arXiv:2603.01476v1 [eess.AS] 2 Mar 2026

ENTROPY-GUI 面向超低比特率神经语音编解码器的 DED GRVQ

*Yanzhou Ren**、*Noboru Harada*†、*Daiki Takeuchi*†、*Siyu Chen**、*Wei Liu**、*Xiao Zhang**、*Liyuan Zhang**、*Takehiro Moriya*†和 *Shoji Makino**

* 日本早稻田大学

† 日本电信电话株式会社

摘要

神经音频编解码器（NAC）对于重建高质量语音信号以及为下游语音语言模型生成离散表示至关重要。然而，在超低比特率限制下，既要确保准确的语义建模，又要保持高保真重建，仍然颇具挑战。我们提出了一种用于超低比特率神经语音编解码器的熵引导分组残差向量量化（EG-GRVQ），该方法保留了用于语言信息的语义分支，并在声学分支中采用熵引导分组策略。假设各通道的激活近似服从高斯统计，则每个通道的方差可作为其信息量的合理代理指标。基于这一假设，我们对编码器输出进行划分，使每个组承载相同份额的总信息量。这种均衡分配提高了码本效率并减少了冗余。我们的模型在 LibriTTS 和 VCTK 上训练，在超低比特率条件下提升了感知质量和可懂度指标，并着重满足面向通信场景的编解码器级保真度要求。

索引词—— 神经音频编解码器、离散表示、语义信息、熵引导量化、码本效率。

1. 引言

神经音频编解码器（NAC）通过生成用于语音应用中不同目的的离散表示，在现代语音处理中日益重要 [1–6]。先前的研究将离散语音表示分为两大类：语义令牌和声学令牌 [7]。语义令牌捕捉高层语言信息，广泛用于下游语音语言建模任务 [8]。它们通常从 HuBERT 和 WavLM 等自监督预训练模型中提取 [9, 10]。随后，可通过语义感知量化将其纳入神经编解码器 [11–13], ，从而为语音转文本、语音翻译和零样本文本转语音等应用实现精准的内容对齐 [14–17]。声学令牌通常由神经音频编解码器生成，旨在保留细粒度的信号细节以重建波形。SoundStream [18], 、EnCodec [19],和 Stable Codec [20] 等神经编解码器依赖残差向量量化（RVQ）或有限尺度量化（FSQ）；而 HIFI-Codec [21] 进一步引入了分组残差向量量化（GRVQ），通过将信息分配到不同量化组来提升感知质量。值得注意的是，早在 35 多年前，超低比特率编解码器中的系数量化就是一个重要研究课题 [22], ；如今，在神经语音与音频编码领域，这一相似课题再次受到关注。

尽管语义令牌和声学令牌各有优势，但在超低比特率限制下兼顾二者仍是一项未解决的挑战。近期研究尝试在单一编解码器中融合这两个方面。例如，Mimi [13] 采用并行量化设计，设置专门的语义和声学信息分支。它使用 RVQ 对声学信息进行编码，同时也为语义令牌分配比特率容量，这不可避免地减少了用于声学细节的资源 [23], ，因此限制了可懂度与重建质量的同步提升，尤其是在超低比特率条件下。

本文采用 Mimi [13] 作为基线模型，因其具备适用于实时通信场景的因果低延迟架构，并特别着重于增强声学分支，以实现更高的重建保真度。我们提出了一种熵引导分组残差向量量化（EG-GRVQ）策略，该策略扩展了最初在 HiFi-Codec [21]中引入的分组残差向量量化框架。EG-GRVQ 的设计受到方差与熵之间信息论联系的启发，其中通道方差可作为信息量的替代指标。我们对编码器输出进行分组，使各组的累积方差趋于均衡，从而实现更均衡的信息分配并提高码本效率。在明确保留 Mimi 的语义分支的同时，我们的设计确保声学分支在低比特率条件下实现更高的保真度。在 LibriTTS [24] 和 VCTK [25] 上的实验表明，我们的方法在感知质量和可懂度方面均有所提升。

2. 提出的方法

2.1. 编码器和解码器

我们所提方法的编码器和解码器遵循 Mimi [13] 的整体架构。如图1所示，输入为 24 千赫兹波形，首先依次通过四个残差卷积块，最后经过编码器中的一维卷积。这些模块逐步降低时间分辨率并增加通道维度，最终将波形转换为 512 维潜在表示，速率为每秒 12.5 帧。卷积块之后，在瓶颈层应用 Transformer 块，以捕获长程依赖关系并增强潜在表示的紧凑性。解码器采用与编码器对称的架构，反向运行以重建波形。它以量化后的潜在特征为输入，通过转置卷积块重建波形；这些卷积块逐步提高时间分辨率并降低通道维度。

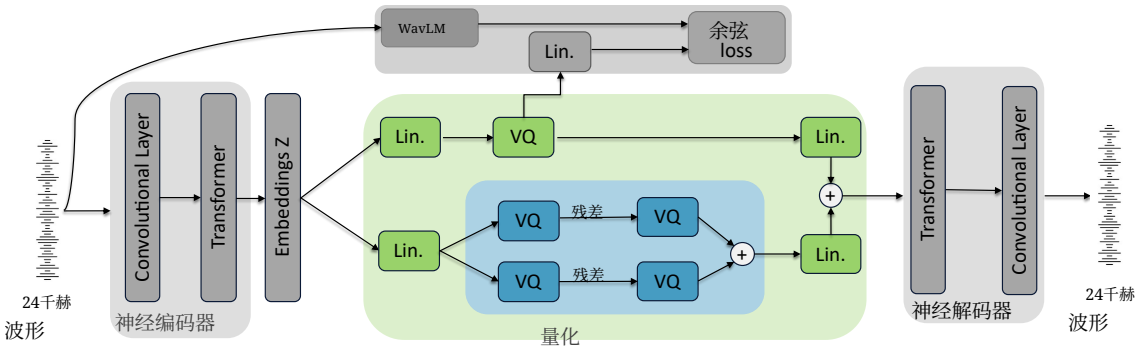


图1：所提模型的结构。

2.2. 量化

如图1所示，编码后，潜在特征被复制为两个分支。第一个分支由语义量化器处理，该量化器旨在捕获对语音可懂度至关重要的高层语言学特征。该语义层采用单一码本，以尽量降低比特率消耗，同时保留必要的语义信息。

第二个分支由熵引导分组残差矢量量化器处理。图2展示了针对声学信息的三种不同量化器结构配置：(a) 残差矢量量化，(b) 分组残差矢量量化，以及 (c) EG-GRVQ（所提方法）。残差矢量量化依次利用一系列残差码本对目标声学向量 Z_2 进行量化，而传统的分组残差矢量量化则将声学目标向量 Z_2 的通道划分为两个编码组。

与将通道均匀划分的传统分组残差矢量量化方法不同，我们的方法 EG-GRVQ 利用编码器输出的全局方差分布来指导分组。令 $z_{k,\,t}$ 表示通道 $k \in \{1,\ldots,C\}$ 在时间帧 $t \in \{1,\ldots,T\}$ 处的编码器激活值。训练集上通道 k 的方差定义为

$$\sigma_k^2 = \frac{1}{T}\sum_{t=1}^{T}\left(z_{k,t} - \mu_k\right)^2, \tag{1}$$

其中， μ_k 是通道 k的平均输出 .

从信息论视角来看，我们假设每个编码器通道减去其均值偏置常数后，近似服从高斯分布 [26]。因此，微分熵可表示为

$$H(X) = \tfrac{1}{2}\ln\left(2\pi e\sigma^2\right), \tag{2}$$

这表明熵是方差的单调函数。因此，通道方差可作为衡量每个通道所承载信息量的替代指标。

然后，我们确定满足以下条件的最小索引 k^*

$$\sum_{i=1}^{k^*}\sigma_i^2 \;\geq\; \tfrac{1}{2}\sum_{j=1}^{C}\sigma_j^2, \tag{3}$$

确保第1组占总方差的约一半。在我们的实现中，使用 $C = 512$ 时，分割点为 $k = 237$，由此形成包含 237 个通道的第1组（如图2-(c)所示的码本 1 和 3）和包含 275 个通道的第2组（如图 2-(c)所示的码本 2 和 4）。在 $k = 256$ 处进行均分后，前半部分分配到 55.30 % 的方差，后半部分分配到 44.70 %。这一策略缓解了高方差通道占主导地位的问题

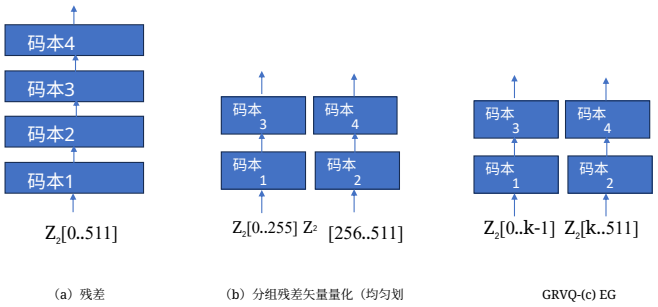


图2：量化器结构配置：(a) 残差矢量量化，(b) 分组残差矢量量化，以及(c) EG-GRVQ（所提方法）。

以及低方差通道利用不足的问题，从而提高量化效率和重建保真度。

请注意，分配了一些信令比特时，分割位置 k 可在每一帧中自适应调整。然而，考虑到额外比特开销与编码增益提升之间的权衡，本方法未采用自适应调整。由此得到的分组在训练和推理期间对所有数据均保持固定，并作为超参数。

最后，将语义分支和声学分支的量化特征相加，并传递给解码器以重建波形。

3. 实验

3.1. 基线

在实验中，我们选取四种系统作为对比基线。我们将未经重新训练的原始 Mimi 编解码器称为 Mimi（官方版），将采用残差矢量量化的重新训练版本称为 Mimi（重新训练版），并将我们在 Mimi 框架中重新实现的分组残差矢量量化称为 Mimi（GRVQ）。我们基于熵引导分组提出的方法称为 Proposal（EG-GRVQ）。

3.2. 数据集与实验设置

我们在合并后的 LibriTTS [24]和VCTK [25], 数据集上训练了所提出的方法，该数据集包括 LibriTTS 的 train-clean-100、train-clean-360 和 train-other-500 子集，以及完整的 VCTK 语料库。为确保比较公平、客观，我们重新训练了

表 1：0.6875 千比特每秒下的客观评估结果。

方法	矢量量化方案（结构）	SDR↑	感知语音质量评估↑	短时客观可懂度↑	ViSQOL↑
Mimi（官方版）	残差矢量量化（1x4）	**3.451**	1.872	0.876	2.010
Mimi（重新训练版）	残差矢量量化（1x4）	-6.969	1.779	0.886	**2.546**
Mimi（GRVQ）	分组残差矢量量化（2x2）	-7.294	1.852	0.889	2.464
所提方法	EG-GRVQ (2x2)	-7.309	**1.881**	**0.890**	2.496

表 2：各声学量化器的归一化均方误差（NMSE）。

方法	码本1	码本2	码本3	码本4	总计
RVQ	0.928	0.908	0.894	0.884	0.884
GRVQ	0.890	0.895	0.842	0.862	0.852
Proposal（EG-GRVQ）	**0.865**	**0.855**	**0.831**	**0.813**	**0.819**

Mimi 基线模型，并使其采用相同的数据条件。作为比较，我们还在 Mimi 框架中重新实现了先前提出的 GRVQ 量化方法 [21]，并将其记为 Mimi（GRVQ）。所有训练和评估均在相同条件下采用 5 个码本的设置进行。所有模型均使用 8 张 NVIDIA A6000 GPU（每张 48 GB）进行训练，每张 GPU 的批量大小为 12。

3.3. 训练策略

为训练所提出的模型，我们采用一种结合特征损失与对抗损失的多目标策略，并使用基于判别器的方案来提升感知质量。生成器通过最小化以下复合损失进行优化：

$$\mathcal{L}_{\text{gen}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{feat}}\mathcal{L}_{\text{FM}}, \tag{4}$$

其中：

- $\mathcal{L}_{\text{adv}}$ 表示对抗损失，计算判别器预测值与目标标签 1（真实）之间的均方误差（MSE）：

$$\mathcal{L}_{\text{adv}} = \frac{1}{N}\sum_{i=1}^{N}(D(\hat{x}_i) - 1)^2. \tag{5}$$

- $\mathcal{L}_{\text{FM}}$ 表示特征匹配损失，定义为真实特征与生成的判别器中间特征之间的 L_1 距离：

$$\mathcal{L}_{\text{FM}} = \frac{1}{N}\sum_{l=1}^{N}\left\|D^{(l)}(x) - D^{(l)}(\hat{x})\right\|_1, \tag{6}$$

其中，N 表示所有层和尺度上的特征图总数。

我们设置 $\lambda_{\text{adv}} = 1$ 和 $\lambda_{\text{feat}} = 15$。此外，还加入了权重为 $\lambda_{\text{commit}} = 1$的向量量化承诺损失 $\mathcal{L}_{\text{commit}}$，以稳定码本使用。

在语义学习方面，我们沿用了 Speechtokenizer [12] 的做法，使用投影后的语义量化器输出与 WavLM 提取的嵌入之间的蒸馏损失。语义量化器的潜在特征被投影到 1024 维空间，然后使用余弦相似度损失 L_{sem}与 WavLM 提取的语义嵌入进行比较。语义蒸馏损失 L_{sem} 计算如下：

$$\mathcal{L}_{\text{sem}} = 1 - \cos(z_s, z_{\text{WavLM}}). \tag{7}$$

3.4. 客观评估

在客观评估中，我们从 LibriTTS [24] test-clean子集中随机选取了 200 个样本，该子集严格排除在训练之外。为防止信息泄漏，我们确保训练集与评估集之间没有数据重叠。我们采用四种客观指标来评估重建质量：语音质量感知评估（感知语音质量评估）(PESQ) [27], 短时客观可懂度（STOI） [28], 虚拟语音质量客观听测器（ViSQOL） [29], 以及信号失真比（SDR） [30]。具体而言，我们采用 PESQ 的宽带版本评估感知质量，采用 STOI 评估低比特率条件下的可懂度，并采用 ViSQOL 衡量基于参考的感知相关性。尽管 SDR 广泛用于分离和增强任务，但它对相位失配较为敏感；因此，我们仅将其作为补充指标报告。评估时，所有模型输出最初均为 24 kHz、16 位分辨率，并使用 7.5 kHz 低通滤波器下采样至 16 kHz，以满足 PESQ 和 STOI 的输入要求。

如表1所示，我们提出的EG-GRVQ在感知语音质量评估和短时客观可懂度上取得了最佳性能，同时在ViSQOL上保持了具有竞争力的结果。与Mimi（官方版）相比，我们的方法将感知语音质量评估提高了0.01，ViSQOL提高了0.49，短时客观可懂度提高了0.01，表明感知质量有显著提升。相对于在相同数据上训练的Mimi（重新训练版），我们的方法使感知语音质量评估提高了0.1，短时客观可懂度提高了0.01，同时信号失真比相当（-7.31对-6.97）。与Mimi（GRVQ）相比，我们的方法进一步将感知语音质量评估从1.85提高到1.88，将ViSQOL从2.46提高到2.50。这些结果验证了所提出的熵引导设计在低比特率条件下提升感知质量的有效性。

此外，我们评估量化重建与 oracle 编码器输出之间的归一化均方误差（NMSE），该指标能够提供与尺度无关的失真度量。表 2 展示了不同声学量化器的 NMSE 结果。所提出的方法始终能实现更低的 NMSE，证实我们的分组策略相对于潜变量方差能够带来更准确的重建。

3.5. 码本效率与分组分析

我们比较声学分支中的码本利用率，以评估不同量化策略的有效性。如图 3 所示，残差矢量量化基线的码本利用率在较深层中急剧下降

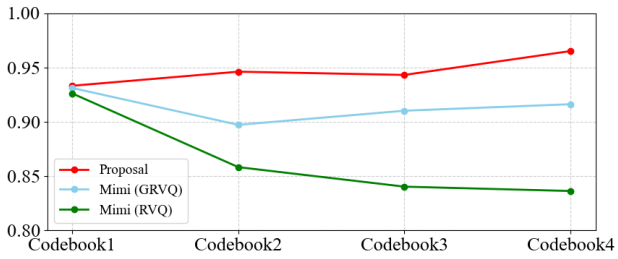


图 3：声学分支中的码本利用率。

表 3：0.6875 千比特每秒下不同分组数量的客观结果。

方法	SDR↑	感知语音质量评估↑	短时客观可懂度↑	ViSQOL↑
1×4 (残差矢量量化)	**-6.969**	1.779	0.886	**2.546**
4×1	-7.420	1.736	0.875	2.026
2×2 (分组残差矢量量化)	-7.294	**1.852**	**0.889**	2.464

，这表明码字使用效率低下，仍有改进空间。Mimi（GRVQ）基线通过平衡各组间的分配，在一定程度上缓解了这一问题，但仍存在逐层失衡。相比之下，所提出的 EG-GRVQ 在所有层中都保持了持续较高且稳定的码本利用率，表明其能更有效地利用量化器容量。这说明，我们的熵引导分组不仅保留了重建准确度，还能在比特率受限的情况下最大限度地提高表示效率。

为评估声学分支中分组数量的影响，我们在量化器总数固定为 4 的条件下比较了三种配置： $2groups \times 2quantizers$, $4groups \times 1quantizer$, $1group \times 4quantizers$ 3， 2×2 。如表所示，该配置在所有指标上均持续优于 4×1 变体。双组结构使每组能够进行更深层次的量化，从而提升重建质量，尤其是感知相似度。这些结果表明，在比特率固定的情况下，与对更多组进行浅层量化相比，对较少的组进行更深层次的量化更为有效。与 1×4 配置相比， 2×2 变体的 PESQ 和 STOI 显著更高，ViSQOL 相同，但信号失真比略低。这表明，在单组中简单地堆叠量化器可能导致信息分配效率低下，进而降低感知质量。这些结果表明，在比特率固定的情况下，对较少的组进行更深层次的量化，比浅层多组量化和缺乏结构分离的单组量化都更有效。

3.6. 主观评估

除客观指标外，我们还进行了带隐藏参考和锚点的多重刺激（MUSHRA）测试，从人类听众的角度评估感知质量。MUSHRA 广泛应用于音频编码和语音增强研究。该测试要求参与者根据隐藏参考和明确设置的劣化锚点，对多个信号变体的质量进行评分，从而提供可靠的主观评估。

我们从 LibriTTS test-clean子集中随机选取了八段语音样本，其中男性和女性各四段，时长从 1.4 秒到 13 秒不等。这些语音样本录制

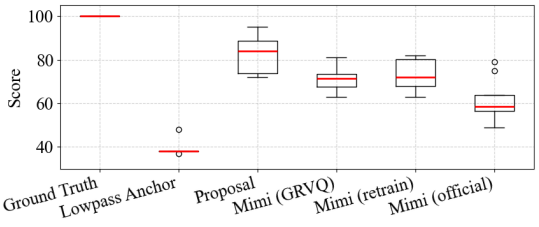


图 4：不同方法的 MUSHRA 得分分布。

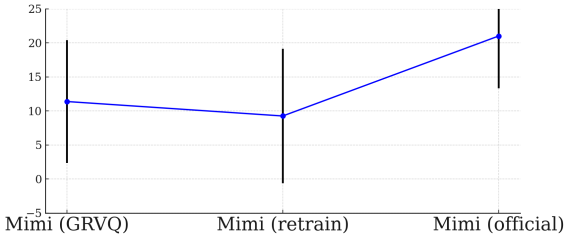


图 5：所提方法（EG-GRVQ）与基线方法之间的 MUSHRA 平均得分差异及 95% 置信区间

时的采样率为 24 kHz，分辨率为 16 位，用于评估。测试系统包括干净的理想参考音频、通过对参考音频应用 3.4 kHz 低通滤波器生成的低通基准音频、Mimi（GRVQ）、Proposal（EG-GRVQ）、Mimi（重新训练版）和 Mimi（官方版）。听音测试由八名参与者完成（五男三女），他们均有音频感知评估经验。

如图 4 所示，所提方法的 MUSHRA 得分比官方 Mimi 高 21 分，比采用 GRVQ 的 Mimi 高 11 分，表明听众明显更偏好所提方法。图 5 进一步展示了 MUSHRA 平均得分差异及 95% 置信区间。结果表明，所提方法相较于官方基线和 Mimi（GRVQ）均取得了具有统计显著性的提升，因为置信区间范围未与零重叠。这些发现凸显了所提方法在低比特率条件下具有明显的感知偏好优势。

4. 结论

本文提出了一种熵引导分组残差向量量化框架，在超低比特率神经语音编码方面展现出明显优势。通过将通道方差作为熵的代理指标，我们的方法能够确保信息在各量化组之间分配得更加均衡，从而同时提升重建保真度和码本利用率。客观评估结果显示，感知语音质量评估（PESQ）和短时客观可懂度（STOI）均持续提升，同时 ViSQOL 和信号失真比（SDR）得分保持竞争力；归一化均方误差（NMSE）分析进一步证实，相较于潜在方差，该方法能够实现更准确的重建。主观 MUSHRA 测试也表明，相较于基线系统，受试者显著偏好本方法。此外，码本使用分析显示，所提出的设计提高了码本利用率，使其在比特率受限的条件下能够实现更高效的表示。这些结果凸显了 EG-GRVQ 在超低比特率神经语音编解码器中的应用潜力。

5. 参考文献

[1] A. Van Den Oord、O. Vinyals 等，“神经离散表示学习”，《神经信息处理系统进展》，第 30 卷，2017 年。[2] H. Yang、K. Zhen、S. Beack 和 M. Kim，“面向噪声语音压缩的源感知神经语音编码”，载于 IEEE 国际声学、语音与信号处理会议论文集，2021 年，第 706–710 页。[3] Y. Yang、F. Shen、C. Du、Z. Ma、K. Yu、D. Povey 和 X. Chen，“迈向通用语音离散令牌：自动语音识别与文本转语音的案例研究”，载于 IEEE 国际声学、语音与信号处理会议论文集，2024 年，第 10 401–10 405 页。[4] S. Chen、C. Wang、Y. Wu、Z. Zhang、L. Zhou、S. Liu、Z. Chen、Y. Liu、H. Wang、J. Li 等，“神经编解码器语言模型是零样本文本转语音合成器”，《IEEE 音频、语音与语言处理汇刊》，2025 年。[5] Y.-C. Wu、I. D. Gebru、D. Markovi´c 和 A. Richard，“Audiodec：开源流式高保真神经音频编解码器”，载于 IEEE 国际声学、语音与信号处理会议论文集，2023 年，第 1–5 页。[6] Z. Borsos、M. Sharifi、D. Vincent、E. Kharitonov、N. Zeghi-zour 和 M. Tagliasacchi，“SoundStorm：高效并行音频生成”，arXiv 预印本 arXiv:2305.09636，2023 年。[7] Z. Borsos、R. Marinier、D. Vincent、E. Kharitonov、O. Pietquin、M. Sharifi、D. Roblek、O. Teboul、D. Grangier、M. Tagliasac-chi 等，“AudioLM：一种音频生成的语言建模方法”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 31 卷，第 2523–2533 页，2023 年。[8] T. Brychc´ın 和 M. Konop´ík，“用于改进语言建模的语义空间”，《计算机语音与语言》，第 28 卷，第 1 期，第 192–209 页，2014 年。[9] A. Baevski、Y. Zhou、A. Mohamed 和 A. Auli，“wav2vec 2.0：语音表示自监督学习框架”，《神经信息处理系统进展》，第 33 卷，第 12 449–12 460 页，2020 年。[10] 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 信号处理专题期刊》，第 16 卷，第 6 期，第 1505–1518 页，2022 年。[11] H. Liu、X. Xu、Y. Yuan、M. Wu、W. Wang 和 M. D. Plumbley，“SemantiCodec：面向通用声音的超低比特率语义音频编解码器”，《IEEE 信号处理专题期刊》，2024 年。[12] X. Zhang、D. Zhang、S. Li、Y. Zhou 和 X. Qiu，“SpeechTokenizer：面向语音大语言模型的统一语音分词器”，载于国际学习表征会议论文集，2024 年。[13] A. D´efossez、L. Mazar´e、M. Orsini、A. Royer、P. P´erez、H. J´egou、E. Grave 和 N. Zeghidour，“Moshi：用于实时对话的语音文本基础模型”，arXiv 预印本 arXiv:2410.00037，2024 年。[14] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhut-dinov 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示学习”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 29 卷，第 3451–3460 页，2021 年。[15] X. Zhu、Y. Lv、Y. Lei、T. Li、W. He、H. Zhou、H. Lu 和 L. Xie，“Vec-Tok speech：用于神经语音生成的语音向量化与分词”，《IEEE 音频、语音与语言处理汇刊》，2025 年。[16] Y. Wang、H. Zhan、L. Liu、R. Zeng、H. Guo、J. Zheng、Q. Zhang、X. Zhang、S. Zhang 和 Z. Wu，“MaskGCT：采用掩码生成式编解码器变换器的零样本文本转语音”，载于国际学习表征会议论文集，2025 年。[17] D. Zhang、S. Li、X. Zhang、J. Zhan、P. Wang、Y. Zhou 和 X. Qiu，“SpeechGPT：赋予大型语言模型内在的跨模态对话能力”，载于自然语言处理实证方法会议论文集，2023 年。[18] N. Zeghidour、A. Luebs、A. Omran、J. Skoglund 和 M. Tagliasacchi，“SoundStream：端到端神经音频编解码器”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 30 卷，495–507 页，2021 年。[19] A. D´efossez、J. Copet、G. Synnaeve 和 Y. Adi，“高保真神经音频压缩”，*arXiv* 预印本 *arXiv:2210 13438*，2022。。[20] J. D. Parker、A. Smirnov、J. Pons、C. Carr、Z. Zukowski、Z. Evans 和 X. Liu，“面向低比特率高质量语音编码的变换器扩展”，载于国际学习表征会议论文集，2025 年。[21] D. Yang、S. Liu、R. Huang、J. Tian、C. Weng 和 Y. Zou，“Hifi-codec：用于高保真音频编解码器的组残差矢量量化”，*arXiv* 预印本 *arXiv:2305.02765*，2023 年。[22] Y. Shiraki 和 M. Honda，“基于可变长度分段量化的 LPC 语音编码”，*IEEE* 声学、语音与信号处理汇刊，第 36 卷，第 9 期，1437–1444 页，1988 年。[23] S. Ji、Z. Jiang、W. Wang、Y. Chen、M. Fang、J. Zuo、Q. Yang、X. Cheng、Z. Wang、R. Li、Z. Zhang、X. Yang、R. Huang、Y. Jiang、Q. Chen、S. Zheng 和 Z. Zhao，“WavTokenizer：用于音频语言建模的高效声学离散编解码器分词器”，载于国际学习表征会议论文集，2025 年。[24] H. Zen、V. Dang、R. Clark、Y. Zhang、R. J. Weiss、Y. Jia、Z. Chen 和 Y. Wu，“LibriTTS：从 LibriSpeech 衍生的文本转语音语料库”，*arXiv* 预印本 *arXiv:1904 02882*，2019。。[25] J. Yamagishi、C. Veaux、K. MacDonald 等，“CSTR VCTK 语料库：用于 CSTR 语音克隆工具包的英语多说话人语料库（0.92 版）”，爱丁堡大学语音技术研究中心（*CSTR*），271–350 页，2019 年。[26] A. Jacot、F. Gabriel 和 C. Hongler，“神经切线核：神经网络中的收敛与泛化”，神经信息处理系统进展，第 31 卷，2018 年。[27] A. W. Rix、J. G. Beerends、M. P. Hollier 和 A. P. Hekstra，“语音质量感知评估（PESQ）——一种用于评估电话网络和编解码器语音质量的新方法”，载于*IEEE* 国际声学、语音与信号处理会议，第 2 卷，2001 年，749–752 页。[28] C. H. Taal、R. C. Hendriks、R. Heusdens 和 J. Jensen，“一种针对时频加权噪声语音的短时客观可懂度测量方法”，载于*IEEE* 国际声学、语音与信号处理会议，2010 年，4214–4217 页。[29] A. Hines、J. Skoglund、A. C. Kokaram 和 N. Harte，“ViSQOL：一种客观语音质量模型”，*EURASIP* 音频、语音与音乐处理期刊，2015 年，第 2015 卷，第 1 期，第 13 页。[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*IEEE* 国际声学、语音与信号处理会议，2019 年，626–630 页。