_d 的幅值约为 SE 损失函数 L_{se}的 100 倍。我们使用 Adam 优化器训练语音增强模型 1M 次迭代，批量大小为 30，最大学习率为 2×10^{-4}。所有模型均在 4 个 Tesla-V100-32G GPU 上进行训练。

[1]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec [2] https://github.com/facebookresearch/fairseq/tree/main/examples/hubert [3] https://github.com/facebookresearch/fairseq/tree/main/examples/data2vec

4. 实验结果

对比方法：我们首先在含噪测试集上测量客观评价指标，作为参考。时域对比方法包括基于生成对抗网络（GAN）的SEGAN [27], 、SEGAN-D [29] 和MetricGAN [31], 。时频域对比方法包括MMSE-GAN [30] 和DeepMMSE [32],，后者依据最小均方误差（MMSE）准则估计噪声谱密度。此外，还比较了Wave U-Net [28], 、DEMUCS [7] 和 CleanUNet [8]，这些方法依赖于时域中的U-net网络。

表1展示了不同模型在 Valentini 数据集上的实验结果。可以看出，所提出的方法结合预训练模型和多粒度VQ，分别取得了3.18的感知语音质量评估分数和96.7的短时客观可懂度，在所有对比方法中表现最佳，也优于我们之前的方法 [17], ；后者采用固定码本，并使用预训练模型初始化编码器。这表明，相较于单一固定码本，采用多粒度VQ模块和预训练模型作为特征提取器更加有效。此外，部署多个码本仅使参数量略有增加。（码本中的码向量是可学习的，因此新增的模型参数数量可计算为（$320+640+960+2560+5120+320\times2）\times128/10^6 \approx 1.3M$。）为了更清楚地了解各个模块的影响，我们进行了对比实验，如表2所示。基础模型不使用VQ或预训练模型，感知语音质量评估分数为3.07，短时客观可懂度为95.2，与CleanUNet的结果相近，因为二者的模型结构十分相似。如图1所示，在基础模型中加入多粒度VQ模块后，各项指标上的语音增强性能均有明显提升，表明使用多个码本是有效的。加入多个VQ，并使用任意一种预训练模型（wav2vec2.0、HuBERT或data2vec）作为特征提取器（即将提取的上下文特征与语音增强编码器提取的局部特征融合）后，性能还能在一定程度上进一步提升；其中，data2vec模型提取的特征效果似乎最佳，而使用wav2vec和HuBERT带来的收益几乎相同。

表 2：不同模块对语音增强性能的影响。

模型	PESQ (WB)	STOI (%)	预测 CSIG	预测 CBAK	预测 COVL
Base	3.07	95.2	4.31	3.43	3.64
+ 多重向量量化	3.12	96.2	4.42	3.53	3.77
+Wav2vec2.0	3.14	96.3	4.48	3.58	3.83
+HuBERT	3.14	96.4	4.48	3.59	3.83
+Data2vec	3.18	96.7	4.53	3.64	3.85

表 3： 不同层的 *VQ* 对性能的影响。

VQ0	VQ1	VQ2	VQ3	VQ4	VQ5	PESQ (WB)	STOI (%)
✓	✓	✓	✓	✓	✓	3.12	96.2
✗	✓	✓	✓	✓	✓	3.09	95.6
✓	✗	✓	✓	✓	✓	3.12	96.0
✓	✓	✗	✓	✓	✓	3.11	96.0
✓	✓	✓	✗	✓	✓	3.10	95.9
✓	✓	✓	✓	✗	✓	3.10	95.8
✓	✓	✓	✓	✓	✗	3.10	95.8

使用任意一种预训练模型（wav2vec2.0、HuBERT或data2vec）作为特征提取器（即将提取的上下文特征与语音增强编码器提取的局部特征融合）后，性能还能在一定程度上进一步提升；其中，data2vec模型提取的特征效果似乎最佳，而使用wav2vec和HuBERT带来的收益几乎相同。

此外，我们在表 3 中通过实验分析了不同层的 VQ 所产生的影响。显然，当每一层都包含 VQ 模块时，语音增强性能最佳。移除任意一层的 VQ 都会导致性能下降，这表明所提出的多粒度向量量化操作是必要的。移除 VQ_0 似乎对性能的影响更为显著，这说明离散化的上下文特征更有利于模型去噪。 VQ_1 在所有 VQ 模块中影响最小，而较高层的 VQ 对性能的影响大致更大。由于该数据集规模有限，这些结果可能无法全面反映分析结论，因此未来我们将在规模更大的数据集上评估所提出的方法。

最后，我们使用 t-SNE 对学习得到的 128 维码本进行可视化，将其降至二维，结果如图 2 所示。从上到下依次展示了包含 320 个（VQ_1）、640 个（VQ_2）和 960 个（VQ_3）可学习向量的单个码本，以及包含 320 个可学习向量的两个码本（总计 320×320 个向量，即 VQ_0）。显然，VQ 码本中使用的可学习向量越多，就能提取越细粒度的语音嵌入。从 t-SNE 分布来看， VQ_0提取的语音嵌入呈现出更明显的聚类特性，这有利于分类。另一方面，多个嵌入在空间上并不重叠，从而形成所谓的多粒度语音特征；它们在表示语音域时具有一定的互补性。这正是表 3 中每个 VQ 模块的引入都能对性能产生积极贡献的原因。

5. 结论

本文研究了多粒度向量量化和预训练模型在语音增强中的应用。结果表明，在不同层使用不同的向量量化模块，相较于使用单一固定向量量化，

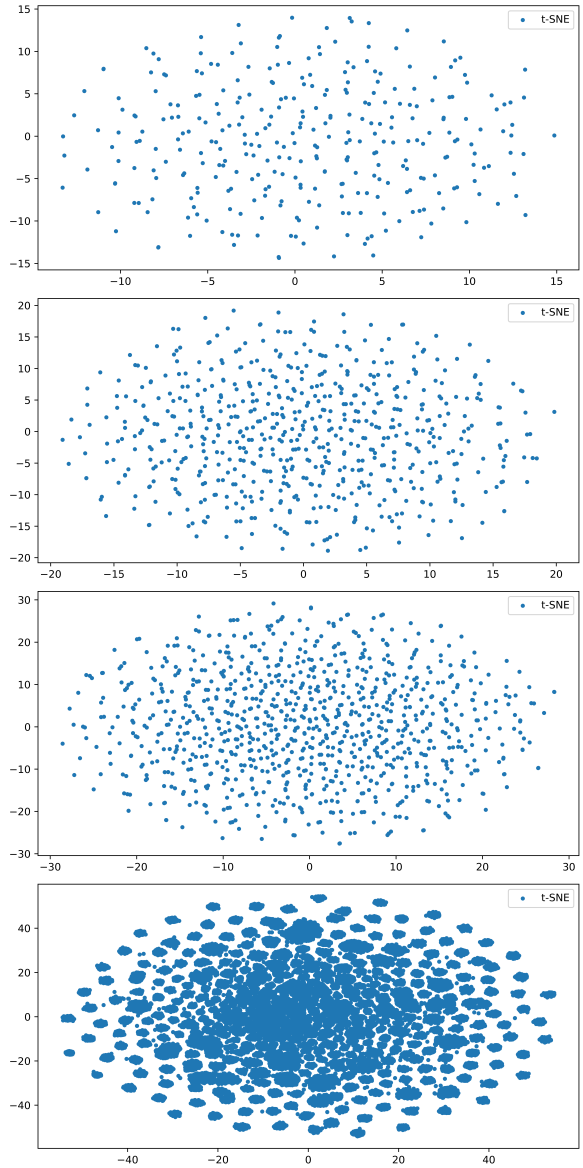

图2： 不同可学习向量数量下的码本可视化，其中从上到下依次展示了包含 *320*、*640* 和 *960* 个可学习向量的单一码本，以及包含 *320* 个可学习向量的两个码本。

能够提升语音增强性能。融合预训练模型提取的上下文特征与编码器提取的局部特征，有助于模型更好地建模全局和局部信息。与专用于语音表征学习的自监督预训练模型 wav2vec2.0 和 HuBERT 相比，data2vec 似乎更适用于语音增强，因为它能够从通用数据（例如语音、图像和视频）中学习表示。

6. 致谢

本研究得到中国国家自然科学基金（62101523）、合肥市自然科学基金（2022012）和中央高校基本科研业务费专项资金的资助。

7. 参考文献

[1] Z. Wang 和 D. Wang，《一种面向鲁棒自动语音识别的联合训练框架》，*IEEE/ACM* 音频、语音与语言处理汇刊，第24卷，第4期，第796–806页，2016年。[2] Q. Zhu、J. Zhang、Z. Zhang 和 L. Dai，《语音增强与自监督模型的联合训练，实现抗噪自动语音识别》，*arXiv* 预印本 *arXiv:2205.13293*，2022年。[3] J. Zhang 和 G. Zhang，《一种基于参数化无约束波束形成器的辅助听觉双耳降噪方法》，*IEEE/ACM* 音频、语音与语言处理汇刊，第30卷，304，2022 292页。–。[4] H. Taherian、Z. Wang、J. Chang 和 D. Wang，《基于单通道和多通道语音增强的鲁棒说话人识别》，*IEEE/ACM* 音频、语音与语言处理汇刊，第28卷，第1293–1302页，2020年。[5] S. Pascual、A. Bonafonte 和 J. Serr`a，《SEGAN：语音增强生成对抗网络》，载于*ISCA Interspeech*，2017年，第3642–3646页。[6] D. Michelsanti 和 Z. Tan，《用于语音增强和抗噪说话人验证的条件生成对抗网络》，载于*ISCA Interspeech*，2017年，第2008–2012页。[7] A. D´efossez、G. Synnaeve 和 Y. Adi，《波形域实时语音增强》，载于*ISCA Interspeech*，2020年，第3291–3295页。

[8] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, “语音 de - 在波形域中使用自注意力进行去噪，”发表于*ICASSP*，2022年，第7867–7871页。

[9] N. Shah、H. Patil 和 M. Soni，“使用卷积生成对抗网络的基于时频掩码的语音增强”，载于*APSIPA ASC* 论文集，2018年，第1246–1251 页。[10] Y. Hu、Y. Liu、S. Lv、M. Xing、S. Zhang、Y. Fu、J. Wu、B. Zhang 和 L. Xie，“DCCRN：用于相位感知语音增强的深度复数卷积循环网络”，载于 *ISCA Inter- speech*，2020年，第2472–2476 页。[11]Y. Luo 和 N. Mesgarani，“Conv-tasnet：超越用于语音分离的理想时频幅度掩蔽”，*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。[12] A. Razavi、A. Van den Oord 和 O. Vinyals，“利用 vq-vae-2 生成多样化的高保真图像”，*NIPS* 论文集，第 32 卷，2019年。[13] S. Gu、D. Chen、J. Bao、F. Wen、B. Zhang、D. Chen、L. Yuan 和 B. Guo，“用于文生图合成的向量量化扩散模型”，载于 *CVPR* 论文集，2022，第 10 696–10 706 页。[14] C. Du、Y. Guo、X. Chen 和 K. Yu，“Vqtts：基于自监督 VQ 声学特征的高保真文本到语音合成”，*arXiv* 预印本 *arXiv:2204.00768*，2022。[15] Y. Liu、R. Xue、L. He、X. Tan 和 S. Zhao，“Delightfultts 2：基于对抗性向量量化自编码器的端到端语音合成”，*arXiv* 预印本 *arXiv:2207.04646*，2022。[16] X. Jiang、X. Peng、H. Xue、Y. Zhang 和 Y. Lu，“用于可扩展神经语音编码的跨尺度矢量量化”，*arXiv* 预印本 *arXiv:2207.03067*，2022。[17] X. Zhao、Q. Zhu 和 J. Zhang，“基于自监督预训练模型和向量量化的语音增强”，*arXiv* 预印本 *arXiv:2209.14150*，2022。[18] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，“wav2vec 2.0：用于语音表征自监督学习的框架”，*NIPS* 论文集，第 33 卷，第 12 449–12 460 页，2020年。[19] Q.-s. Zhu、J. Zhang、M.-h. Wu、X. Fang 和 L.-R. Dai，“一种改进的 wav2vec 2.0 预训练方法：通过增强局部依赖建模实现语音识别。”载于 *ISCA Inter- speech*，2021年，第 4334–4338 页。[20] W. Hsu、B. Bolte、Y. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过对隐藏单元进行掩码预测实现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年。

[21] A. Baevski、W. Hsu、Q. Xu、A. Babu、J. Gu 和 M. Auli，“Data2vec：用于语音、视觉和语言自监督学习的通用框架”，*arXiv* 预印本 *arXiv:2202 03555*，2022。。[22] Q.-S. Zhu、J. Zhang、Z.-Q. Zhang、M.-H. Wu、X. Fang 和 L.- R. Dai，“基于抗噪自监督预训练模型的语音表征学习，用于自动语音识别”，载于 *ICASSP*，2022，第 3174–3178 页。[23] Q.-S. Zhu、L. Zhou、J. Zhang、S.-J. Liu、Y.-C. Hu 和 L.-R. Dai，“鲁棒 Data2vec：结合回归与改进的对比学习，实现面向自动语音识别的抗噪语音表征学习”，*arXiv* 预印本 *arXiv:2210.15324*，2022。[24] Z. Huang、S. Watanabe、S. 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，第 6837–6841 页。[25] K. Hung、S. Fu、H. Tseng、H. Chiang、Y. Tsao 和 C. Lin，“提升用于语音增强的自监督嵌入”，*arXiv* 预印本 *arXiv:2204.03339*，2022。

[26] E. Jang, S. Gu, 和 B. Poole，“使用 Gumbel-Softmax 进行类别重参数化”，*arXiv* 预印本 *arXiv:1611.01144*，2016年。

[27] S. Pascual, A. Bonafonte, and J. Serr`a，“Segan：语音增强 - 生成对抗网络”，载于*ISCAInterspeech*，2017年，第3642–3646页。

[28] C. M. Wendy, 等，“使用 wave-u-net 的改进语音增强”，*arXiv* 预印本 *arXiv:1811.11307*，2018年。

[29] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, 和 A. Mertins，“改进用于语音增强的生成对抗网络”，*IEEE* 信号处理快报，第 27 卷，第 1700–1704 页，2020年。

[30] M. Soni, N. Shah 和 H. Patil，“时间-频率掩蔽 - 基于生成对抗网络的语音增强，”载于*ICASSP*，2018年，第5039–5043页。

[31] S. Fu, C. Liao, Y. Tsao, 和S. Lin，“MetricGAN：生成对 - 抗网络驱动的黑盒度量分数优化，用于语音增强”，载于机器学习国际会议（*ICML*），第 97 卷。PMLR，2019年，第 2031–2041 页。

[32] Q. Zhang、A. Nicolson、M. Wang、K. Paliwal 和 C. Wang，“Deepmmse：一种基于深度学习的 MMSE 噪声功率谱密度估计方法”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 28 卷，第 1404–1415 页，2020年。[33] C.Valentini-Botinhao，“用于训练语音增强算法和文本转语音模型的噪声语音数据库”，2017年。[34] I.-T. Recommendation，“语音质量感知评估（感知语音质量评估）：一种用于窄带电话网络和语音编解码器端到端语音质量评估的客观方法”，*Rec. ITU-T P. 862*，2001年。[35] C. Taal、R. Hendriks、R. Heusdens 和 J. Jensen，“时频加权噪声语音可懂度预测算法”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 19 卷，第 7 期，第 2125–2136 页，2011年。[36] Y. Hu 和 P. Loizou，“语音增强客观质量度量的评估”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 16 卷，第 1 期，第 229–238 页，2008年。