TQCodec：迈向　　面向高保真音乐流媒体的神经音频编解码器　　G

Lixing He[1,2] *Zhouxuan Chen*[1] *Mingshuai Liu*[1] *Xinran Sun*[1,3] *Wucheng Wang*[1]*Minfu Li*[1,4] *Lingcheng Kong*[1*] *Weifeng Zhao*[1] *Wenjiang Zhou*[1]

[1] 腾讯音乐娱乐，[2] 香港中文大学，[3] 东南大学[4] 清华大学

arXiv:2603.01592v1 [cs.SD] 2 Mar 2026

摘要

我们提出了 TQCodec，这是一种专为高比特率、高保真音乐流媒体设计的神经音频编解码器。不同于现有主要面向超低比特率（$\leq$ $16kbps$）的神经编解码器，TQCodec 以 44.1 kHz 运行，支持 32 kbps 至 128 kbps 的比特率，与现代音乐流媒体平台的标准音质相匹配。该模型采用基于 SEANet 的编码器-解码器架构，以实现高效的端侧计算，并引入了多项改进：通过非均衡网络设计，以较低开销提升音质；采用 SimVQ 保留中频细节；以及使用相位感知波形损失函数。此外，我们还提出了一种由感知驱动的分频段比特分配策略，以优先分配比特至感知上至关重要的低频段。在多种音乐数据集上的评估表明，TQCodec 在目标比特率下能够实现卓越的音频质量，因此非常适用于高质量音频应用。

索引词— 神经音频编解码器，音乐

1. 引言

兼具高保真与紧凑表示的音频编解码器，对于音频压缩和生成都至关重要。尽管 Opus 和 MP3 等传统编解码器已在众多应用中得到广泛采用，深度神经网络的兴起推动了神经音频编解码器（NAC）的发展。SoundStream、Encodec 和 DAC 等模型的性能已可与传统编解码器相媲美。尽管 NAC 无需手工特征工程，但大多数现有编解码器面向相对较低的比特率。在音乐流媒体等场景中，可用带宽可达 32 kbps 或更高，此时音频质量比压缩效率更为重要。这促使我们需要一种面向高比特率、高质量应用的新型神经音频编解码器。

本文提出 TQCodec，这是一种神经编解码器，可用 32 kbps 至 128 kbps 的较高比特率对 44.1 kHz 音乐进行编码，与音乐流媒体的标准比特率相匹配。相比之下，大多数神经编解码器的比特率要低得多（主要为 16 kbps 或更低）。因此，有必要探索能够在高比特率下仍使编解码器受益的设计。TQCodec 在时域对音频建模，沿用成熟的编码器-解码器架构，与 DAC [1] 和 Encodec [2]相同。不同的是，我们重新设计了编解码器，使其更适合终端设备（即移动电话）的限制：以 SEANet[3], 取代 DAC 的编码器和解码器，后者具有较短的感受野和较低的计算开销。此外，我们提出了几项技术以进一步提升性能，包括一对不均衡的网络，在保持计算轻量的同时提升质量；保留中频细节的 SimVQ；以及用于补偿相位的波形损失。最后，我们提出一种由感知驱动的分频段编解码器分配方案，可将更多资源分配给对听感影响最大的低频。

TQCodec 在大规模、多样化的音乐数据集上，以 32 kbps 至 128 kbps 的不同比特率进行评估。我们同时给出客观评估和频谱图可视化，以验证本设计的优势。

2. 相关工作

神经音频编解码器通常采用 VQ-GAN 框架，包含编码器-解码器结构和量化瓶颈，并且常使用残差向量量化（残差向量量化）。代表性工作包括 SoundStream [4], Encodec [2] （采用多尺度 STFT 判别器）以及 DAC [1]（采用增强型 RVQ 设计）。其中，尽管 DAC 发布较早，仍然是一个强有力的基线。

音乐凭借其宽带信号和丰富的谐波，为音频编解码器带来了独特挑战，既要求清晰度，也要求高保真。近期音乐生成研究的兴起凸显了对先进音频编解码器的需求。虽然 [5]等系统利用 SpectroStream [6] 进行音乐合成，但其 16kbps 的比特率通常难以保留音乐的保真度。采用神经声码器的替代方法 [7] 可以提升质量，但会带来高昂的计算成本，使其不适用于实时流媒体传输。TQCodec 弥合了这一差距，为音乐应用提供了兼具效率和高保真重建能力的解决方案。

* 通讯作者：lesterkong@tencent.com

另一方面，相当一部分编解码器研究聚焦于语音领域，在该领域，低带宽至关重要且已足够。将这些编解码器直接应用于生成任务时，通常希望采用低比特率和单一码本[8, 9]。其他工作则通过引入语义知识 [10], 来扩展 NAC；尽管它覆盖包括音乐在内的通用音频，但仍侧重于超低比特率。此外，研究者还提出了 Score-Dec [11] 和 FlowDec [12] ，通过使用扩散模型进行后滤波来提升 NAC 的感知质量。然而，额外的模块可能会大幅增加实时因子，使其不适用于音乐流媒体。

3. 方法

我们的模型基于 DAC [1], 构建，遵循编码器-量化器-解码器的端到端流程。如表 1 所示，TQCodec 的工作比特率要高得多。

表 1：压缩方法比较（SR：采样率，BR：比特率，CB：码本，FR：帧率）。

编解码器	SR（kHz）	BR（kbps）	CB	FR
TQCodec	44.1	32, 64, 128	5/ 10/ 20	689
DAC [1]	44.1	16	18	86
Encodec [2]	24/ 48	24	16	75 /150

3.1. 改进的神经音频编解码器

3.1.1. 编码器和解码器

音频编解码器包含两个阶段：编码和解码。在音乐流媒体场景中，编码阶段通常在云服务器上执行一次，其计算资源充足，因此复杂度易于应对，并非主要瓶颈。相比之下，解码阶段发生在终端用户设备上，这些设备主要是计算能力各异且往往有限的智能手机。为确保广泛的可及性和流畅的用户体验，解码过程必须保持在实际可行的计算限制之内。

对于 TQCodec，我们同时考虑计算量和延迟，其中计算量受到每秒不超过 10 十亿次乘加运算（GMACs）的限制。遗憾的是，DAC 解码器的默认配置约为 365 GMACs，远超我们的限制。降低乘加运算次数的一种直接方法是缩减模型维度。例如，我们可以将解码器维度从 1536 降至 128，从而有效减少计算量。然而，我们注意到，感受野仍保持在 17704 个采样点，这主要是由 DAC 的膨胀和较高的下采样率导致的。如此大的感受野可能造成很高的延迟，尤其是在流式窗口较短时。

为了解决这一问题，我们用 Encodec 的编码器和解码器结构替换了原有结构。Encodec 采用高效的 SEANet架构 [3] ，并通过 LSTM层加以增强。该架构专为提高计算效率、降低延迟以及兼容流式应用而设计。因此，解码器的计算量成功降至约 6.31 GMACs，而感受野也缩小至仅 2410 个采样点。此外，我们发现编码器默认需要约 2 GMACs。由于这不会影响解码过程，我们决定根据经验将编码器的计算量提高至约 80 GMACs。

3.1.2. 量化器

沿用 DAC 的设计，我们保留了其改进的残差向量量化（RVQ）框架，该框架具有较高的码本利用率。然而，在听音测试中，我们发现即使尺度不变信噪比（SI-SNR）已经很高，解码音频中仍含有可感知的噪声。对分频段重建误差的进一步分析表明，RVQ 难以准确建模中频范围（3000–8000 Hz）。为解决这一问题，我们提议将 RVQ 结构中的标准向量量化层替换为 SimVQ 模块，如 [13], 中所述；该模块的码本被冻结，编码则通过线性投影隐式生成。为实现更高的比特率，我们将其扩展为残差SimVQ，其中每个 SimVQ 都对前一个模块产生的残差进行量化。

3.1.3. 损失函数

DAC 已引入判别器、损失函数及相应权重的组合，以促进模型稳定收敛。具体而言，生成器采用梅尔频谱图损失，其中包含幅度和对数幅度的 L1 损失。对抗损失则包括多周期判别器和多频带多尺度 STFT 判别器，并采用 HingeGAN [14]对抗损失公式和特征损失 [15]。此外，我们还采用简单的码本损失和承诺损失，并使用原始 VQ-VAE 公式中的停止梯度 [16]。

然而，正如 [11] 所指出的，生成器所用的梅尔频谱图损失难以解决相位混淆问题。虽然 DAC 模型可以通过结合判别器进行对抗训练来缓解这一问题，但我们发现，加入波形损失能带来更进一步的改进。对于各项损失函数的权重，我们采用如下设置：多尺度梅尔损失为 15.0，波形损失为 1.0，特征匹配损失为 2.0，对抗损失为 1.0。此外，码本损失的权重为 1.0，承诺损失的权重为 0.25。这些权重设置与近期关于梅尔损失的研究 [1] 一致。

3.2. 感知驱动子带建模

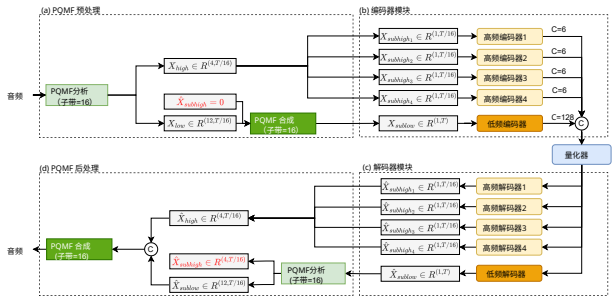


图 1：子带建模示意图。

人类听觉系统通常能够感知 20 Hz 至 20,000 Hz 范围内的声音频率。然而，人类对这一范围的感知并非线性的——与较高频段相比，人耳对较低频段的频率差异更为敏感。为更贴近人类的感知特性，人们开发了多种心理声学频率尺度，包括梅尔尺度和巴克尺度。以 Opus [17] 为例，它采用巴克尺度；当比特率不足时，算法会舍弃高频部分。

在 TQCodec 的设计中，如图 1 所示，我们采用了与神经音频编解码器相似的结构。值得注意的是，NAC 通过对原始波形数据进行一维卷积来运行，这使其有别于显式建模各个频带的基于频谱图的编解码器。相反，我们采用多相正交调制滤波器（PQMF），将输入波形分解为 16 个子带。前 12 个子带被指定为核心频带，对人类听觉感知最为重要，因此分配了大部分计算资源。我们为核心频带保留了与先前配置相同的网络维度，同时将其余子带的维度缩减至四分之一，因此仅额外增加 0.17 MACs。为保持比特率一致，量化前将所有子带的潜在表示拼接起来，其中核心频带的维度设为 128，其余每个子带的维度设为 6。

4. 实验

4.1. 训练方案

我们使用公开和私有数据集对 TQCodec 进行训练和测试，具体信息列于表 2。在默认实验中，我们在一台 8× H20 机器上以 32 的批量大小训练每个模型，迭代 400k 次。实际训练约需 30 小时。训练期间，我们将音频样本的时长设为 1 秒。生成器和判别器均使用 AdamW 优化器，学习率为 0.0001，$\beta_1 = 0.8$ 和 $\beta_2 = 0.9$。

表2：数据集信息。

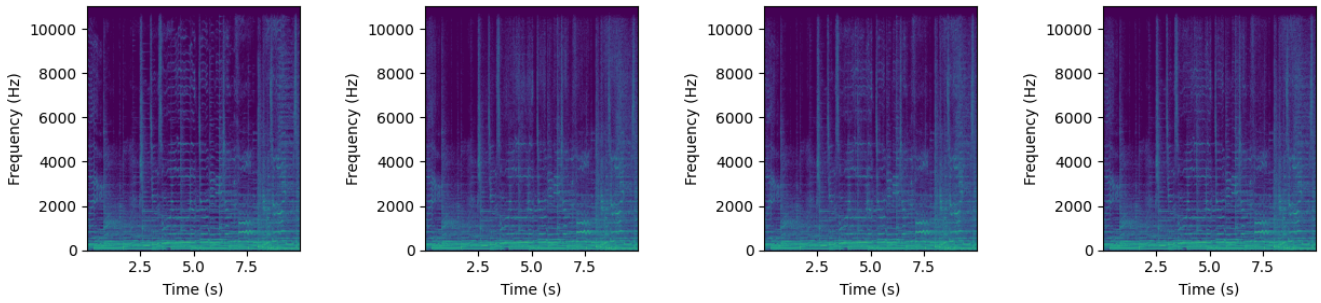
Name	曲目数量
MusDBHQ [18]	150
Jingju [19]	120
Jamendo [20]	55,609
Fma [21]	106,574
私有数据集	100000+

对于 64 kbps 的立体声配置（每个声道 32 kbps），模型采用 3 个编码器模块和 3 个解码器模块，下采样因子为 [2,4,8]。模型使用 5 个码本，每个码本大小为 512，因此比特率为$44100/(2*4*8)*5*\log_2(512) = 31007$。对于 128 kbps 的立体声配置，码本数量增加到 9，因此比特率为 $44100/(2*4*8) * 10 * \log_2(512) = 62015$。编码器维度设为 64，潜在维度设为 128，解码器维度设为 128。请注意，实际比特率并不完全符合预期值，而是低于预期，这主要是因为采样率为 44.1 kHz。对于双通道音频，我们将 TQCodec 应用两次，因为我们通过经验发现，双通道神经网络的表现不佳。由于损失函数和判别器设计与 DAC 类似，因此我们沿用相同的 [1]配置。

评估指标：我们使用对数谱距离（LSD）作为主要指标，因为它更关注全频段性能，而非由低频段主导的性能。具体而言，它衡量重建音频与真实音频之间的频域质量：$\mathrm{LSD}(x, y) = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(X(l,k) - \hat{X}(l,k)\right)^2}$，其中 l 和 k分别为时间和频率索引，$X = \log(|\text{短时傅里叶变换}(y)|^2)$，而 $\hat{X} = \log(|\text{短时傅里叶变换}(x)|^2)$。此外，我们也使用信噪比（SNR）作为指标，它衡量相对于噪声的信号质量，计算方式为 $\mathrm{SNR}(x, y) = 20\log_{10}\left(\left(\frac{y}{x-y}\right)^2\right)$，其中 x 为估计音频，y 为干净音频。

4.2. 评估

TQCodec由多个不同组件构成，因此有必要评估添加每个组件所带来的性能提升，如表3所示。首先，我们在第一行将 Encodec 编码器和解码器与 DAC 中改进的残差向量量化相结合；与基线相比，该方案取得了具有竞争力的性能，同时所需计算量显著更少。接着，我们依次加入 RSimVQ、波形损失函数和不平衡自编码器。研究结果表明，就对数谱距离的改善而言，不平衡自编码器是影响最大的组件，尤其是在低频范围内。值得注意的是，消融实验中的各组件均不会增加解码过程的计算量。在表格最后，我们将 TQCodec 扩展至 64千比特每秒和



(a) 干净 (b) 残差向量量化 (c) RSimVQ (d) 子带建模

图2：频谱图可视化，其中残差向量量化无法清晰地重建中频（高于 4000Hz）。**表3**：所提出的编解码器的消融研究结果，其中对数谱距离-L 表示低于 16000Hz 的频率，对数谱距离-H 表示高于 16000Hz 的频率。

Name	对数谱距离 ↓	低频对数谱距离↓	高频对数谱距离 ↓	编码器乘加运算次数 ↓	解码器乘加运算次数 ↓	感受野 ↓
DAC（16kbps）	0.927	0.877	1.016	25G	365G	17706
DAC-small（32kbps）	0.833	0.781	0.936	25G	365G	17706
TQCodec 消融实验（32kbps）						
Encodec + DAC RVQ	0.854	0.790	0.983	2.4G	6.31G	2410
+ RSimVQ	0.850	0.772	0.997	2.4G	6.31G	2410
+ 波形损失	0.844	0.769	0.995	2.4G	6.31G	2410
+不平衡自编码器	**0.822**	**0.757**	**0.949**	72G	6.31G	2410
64千比特每秒	0.7715	0.647	0.950	72G	6.31G	2410
128kbps	0.6716	0.5166	0.9517	72G	6.31G	2410

128kbps，并观察到低频对数谱距离持续下降，而高频对数谱距离已经趋于饱和。

此外，我们分析了 RSimVQ 对中频重建的影响。如表3所示，RSimVQ 能有效降低低频对数谱距离。更重要的是，如图2所示，我们发现它能够缓解损害聆听质量的伪影。尽管客观指标可能无法充分反映这些伪影，但它们会显著影响主观感知。具体而言，图2b中的残差向量量化输出在 5000–10000 Hz 之间呈现出模糊的频谱图，而 RSimVQ 则成功恢复了该区域中的部分谐波结构，如图2c所示。

在对子带建模进行分析时，我们将 TQ- Codec 与 Ogg-Vorbis [17] 进行比较，如表4所示。我们观察到，在比特率为 32kbps 时，Ogg-Vorbis 会显著丢弃高频内容，导致 LSD-High分数大幅降低。然而，借助我们的子带建模技术，我们在提升低频对数谱距离和信噪比的同时，仅略微增加了高频对数谱距离。如图2d所示，我们的子带建模在大幅提升低频质量的同时，保持了相似的谐波结构。

5. 结论

我们提出了TQCodec，这是一种用于 44.1千赫兹高比特率音乐流媒体（32-128kbps）的神经音频编解码器。关键创新—

表4：与基线方法的比较，用于验证子带建模。

模型	比特率	低频对数谱距离 ↓	高频对数谱距离 ↓	信噪比 ↑
Ogg-Vorbis [17]	48kbps	0.763	2.847	16.998
TQCodec	43kbps	0.718	0.954	16.074
+ 子带	43千比特每秒	0.702	1.163	16.88

表5：主观听音测试：AICodec-64kbps 与 Ogg-96kbps（高质量）

MOS	
平均值	4.18
通过率	100%
优秀率	77.8%
偏好	
高质量（Ogg）更好	39.4%
AICodec更好	39.1%
无显著差异	21.5%

主要贡献包括：(1) 基于 SEANet 的架构，可实现高效的设备端解码（6.31 十亿次乘加运算）；(2) 使用 SimVQ 保留中频；(3) 相位感知波形损失；(4) 感知优化比特分配。评估结果表明，该编解码器在 LSD/SNR指标上优于基线方法，同时保持较高的计算效率。该编解码器尤其适用于要求高保真的音乐流媒体应用。

6. 参考文献

[1] Rithesh Kumar、Prem Seetharaman、Alejandro Luebs、IshaanKumar 和 KundanKumar，《通过改进 rvqgan 实现高保真音频压缩》，神经信息处理系统进展，第 36 卷，第 27980–27993 页，2023 年。

[2] AlexandreD´efossez、JadeCopet、GabrielSynnaeve 和 Yossi Adi，《高保真神经音频压缩》，*arXiv* 预印本 *arXiv:2210.13438*，2022 年。

[3] Marco Tagliasacchi、Yunpeng Li、Karolis Misiunas 和 Dominik Roblek，《Seanet：一种多模态语音增强网络》，*arXiv* 预印本 *arXiv:2009.02095*，2020 年。

[4] Neil Zeghidour、Alejandro Luebs、Ahmed Omran、Jan Skoglund 和 Marco Tagliasacchi，《Soundstream：一种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，第 495–507 页，2021 年。

[5] LyriaTeam、Antoine Caillon、Brian McWilliams、Cassie Tarakajian、Ian Simon、Ilaria Manco、Jesse Engel、Noah Constant、Pen Li、Timo I Denk 等，《现场音乐模型》，*arXiv* 预印本 *arXiv:2508.04651*，2025 年。

[6] Yunpeng Li、Kehang Han、Brian McWilliams、Zalan Borsos 和 Marco Tagliasacchi，“Spectrostream：一种通用音频通用神经编解码器”，*arXiv* 预印本 *arXiv:2508.05207*，2025。

[7]Ge Zhu、Juan-Pablo Caceres、Zhiyao Duan 和 Nicholas J Bryan，“Musichifi：快速高保真立体声声码合成”，*IEEE* 信号处理快报，2024。

[8] Hanzhao Li、Liumeng Xue、Haohan Guo、Xinfa Zhu、Yuanjun Lv、Lei Xie、Yunlin Chen、Hao Yin 和 Zhifei Li，“Single-codec：面向高性能语音生成的单码本语音编解码器”，*arXiv* 预印本 *arXiv:2406.07422*，2024。

[9] Detai Xin、Xu Tan、Shinnosuke Takamichi 和 Hiroshi Saruwatari，“Bigcodec：突破低码率神经语音编解码器的极限”，*arXiv* 预印本 *arXiv:2409.05377*，2024。

[10] Haohe Liu、Xuenan Xu、Yi Yuan、Mengyue Wu、Wenwu Wang 和 Mark D Plumbley，《SemantiCodec：面向通用声音的超低比特率语义音频编解码器》，*IEEE* 信号处理选题期刊，2024。

[11] Yi-Chiao Wu、Dejan Markovi´c、Steven Krenn、Israel D Gebru 和 Alexander Richard，《Scoredec：一种采用广义基于分数的扩散后置滤波器、保留相位的高保真音频编解码器》，发表于 *ICASSP2024-2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024 年，第 361–365 页。[12]Simon Welker、Matthew Le、Ricky TQ Chen、Wei-Ning Hsu、Timo Gerkmann、Alexander Richard 和 Yi-Chiao Wu，《Flowdec：一种具有高感知质量的基于流的全频带通用音频编解码器》，*arXiv* 预印本 *arXiv:2503.01485*，2025 年。[13] Yongxin Zhu、Bocheng Li、Yifei Xin、Zhihua Xia 和 Linli Xu，《利用一个线性层解决向量量化模型中的表示坍塌问题》，*arXiv* 预印本 *arXiv:2411.02038*，2024 年。

[14] Jae Hyun Lim 和 Jong Chul Ye，《几何生成对抗网络》，*arXiv* 预印本 *arXiv:1705.02894*，2017。

[15] Kundan Kumar、Rithesh Kumar、Thibault De Boissiere、Lucas Gestin、Wei Zhen Teoh、Jose Sotelo、Alexandre De Brebisson、Yoshua Bengio 和 Aaron C Courville，《MelGAN：用于条件波形合成的生成对抗网络》，神经信息处理系统进展，第 32 卷，2019。

[16] Aaron Van Den Oord、Oriol Vinyals 等，《神经离散表示学习》，神经信息处理系统进展，第 30 卷，2017。

[17] Jack Moffitt，“Ogg Vorbis——开放、免费的音频——让你的媒体自由，”*Linux Journal*，2001年，第2001卷，第81es期，第9–es页，2001年。

[18]Zafar Rafii、Antoine Liutkus、Fabian-Robert St¨oter、Stylianos Ioannis Mimilakis和Rachel Bittner，“MUSDB18-HQ——MUSDB18的未压缩版本，”（无标题），2019年。

[19] Rong Gong、Rafael Caro Repetto和Xavier Serra，“面向自动京剧演唱评估研究创建无伴奏歌唱音频数据集”，载于第*4*届音乐学数字图书馆国际研讨会论文集，2017年，第37–40页。

[20] Dmitry Bogdanov、Minz Won、Philip Tovstogan、Alastair Porter和Xavier Serra，“用于自动音乐标注的MTG-Jamendo数据集”，ICML，2019年。

[21] Micha¨el Defferrard、Kirell Benzi、Pierre Vandergheynst 和 Xavier Bresson，《Fma：一个用于音乐分析的数据集》，*arXiv* 预印本 *arXiv:1612.01840*，2016年。