arXiv:2602.01793v1 [cs.SD] 2 Feb 2026

ParaGSE：基于组矢量量化神经语音编解码器的并行生成式语音增强

Fei Liu，*Yang Ai**

中国科学技术大学语音与语言信息处理国家工程研究中心，中国合肥

fliu215@mail.ustc.edu.cn，yangai@ustc.edu.cn

摘要

近年来，生成式语音增强受到广泛关注；然而，现有方法受限于过高的复杂度、有限的效率以及不够理想的语音质量。为克服这些挑战，本文提出一种新颖的并行生成式语音增强（ParaGSE）框架，该框架采用基于组矢量量化（GVQ）的神经语音编解码器。基于 GVQ 的编解码器采用独立的矢量量化器，生成彼此独立的标记，从而能够在 ParaGSE 中高效地进行并行令牌预测。具体而言，ParaGSE 利用基于 GVQ 的编解码器将退化语音编码为不同的标记，通过以退化语音的频谱特征为条件的并行分支预测相应的干净令牌，并最终通过编解码器解码器重建干净语音。实验结果表明，在噪声、混响、带宽限制及其混合等多种失真条件下，与判别式和生成式基线相比，ParaGSE 始终能生成更优的增强语音。此外，得益于令牌预测中的并行计算，与串行生成式语音增强方法相比，ParaGSE 在 CPU 上的生成效率提高了约 1.5 倍。

索引词— 生成式语音增强、神经语音编解码器、并行令牌预测、分组矢量量化

1. 引言

在真实场景中，语音经常受到噪声、混响和带宽限制等失真的影响 [1]。语音增强（SE）通过优化退化语音、减少干扰并恢复清晰度来应对这些挑战，最终改善听觉体验以及通信 [2, 3] 和语音识别 [4, 5]等应用中的性能。

随着深度学习的兴起，基于神经网络的语音增强已成为主流方法，在提升语音质量方面超越了传统方法。早期的神经语音增强模型（如 DE-MUCS） [6], 通过回归损失直接最小化预测语音与原始干净语音之间的差异，但性能并不理想。随后，人们提出了判别式语音增强方法，采用生成对抗网络（GAN）来提升语音质量。时域判别式语音增强方法（如 SEGAN） [7], 使用全卷积网络，根据含噪输入预测干净语音，并通过对抗损失进行训练，但仍存在性能有限、计算成本高的问题。为应对这些挑战，频域判别式语音增强方法已逐渐受到青睐。例如，CMGAN [8] 通过预测复数频谱来重建干净语音波形，而 MP-SENet[9, 10] 则通过并行建模显式预测幅度谱和相位谱，从而提升性能。

近年来，神经语音编码的快速发展推动了生成式语音增强方法的发展，旨在利用由神经语音编解码器量化得到的离散令牌来提升语音质量 [11, 12, 13, 14, 15]。在生成式语音增强方法中，典型做法是使用神经网络从退化输入预测干净语音的离散令牌，再通过编解码器解码器重建干净语音。优化目标由此转变为令牌级分类损失，而非回归损失或对抗损失。然而，这类方法仍处于起步阶段，在效率、复杂度和性能方面面临诸多挑战。例如，GenSE [15] 使用大语言模型（LLM）进行干净语音令牌预测以解码干净语音，这无疑会导致较高的模型复杂度和较低的生成效率。Genhancer[14], 虽不依赖大语言模型，但采用基于残差矢量量化器（RVQ）的编解码器，从而产生自回归干净语音令牌预测，进而限制了计算效率。此外，GenSE和Genhancer都需要来自预训练模型的语义令牌输入，这也增加了其复杂度。

因此，本文提出了 ParaGSE，这是一种由基于分组向量量化器（GVQ）的神经语音编解码器支持的新型并行生成式语音增强方法，无需 LLM 或语义标记辅助即可运行，从而实现高效的并行计算。本文的主要贡献有两点。首先，我们介绍了 G-MDCTCodec，这是一种基于 GVQ 的新型编解码器，源自我们此前提出的 MDCTCodec [16], ，可通过分组和独立量化支持 ParaGSE 的并行标记预测与训练。另一项贡献是为生成式语音增强领域开发了 ParaGSE 中新颖的并行令牌预测框架。在该框架中，G-MDCTCodec 的编码器从退化语音中提取独立令牌，随后由轻量级并行预测分支处理并输出干净令牌。最后，通过 G-MDCTCodec 的解码器重建干净语音。客观和主观实验均表明，我们提出的 ParaGSE 能够在各种失真类型（如噪声、混响、带宽限制及其混合）下恢复出更高质量的语音，优于生成式和判别式语音增强基线。此外，得益于并行计算，ParaGSE 具有显著的效率优势，在 CPU 上的运行速度比串行生成式语音增强快 1.5× 倍以上。

*通讯作者。本研究由中国国家自然科学基金（项目编号62301521）资助。

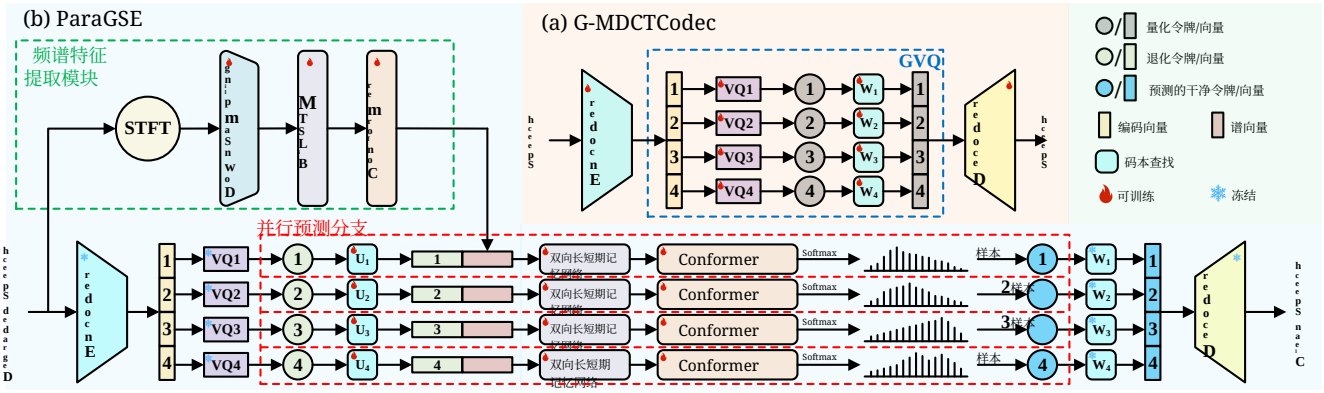


图1。（a）G-MDCTCodec 和（b）ParaGSE 的细节，以 $N = 4$ 为例。

2. 所提方法

所提出的 ParaGSE 利用 G-MDCTCodec 实现生成式语音增强中的并行令牌预测。图1展示了这两个组件的结构，下面将进行详细介绍。

2.1. G-MDCTCodec：基于GVQ的神经语音编解码器

如图1 (a)所示，G-MDCTCodec主要由编码器、GVQ和解码器组成。G-MDCTCodec决定了ParaGSE的性能上限，其所有组件均已集成并参与训练目标的构建。下文将详细介绍G-MDCTCodec的各个组件。

• **编码器与解码器：**我们保留MDCTCodec [16], 的编码器和解码器架构， [17]以一维ConvNeXt v2模块为主干网络，对输入语音的修正离散余弦变换（MDCT）频谱进行编码和解码。

• **GVQ：**GVQ沿维度轴对编码特征进行分组，并使用独立的VQ量化每个分组。与大多数编解码器中使用的RVQ相比 [16, 18, 19, 20], ，GVQ采用并行而非串行结构。此外，与HiFi-Codec使用的分组RVQ（GRVQ）串并行混合结构相比 [21], ，GVQ效率更高，也更适合生成式语音增强。假设编码器输出的单帧编码向量为 $\boldsymbol{e} \in \mathbb{R}^K$，则首先沿维度轴将其划分为 N组，得到 N个特征向量$\boldsymbol{e}_n \in \mathbb{R}^{K/N}$、 $n = 1$、.....、 N，其中 K表示向量维度。每个分组向量 $\boldsymbol{e}_n$ 都由对应的VQ进行量化：从关联码本 $\mathbb{W}_n = \{\boldsymbol{w}_{(m}_n \in \mathbb{R}^{K/N} \mid m = 1)$、.....、 $M\}$ 中选取与 $\boldsymbol{e}_n$ 欧氏距离最小的码向量作为量化向量，并将其索引视为标记，即

$$\hat{\boldsymbol{e}}_n, d_n = \arg \min_{\boldsymbol{w}_n^{(m)}, m} \|\boldsymbol{e}_n - \boldsymbol{w}_n^{(m)}\|_2, \tag{1}$$

其中$\hat{\boldsymbol{e}}_n \in \mathbb{R}^{K/N}$和 $d_n \in \{1, \ldots, M\}$ 分别表示量化向量和标记， M为码本大小。最后，将所有组的量化向量沿维度轴拼接，得到最终量化结果$\hat{\boldsymbol{e}} \in \mathbb{R}^K$，即

$$\hat{\boldsymbol{e}} = \left[\hat{\boldsymbol{e}}_1^\top, \hat{\boldsymbol{e}}_2^\top, \ldots, \hat{\boldsymbol{e}}_N^\top\right]^\top, \tag{2}$$

随后将其解码为语音。

•**训练准则：**在G-MDCTCodec的训练过程中，我们保留了MDCTCodec [16]中的对抗损失、MDCT频谱损失和梅尔频谱图损失。此外还引入了GVQ损失，以优化可学习码本并降低量化误差，其定义为各矢量量化输入与输出之间均方误差（MSE）之和，即，

$$\mathcal{L}_{\text{GVQ}} = \sum_{n=1}^{N} \mathbb{E}_{(\hat{\boldsymbol{e}}_n, \boldsymbol{e}_n)} \|\hat{\boldsymbol{e}}_n - \boldsymbol{e}_n\|_F^2, \tag{3}$$

其中 $\|\cdot\|_F$为弗罗贝尼乌斯范数。

2.2. ParaGSE：并行生成式语音增强

如图1（b）所示，ParaGSE由频谱特征提取模块 $\phi^{(SFEM)}$、 N 个并行预测分支$\phi_1^{(PPB)}, \ldots, \phi_N^{(PPB)}$ 以及G-MDCTCodec的各组件构成。ParaGSE首先将退化语音输入G-MDCTCodec的编码器和GVQ，以获得退化令牌。在频谱特征提取模块所提取频谱特征的条件下，并行预测分支将这些退化令牌并行转换为干净令牌，再由G-MDCTCodec的解码器解码，以重建干净语音。各组件的细节如下。

• **频谱特征提取模块：**以单帧特征为例，频谱特征提取模块 $\phi($ $_{SFEM)}$从退化语音片段 $\boldsymbol{y} \in \mathbb{R}^T$导出频谱特征向量$\hat{}\boldsymbol{s} \in \mathbb{R}^C$，即，

$$\hat{\boldsymbol{s}} = \phi^{(SFEM)}(\boldsymbol{y}), \tag{4}$$

其中 C 表示频谱特征维度。 T表示语音片段长度，它也对应于该模块和 G-MDCTCodec 编码器的下采样倍数。具体而言，首先对$\boldsymbol{y}$ 进行短时傅里叶变换（STFT），以获得幅度谱和相位谱，其中 STFT 的帧移为 w_s。将这两个频谱拼接后，输入由多个卷积层组成的下采样模块，各层步幅为 > 1，将时间分辨率降低 R倍。因此，我们有 $T = w_s \cdot R$。下采样后，应用两个双向长短期记忆网络（BiLSTM）层来捕获全局依赖关系，随后通过 Conformer模块 [22] 进一步建模全局和局部信息，从而得到频谱特征$\hat{}\boldsymbol{s}$。

• **并行预测分支：**以第 n个分支 $\phi_{(PPB)n}$为例，该分支以 G-MDCTCodec 的 GVQ 中第 n个 VQ 生成的标记 $d_{(y)n}$ （由 $\boldsymbol{y}$ 生成）以及$\hat{}\boldsymbol{s}$ 作为辅助输入，并输出干净标记$\hat{}d_{(x)n}$，即，

$$\hat{d}_n^{(x)} = \phi_n^{(PPB)}(d_n^{(y)}; \hat{\boldsymbol{s}}). \tag{5}$$

具体而言，首先通过查找可训练的嵌入表（即码本） $\mathbb{U}_n = \{\boldsymbol{u}_n^{(m)} \in \mathbb{R}^C \mid m = 1, \ldots, M\}$，将退化标记 $d_n^{(y)}$映射为潜向量 $\boldsymbol{v}_n \in \mathbb{R}^C$。接着，将潜向量 $\boldsymbol{v}_n$ 与频谱特征向量$\hat{\boldsymbol{s}}$拼接，得到 $\hat{}\boldsymbol{o}_n \in \mathbb{R}^{2C}$，再由两个双向长短期记忆网络层和一个 Conformer模块进行处理，生成输出向量 $\hat{}\boldsymbol{z}_n \in \mathbb{R}^M$。然后，将 $\hat{}\boldsymbol{z}_n$ 输入 softmax 层，以计算概率分布 $\hat{}\boldsymbol{p}_n^{(x)} \in \mathbb{R}^M$，即:

$$\hat{\boldsymbol{p}}_n^{(x)} = softmax(\hat{\boldsymbol{z}}_n). \tag{6}$$

通过对分布 $\hat{}\boldsymbol{p}_n^{(x)} = [\hat{}p_{n\,1}^{(x)}, \ldots\ldots, \hat{}p_{n\,m}^{(x)}, \ldots\ldots, \hat{}p_{n\,M}^{(x)}]^\top$进行最大值采样（即 argmax），即可得到预测的干净标记 $\hat{}d_n^{(x)}$，即:

$$\hat{d}_n^{(x)} = arg\max_m(\hat{p}_{n,m}^{(x)}). \tag{7}$$

最后，预测的干净语音标记 $\hat{}d_n^{(x)}$通过 G-MDCTCodec 的 GVQ 中第 n 个 VQ 的码本 $\mathbb{W}_n$ 还原为向量$\hat{\boldsymbol{e}}_n^{(x)}$。随后，按照公式 2 将所有 N 个分支的向量拼接起来，并输入解码器以重建干净语音。

• **训练准则：**在 ParaGSE 中，G-MDCTCodec 的所有组件均被冻结，仅训练频谱特征提取模块和 N 个并行预测分支。具体而言，首先将目标干净语音 $\boldsymbol{x} \in \mathbb{R}^T$ 输入 G-MDCTCodec 进行编码和量化，以生成目标干净语音标记 $d_{(x)\,1}, \ldots, d_{(x)\,N}$。随后，将这些标记分别转换为独热目标概率分布 $\boldsymbol{p}_{(x)\,1}, \ldots, \boldsymbol{p}_{(x)\,N}$，并用其与预测分布 $\hat{}\boldsymbol{p}_{(x)\,1}, \ldots, \hat{}\boldsymbol{p}_{(x)\,N}$ 计算交叉熵损失，即:

$$\mathcal{L} = \sum_{n=1}^{N} CrossEntropy(\hat{\boldsymbol{p}}_n^{(x)}, \boldsymbol{p}_n^{(x)}). \tag{8}$$

3. 实验

3.1. 实验设置

在我们的实验中[1]，，我们使用 VoiceBank语料库 [23], 构建了一个干净语音数据集，其中包含来自 56 位说话人的 23,075 条话语用于训练，以及来自 2 位未见说话人的 824 条话语用于测试，采样率为 16 kHz。我们基于干净语音语料库构建了针对不同任务的退化数据集，纳入三种失真类型：噪声、混响和混合失真（噪声 + 混响 +带宽限制）。

• **降噪：** 对于降噪任务，我们通过将 DEMAND 数据集 [24] 中的噪声添加到干净语音中，构建了噪声数据集。训练时使用了 10 种噪声，信噪比（SNR）范围为 0 至 15 分贝，间隔为 5 分贝；测试时则使用了 5 种未见过的噪声类型，信噪比范围为 2.5 至 17.5 分贝，间隔为 5 分贝。

[1]代码和语音样本可在此处获取：https://anonymity225.github.io/ParaGSE/。

表 1。MDCT-Codec 与 G-MDCTCodec 的解码语音质量对比。

编解码器	LSD↓	短时客观可懂度↑	语音质量客观评估↑
MDCTCodec	0.76	**0.94**	**4.42**
G-MDCTCodec	**0.75**	**0.94**	4.39

• **去混响：** 对于去混响任务，我们使用 DNS 挑战赛 [25], 的房间脉冲响应（RIR）数据集构建了混响数据集，该数据集包含 248 个真实 RIR 和约 60,000 个模拟 RIR。我们通过随机选择 RIR 并将其与干净语音进行卷积来生成混响语音，并确保测试集中使用的 RIR 在训练期间未曾出现。

• **混合失真抑制：**对于混合失真抑制任务，首先将混响添加到干净语音中，随后添加噪声，这两个步骤均遵循上述流程。最后，将语音信号下采样至 8 kHz（即带宽缩减），以构建混合失真数据集。

ParaGSE 中的 G-MDCTCodec 使用了 4 个矢量量化器（即 $N = 4$），每个矢量量化器的码本大小为 256（即 $M = 256$），码向量维度为 8（即 $K = 32$）。其他所有配置均与 MDCTCodec [16] 相同。ParaGSE 中的频谱特征提取模块采用短时傅里叶变换，帧长为 320，帧移为 40（即 $w_s = 40$），快速傅里叶变换点数为 1024。后续的下采样模块由 3 个卷积层组成，每层的卷积核大小为 3、步幅为 2，整体下采样因子为 8（即 $R = 8$）。频谱特征提取模块和并行预测分支均采用 Conformer 模块，通道数为 512（即 $C = 512$），注意力头为 8。我们在单张 Nvidia A800 GPU 上使用 AdamW 优化器训练 ParaGSE，$\beta_1 = 0.9$, $\beta_2 = 0.95$, 权重衰减为 0.01，共训练 100 轮。初始学习率设为 0.0005，采用余弦退火策略进行衰减，并在前 10k 步使用预热训练调度器。

我们将 ParaGSE 与多个具有代表性的基线模型进行了比较，包括早期的时域回归语音增强模型 DEMUCS [6],、频域判别式语音增强模型 CMGAN [8] 和 MP-SENet [10], ，以及生成式语音增强模型 Genhancer [14]。为确保比较公平，我们使用其官方实现重新制作了 DEMUCS 、 CMGAN 和 MP-SENet ； 由于 Genhancer 没有官方代码，我们根据其论文中的描述重新实现了该模型。

3.2. 可行性研究

如第 2.1 节所述，G-MDCTCodec 决定了 ParaGSE 增强语音质量的上限。因此，ParaGSE 的可行性取决于 G-MDCTCodec 能否达到足够高的编码质量。本节将 G-MDCTCodec 与其前身基于RVQ的MDCTCodec [16]的编码结果进行了比较。我们采用了语音编码中常用的评估指标，包括对数谱距离（LSD）、短时客观可懂度（STOI）和虚拟语音质量客观评估（ViSQOL）。实验结果见表 1。可以看出，G-MDCTCodec 的解码语音质量与 MDCTCodec 相当，这表明尽管 GVQ 中的矢量量化器彼此独立、相关性弱于 RVQ，但这并不会导致语音质量下降。这是 ParaGSE 实现并行令牌预测的先决条件，也验证了并行生成式语音增强的可行性。

表 2. ParaGSE 与基线语音增强方法在三种语音增强任务上的客观实验结果。

语音增强任务	模型	LSD↓	NISQA↑	DNSMOS↑	UTMOS↑
降噪	**DEMUCS**	1.17	3.57	3.07	3.59
	CMGAN	**0.74**	**4.67**	3.23	**4.01**
	MP-SENet	0.79	4.66	3.23	4.00
	Genhancer	1.10	3.40	2.65	2.73
	ParaGSE	0.95	4.57	**3.26**	3.95
去混响	**DEMUCS**	1.78	1.67	2.68	1.39
	CMGAN	**0.72**	4.45	3.13	3.54
	MP-SENet	0.91	4.33	3.23	3.41
	Genhancer	1.09	2.87	1.90	1.70
	ParaGSE	0.84	**4.60**	**3.26**	**3.86**
混合失真抑制	**DEMUCS**	2.27	0.96	1.72	1.44
	CMGAN	1.41	3.34	3.00	2.28
	MP-SENet	1.62	3.40	3.10	2.27
	Genhancer	1.40	3.47	2.01	1.95
	ParaGSE	**1.10**	**4.64**	**3.30**	**3.80**

3.3. 与基线语音增强方法的比较

为将 ParaGSE 与基线语音增强方法进行比较，我们采用了客观和主观指标进行评估。客观评估方面，我们使用了对数谱距离（LSD）以及三种在语音增强中广泛使用的非侵入式指标，即 NISQA [26], 、DNSMOS [27], 和 UTMOS [28], ，以评估整体语音质量。主观评估方面，我们在亚马逊土耳其机器人上进行了 ABX偏好测试，将 ParaGSE 与两种表现最强的基线方法 CMGAN 和 MP-SENet 进行比较。每次测试中，我们从测试集中随机选取由待比较模型增强的 20 段话语，并由至少 30 位以英语为母语的听众进行评估。听众判断每对样本中哪个质量更好，或表示没有偏好（N/P）。我们报告了平均偏好得分，以及通过 t 检验得到的 p-值，以评估结果的显著性。

客观和主观实验结果分别如表 2 和表 3 所示。在降噪和去混响任务中，如表 2 所示，ParaGSE 在所有客观指标上均始终优于时域模型 DEMUCS 和生成式语音增强模型 Genhancer。与判别式模型 CMGAN 和 MP-SENet 相比，ParaGSE 在降噪任务的三项非侵入式客观指标上表现相当，在去混响任务中甚至表现更优。然而，在侵入式对数谱距离指标上，ParaGSE 的表现不及 CMGAN 和 MP-SENet，这可以归因于生成式模型旨在近似真实分布，而非直接重建干净语音。表 3 进一步展示了主观结果。在降噪任务中，ParaGSE 的表现与 CMGAN（$p = 0.625$）相当，且显著优于 MP-SENet（$p < 0.01$）；而在去混响任务中，其表现明显优于两种基线方法（$p < 0.01$）。这些结果表明，尽管生成式模型在对数谱距离等侵入式指标上的表现较弱，但它们能够更好地捕捉感知重要特征。这说明 ParaGSE 在降噪和去混响这两项经典任务中均具有竞争力。

在混合失真抑制任务中，ParaGSE 的优势更加显著，表 2 和表 3 中的客观与主观结果均显示其相比所有基线方法都有显著提升（$p < 0.01$）。值得注意的是，它在包括对数谱距离（LSD）在内的多项指标上甚至超越了表现强劲的基线方法 CMGAN 和 MP-SENet，而听众始终更偏好 ParaGSE 增强语音的感知质量。这些发现表明，ParaGSE 在处理复杂失真时具有很强的适应性和鲁棒性，凸显了其在实际应用中的潜力。

表 3。ParaGSE 与 CMGAN 和 MP-SENet 在三种语音增强任务中的主观 ABX偏好得分（%）。

语音增强任务	ParaGSE	CMGAN	MP-SENet	N/P	p
降噪	32.63	34.08	-	33.29	0.625
	66.84	-	17.76	15.40	$<$ **0.01**
去混响	**42.41**	32.04	-	25.55	$<$ **0.01**
	56.22	-	24.46	19.32	$<$ **0.01**
混合失真抑制	**77.63**	15.00	-	7.37	$<$ **0.01**
	82.97	-	10.00	7.03	$<$ **0.01**

表4。ParaGSE、Genhancer 和串行 GSE 在混合失真抑制任务中的实时因子分析结果。此处“$a\times$”表示 $a\times$ 实时。

模型	Type	UTMOS↑	RTF（GPU）↓	RTF（CPU）↓
Genhancer	串行	1.95	0.1160 (8.6×)	0.6535 (1.5×)
SerialGSE	串行	3.21	0.0308 (32.5×)	0.0696 (14.4×)
ParaGSE	并行	**3.80**	**0.0228 (43.9×)**	**0.0466 (21.5×)**

3.4. 效率分析实验结果

为验证 ParaGSE 并行预测策略的效率优势，我们将其与基于串行预测的生成式语音增强方法（例如 Genhancer）进行了比较。然而，Genhancer 采用 DAC [19], 并依赖额外的语义令牌输入，因此与 ParaGSE 的比较并不公平。为此，我们基于 ParaGSE 构建了一个串行基线方法 SerialGSE。与 ParaGSE 相比，SerialGSE 仅将 G-MDCTCodec 替换为 MDCTCodec [16] ，并采用基于RVQ的顺序预测方案。具体而言，每个矢量量化器的干净令牌都根据其噪声令牌以及所有前序矢量量化器的干净令牌进行预测，同时每个预测分支均保留与 ParaGSE 相同的结构。

我们使用实时因子（RTF）衡量生成效率，其定义为生成一秒语音所需的时间，并在 Nvidia A800 GPU 和 Intel(R) Xeon(R) Silver 4314 CPU 上进行评估。混合失真抑制任务的实验结果见表4。Genhancer 的效率最低，这归因于其串行预测方案、DAC 的高复杂度以及对语义令牌输入的依赖。其质量（以 UTMOS 衡量）也不理想，可能是因为它未能利用 RVQ 依赖关系，且最初是为降噪而设计的 [14], ，因此在处理复杂失真方面效果较差。与 SerialGSE 相比，ParaGSE 具有明显的效率优势；得益于并行预测和加速，其在 CPU 上的速度提升了 1.5 倍。以 UTMOS 衡量，ParaGSE 的语音质量也更高。这些结果表明，ParaGSE 的并行框架兼顾了生成效率和感知质量，凸显了其在实时和资源受限应用中的实用性。

4. 结论

本文介绍了 ParaGSE，这是一种基于基于GVQ的神经语音编解码器构建的新型并行生成式语音增强模型。该编解码器采用独立的矢量量化器生成相互独立的令牌，使 ParaGSE 能够对退化语音进行编码，基于频谱特征并行预测干净令牌，并通过解码器重建增强语音。大量实验表明，在各种失真情况下，ParaGSE 的表现始终优于判别式和生成式基线模型，同时在 CPU 上的效率较串行方法提高约 1.5 倍。未来，我们计划进一步改进 ParaGSE，并研究其在更广泛应用场景中的有效性，包括更具挑战性的真实场景条件。

5. 参考文献

[1] Jacob Benesty、Jingdong Chen 和 Yiteng Huang，麦克风阵列信号处理，第 1 卷，Springer Science & Business Media，2008年。[2] Lawrence R Rabiner，“语音处理对现代电信的影响”，*Speech Communication*，第 17 卷，第 217–226 页，1995年。[3] Abdulhussain E Mahdi 和 Dorel Picovici，“现代电信中的语音质量测量进展”，*Digital SignalProcessing*，第 19 卷，第 79–103 页，2009年。[4] Peidong Wang、Ke Tan 等，“通过与失真无关的声学建模弥合单声道语音增强与识别之间的差距”，*IEEE/ACM 音频、语音与语言处理汇刊*，第 28 卷，48，2019年 39页。– 。[5] Z Chen、S Watanabe、H Erdogan 和 J Hershey，“使用长短期记忆循环神经网络集成语音增强与识别”，载于*Interspeech* 会议论文集，2015年，第 1–7 页。[6] Alexandre Defossez、Gabriel Synnaeve 和 Yossi Adi，“波形域中的实时语音增强”，载于*Interspeech* 会议论文集，2020年，第 3291–3295 页。[7]Santiago Pascual、Antonio Bonafonte 和 Joan Serra，“SEGAN：语音增强生成对抗网络”，载于*Interspeech* 会议论文集，2017年，第 3642–3646 页。[8] Sherif Abdulatif、Ruizhe Cao 和 Bin Yang，“CMGAN：用于单声道语音增强的基于Conformer的度量GAN”，*IEEE/ACM 音频、语音与语言处理汇刊*，第 32 卷，第 2477–2493 页，2024年。[9] Ye-Xin Lu、Yang Ai 和 Zhen-Hua Ling，“MP-SENet：一种对幅度谱和相位谱进行并行去噪的语音增强模型”，载于*Interspeech* 会议论文集，2023年，第 3834–3838 页。[10] Ye-Xin Lu、Yang Ai 和 Zhen-Hua Ling，“并行显式估计幅度谱和相位谱，实现高质量语音增强”，*Neural Networks*，第 107562 页，2025年。[11] Xiaofei Wang、Manthan Thakker、Zhuo Chen、Naoyuki Kanda、Sefik Emre Eskimez、Sanyuan Chen、Min Tang、Shujie Liu、Jinyu Li 和 Takuya Yoshioka，“SpeechX：作为多功能语音Transformer的神经编解码器语言模型”，*IEEE/ACM 音频、语音与语言处理汇刊*，第 32 卷，第 3355–3364 页，2024年。[12] Huaying Xue、Xiulian Peng 和 Yan Lu，“通过语音令牌生成实现低延迟语音增强”，载于*ICASSP* 会议论文集，2024年，第 661–665 页。[13] Ziqian Wang、Xinfa Zhu、Zihan Zhang、YuanJun Lv、Ning Jiang、Guoqing Zhao 和 Lei Xie，“SELM：使用离散令牌和语言模型进行语音增强”，载于*ICASSP* 会议论文集，2024年，第 11561–11565 页。[14] Haici Yang、Jiaqi Su、Minje Kim 和 Zeyu Jin，“Genhancer：通过对离散编解码器令牌进行生成式建模实现高保真语音增强”，载于*Interspeech* 会议论文集，2024年，第 1170–1174 页。[15] Jixun Yao、Hexin Liu、Chen Chen、Yuchen Hu、Ensiong Chng 和 Lei Xie，“GenSE：通过分层建模，使用语言模型进行生成式语音增强”，载于*ICLR* 会议论文集，2025年，第 2025 卷，第 69229–69249 页。[16] Xiao-Hang Jiang、Yang Ai、Rui-Chen Zheng、Hui-Peng Du、Ye-Xin Lu 和 Zhen-Hua Ling，“MDCTCodec：一种面向高采样率和低比特率场景的轻量级基于 MDCT 的神经音频编解码器”，载于*SLT* 会议论文集，2024年，第 540–547 页。[17] Sanghyun Woo、Shoubhik Debnath、Ronghang Hu、Xinlei Chen、Zhuang Liu、In So Kweon 和 Saining Xie，“ConvNeXt v2：通过掩码自编码器对卷积网络进行协同设计与扩展”，载于*CVPR* 会议论文集，2023年，第 16133–16142 页。[18] Alexandre D´efossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi，“高保真神经音频压缩”，机器学习研究汇刊，2023年。[19] Rithesh Kumar、Prem Seetharaman、Alejandro Luebs、Ishaan Kumar 和 Kundan Kumar，“采用改进 RVQGAN 的高保真音频压缩”，载于*NeurIPS* 会议论文集，2023年，第 36 卷，第 27980–27993 页。[20] Yang Ai、Xiao-Hang Jiang、Ye-Xin Lu、Hui-Peng Du 和 Zhen-Hua Ling，“APCodec：一种采用并行幅度谱和相位谱编解码的神经音频编解码器”，*IEEE/ACM 音频、语音与语言处理汇刊*，第 32 卷，第 3256–3269 页，2024年。[21] Dongchao Yang、Songxiang Liu、Rongjie Huang、Jinchuan Tian、Chao Weng 和 Yuexian Zou，“HiFi-Codec：用于高保真音频编解码器的分组残差矢量量化”，*arXiv* 预印本 *arXiv:2305.02765*，2023年。[22] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Par-mar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zheng-dong Zhang、Yonghui Wu 等，“Conformer：卷积增强型语音识别 Transformer”，载于*Interspeech* 会议论文集，2020年，第 5036–5040 页。[23] Cassia Valentini-Botinhao、Xin Wang、Shinji Takaki 和 Ju-nichi Yamagishi，“探究基于 RNN 的语音增强方法在抗噪文本转语音中的应用”，载于*SSW* 会议论文集，2016年，第 146–152 页。[24]Joachim Thiemann、Nobutaka Ito 和 Emmanuel Vin-cent，“多样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，2013年，第 19 卷。[25] Harishchandra Dubey、Ashkan Aazami、Vishak Gopal、Babak Naderi、Sebastian Braun、Ross Cutler、Alex Ju、Mehdi Zo-hourian、Min Tang、Mehrsa Golestaneh 等，“ICASSP 2023 深度噪声抑制挑战赛”，*IEEE* 信号处理开放期刊，2024年。[26] Gabriel Mittag、Babak Naderi、Assmaa Chehadi 和 Sebas-tian M¨oller，“NISQA：一种利用众包数据集进行多维语音质量预测的深度 CNN 自注意力模型”，载于*Interspeech* 会议论文集，2021年，第 2127–2131 页。[27] Chandan KA Reddy、Vishak Gopal 和 Ross Cutler，“DNS-MOS P. 835：一种用于评估噪声抑制器的非侵入式感知客观语音质量指标”，载于*ICASSP* 会议论文集，2022年，第 886–890 页。[28] Takaaki Saeki、Detai Xin、Wataru Nakata、Tomoki Koriyama、Shinnosuke Takamichi 和 Hiroshi Saruwatari，“UTMOS：东京大学 Sarulab 语音 MOS 挑战赛 2022 系统”，载于*Interspeech* 会议论文集，2022年，第 4521–4525 页。