arXiv:2603.07285v1 [eess.AS] 7 Mar 2026

基于 Vocos 的快速灵活音频带宽扩展

Yatharth Sharma

独立研究者

yatharthsharma3501@gmail.com

摘要

我们提出了一种基于 Vocos 的带宽扩展（BWE）模型，通过生成缺失的高频内容，将 8–48 kHz 音频增强至 48 kHz。输入信号会重采样至 48 kHz，并由神经声码器主干网络处理，使单个网络能够支持任意上采样比率。受 Linkwitz-Riley 分频器启发的轻量级细化器通过平滑分频，将原始低频带与生成的高频合并。在验证中，该模型取得了有竞争力的对数谱距离，同时在 NVIDIA A100 GPU 上的实时因子为 0.0001，在 8核中央处理器上的实时因子为 0.0053，展现了该模型在极高吞吐量下实现实用、高质量带宽扩展的能力。

索引词：带宽扩展、神经声码器、音频超分辨率、语音增强

1. 引言

带宽扩展（BWE）旨在恢复或幻觉生成以有限带宽采集的音频信号中缺失的频率成分，例如旧式录音或电话语音。传统信号处理方法，包括基于插值的上采样和频谱整形，具有较高的效率，但往往无法重建听感上逼真的高频细节。近年来，基于学习的方法将带宽扩展重新表述为条件生成任务，根据观测到的低频带信号预测缺失的高频带内容。

最先进的基于扩散的模型（如 Au- dioSR [1] ）展现出了卓越的生成质量。然而，其迭代采样过程通常具有过高的计算成本，难以用于实时或大规模部署。相比之下，基于 GAN 的方法（如 AP-BWE [2] ）在速度方面具有显著优势，但其中许多方法仍受限于固定的输入/输出采样率对（例如 16 kHz $\rightarrow$ 48 kHz）。这一限制妨碍了其在异构现实世界流程中的灵活性，因为此类流程的输入采样率变化很大。

本文提出了一种灵活的带宽扩展系统，可在单个网络内支持 8–48 kHz 范围内的任意输入采样率。我们以 Vocos [3] 架构为基础，利用其高保真的傅里叶域神经声码器架构来合成缺失的高频细节。通过将所有输入重采样至目标采样率 48 kHz，并采用 Vocos 风格的生成器，我们的模型能够在各种上采样倍率下保持一致性。为进一步提升保真度，我们引入了一种受林克维茨-莱利启发的分频器 [4]启发的轻量级频域细化器。该细化器平滑地融合原始重采样低频带信号与生成的高频带成分，确保在分频频率处实现相位连贯过渡。

在 VCTK 语料库上的实验结果 [5] 表明，我们所提模型的表现可与强大的基线方法媲美，甚至更优；在对数谱距离（LSD）方面表现出色，同时吞吐量比基于扩散模型的方法高出几个数量级。

本文的主要贡献如下：

- 据我们所知，我们首次提出了基于 Vocos 的带宽扩展模型，利用神经声码器为任意输入采样率生成高频内容。
- 我们提出一种受林克维茨-莱利启发的频域细化器，通过将原始低频带与合成高频带无缝结合来提升感知质量。
- 我们展示了更优的质量与速度权衡，在 NVIDIA A100 上实现了 0.0001 的实时因子（RTF），并在标准 8核中央处理器上保持显著快于实时的性能。

2. 所提方法

2.1. 问题表述

令 $x \in \mathbb{R}^T$ 表示采样率为 48 kHz 的波形，令 $x_{\downarrow r}$ 表示带限、降采样后的观测信号，其采样率为 $r \in \{8, 16, 24\}$ kHz（更一般地，适用于任意 $r \in [8, 48]$ kHz）。带宽扩展的目标是估计采样率为 48 kHz 的高频增强波形 $\hat{x}$ ，使其在感知和频谱上都接近 x，同时与 $x_{\downarrow r}$ 所蕴含的低频带内容保持一致。

所提模型首先使用 Sinc 插值将输入重采样至 48 kHz，以获得基带波形 $y \in \mathbb{R}^T$。该波形保留了可用的低频带信息，但缺少真实的高频细节。随后，生成器 G_θ 预测增强波形 $\tilde{x} = G_\theta(y)$，再由频域细化器生成最终输出$\hat{x}$。

2.2. 模型架构

生成器对从重采样至 48 kHz 的输入中提取的时频表示（梅尔频谱图）进行处理，并生成复数短时傅里叶变换表示，再通过逆短时傅里叶变换转换为波形。

特征提取。 我们使用 80 个梅尔频带计算梅尔频谱图，$n_{\text{fft}} = 2048$，在 48 kHz 采样率下帧移为 512。该梅尔表示用作生成器的条件输入。

主干网络。 主干网络从头初始化。其核心由 8 个 ConvNeXt 风格残差块组成，模型维度为 $C = 512$。每个块都采用 7×1 深度卷积进行时序建模，随后通过前馈网络将特征扩展至 1536 的中间维度，再投影回 512 个通道。

整个网络中均采用层归一化和 GELU 激活函数，以确保训练稳定。最终的潜在表征

会经过一个逐点卷积层和一个线性头，后者用于预测基于逆短时傅里叶变换的重建所需的复数系数。这种配置使模型能够将 80 个频带的梅尔频谱图输入映射为高保真度波形，同时保持时间分辨率恒定。

输出头与波形合成。线性头预测短时傅里叶变换分量，并通过逆短时傅里叶变换将其转换为波形。模型从头开始初始化；在此设置下，输入被重采样至 48 kHz，模型经过训练后生成缺失的高频内容，而非仅仅重建输入频带。

2.3. 受 Linkwitz-Riley 启发的频率细化器

神经生成器可能会在输入已包含可靠信息的区域引入细微的不一致或伪影。所提细化器构建了一个分频掩码$M(f) \in [0,1]$，以便将低频锚点$Y(f)$与生成的高频内容$\tilde{X}(f)$无缝融合。在以截止频率f_c为中心的过渡区域内，我们将归一化频率定义为$t = (f - f_{start})/(f_{end} - f_{start})$。该掩码遵循一条平滑的多项式曲线：

$$M(f) = \begin{cases} 0 & f < f_{start} \\ 3t^2 - 2t^3 & f_{start} \le f \le f_{end} \\ 1 & f > f_{end} \end{cases} \quad (1)$$

ˆ随后，通过对两个信号进行线性插值，计算得到细化频谱$X(f)$：

$$\hat{X}(f) = (1 - M(f))\, Y(f) + M(f)\, \tilde{X}(f) \quad (2)$$

最终波形$\hat{x}$通过逆实数快速傅里叶变换（iRFFT）获得。这种形式可确保分频器处的幅度响应平坦，同时抑制连接处的相位不连续。因此，所提细化器显著提升了最终质量。

2.4. 训练目标

所提模型通过结合频谱、感知和对抗损失进行训练，以确保结构准确性和高频真实感。

多分辨率短时傅里叶变换损失（MRSTFT）。短时傅里叶变换损失在$n_{\text{fft}} \in \{512, 1024, 2048\}$下计算，跳跃长度为$n_{\text{fft}}/4$。这种多尺度方法有助于在不同的时频分辨率下实现准确重建，同时捕捉细粒度的时间事件和长期频谱包络。

梅尔频谱图损失。我们对真实值和生成音频的梅尔频谱图计算L_1损失。计算采用 48 kHz、128 个梅尔频带和$n_{\text{fft}} = 2048$，使模型聚焦于感知上相关的频带。

多分辨率判别器（MRD）。MRD [6]由多个子判别器组成，它们在不同尺度上处理波形的复数短时傅里叶变换（STFT）。通过使用不同的窗长在频域分析信号，MRD 能够同时惩罚高频瞬态中的伪影（短窗），并保留浊音语音的谐波结构（长窗）。这可以避免仅使用时域判别器时常见的“相位模糊”现象。

对抗损失与特征匹配损失。为进一步提升稳定性和保真度，我们引入特征匹配损失（$\mathcal{L}_{\text{fm}}$）。该损失用于最小化判别器处理真实音频与生成音频时其中间特征图之间的L_1距离：

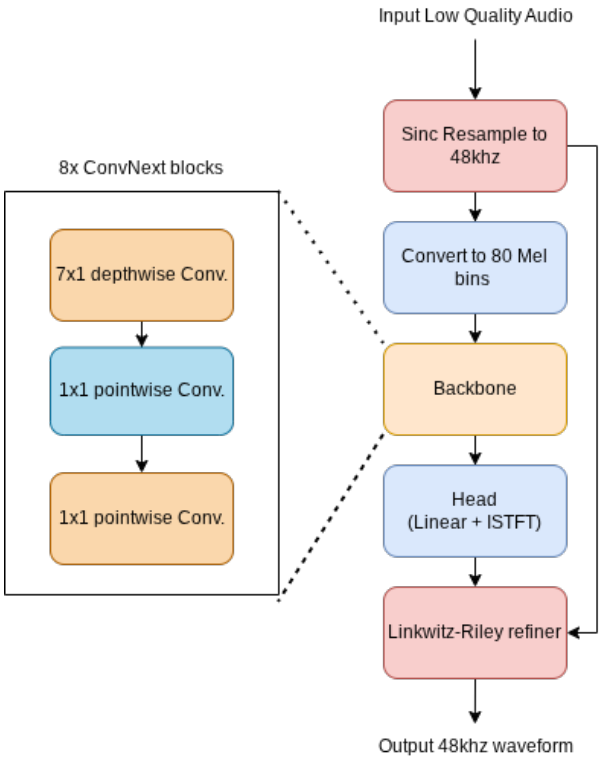

图 1：所提带宽扩展架构的概览，该架构遵循 *Vocos* 框架。输入音频通过 *Sinc* 插值上采样至 *48 kHz*，并转换为 *80* 个频带的梅尔频谱图。主干网络由 *8* 个 *ConvNeXt* 块组成，采用7×1深度卷积和1×1逐点卷积。输出头使用线性层和逆短时傅里叶变换（*ISTFT*）重建波形，随后通过受林克维茨-莱利启发的细化器显著提升音质。

判别器处理真实音频与生成音频时其中间特征图之间的距离：

$$\mathcal{L}_{\text{fm}}(G; D) = \mathbb{E}\left[\sum_{i=1}^{L} \frac{1}{N_i} ||D^{(i)}(x) - D^{(i)}(\tilde{x})||_1\right] \quad (3)$$

其中，$D^{(i)}$和N_i分别表示判别器第i层的特征和元素数量。这会促使生成器生成与真实语音具有相同多尺度统计特性的音频。

优化。使用 AdamW，学习率为10^{-4}，权重衰减为10^{-2}，并采用指数衰减调度器，每 64 步将学习率乘以 0.99。模型以 16 为批量大小进行训练。

3. 实验设置

3.1. 数据集

我们使用 VCTK 语料库（约 44 小时语音）进行训练，随机裁剪约 60,000 个时长为 1.28、2.56 或 3.2 秒的音频片段。

3.2. 退化过程

干净的 48 $r \in \{8, 12, 16\}$ kHz 音频会随机使用 sinc、零阶保持或线性插值降采样至 kHz。可选地，

加入量化噪声以增强模型的鲁棒性。劣化信号会先重采样至 48 kHz，再输入生成器。

3.3. 基线方法

基线方法包括：

• **Sinc 上采样** – 标准插值。

• **AP-BWE**– 受 APNet2 启发、针对特定倍率训练的生成对抗网络。

• **NVSR** [7] – 基于神经生成对抗网络声码器的超分辨率基线方法。

• **AudioSR** – 基于扩散的音频超分辨率。

3.4. 评估指标

对数谱距离（LSD）。LSD 用于衡量参考信号 x 与增强信号$\hat{x}$在对数分贝尺度上的功率谱之间的距离。给定功率谱 $P(k,\ n) = |X(k,\ n)|^2$和 $\hat{}P(k,\ n) = |\hat{}X(k,\ n)|^2$，其中 k和 n 分别表示频率箱和时间帧，距离定义为：

$$\mathrm{LSD}(x,\hat{x}) = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\log_{10}P(k,n) - \log_{10}\hat{P}(k,n)\right)^2}$$

其中， K 是频率箱的数量， N 是时间帧的数量。数值越低，表示频谱重建效果越好。

ViSQOL。ViSQOL [8]使用一种旨在与人类平均意见分相关的相似度度量，比较参考频谱图与处理后频谱图的局部块。评分范围为1–4.75，越高越好。

4. 结果与分析

4.1. 质量基准测试

方法	8→48 千赫兹 （对数谱距离 ↓）	12→48 千赫兹 （对数谱距离 ↓）	16→48 千赫兹 （对数谱距离 ↓）
Sinc 上采样	3.52	3.19	2.94
AudioSR	1.61	1.52	1.44
NVSR	1.22	1.20	1.12
AP-BWE	0.87	0.81	0.74
所提模型	**0.85**	**0.80**	**0.74**

表 1：*VCTK* 上的对数谱距离（*LSD*）对比。我们的模型在所有档位上都实现了极具竞争力的频谱重建效果。

方法	8→48 千赫兹 （ViSQOL ↑）	12→48 千赫兹 （ViSQOL ↑）	16→48 千赫兹 （ViSQOL ↑）
Sinc 上采样	2.10	2.23	2.35
AudioSR	3.15	3.21	3.40
NVSR	2.97	3.00	3.10
AP-BWE	**3.51**	**3.54**	**3.70**
所提模型	**3.51**	3.53	3.69

表2：*VCTK*上的*ViSQOL*分数。我们的模型与*AP-BWE*的感知质量相近。

4.2. 性能分析

表1和表2中的实验结果证明了所提模型在各种上采样尺度上的有效性，评估同时采用了客观谱失真和感知质量指标。

谱保真度：如表 1 所示，我们的模型在所有场景下的对数谱距离（LSD）方面均取得了极具竞争力的表现。值得注意的是，在 $8 \rightarrow 48$千赫兹任务中，所提模型的 LSD 达到 **0.85**，优于基于扩散模型的 AudioSR（1.61）和 NVSR（1.22）。在 12 千赫兹（**0.80**）和 16 千赫兹（**0.74**）下持续领先，并达到或超越 AP-BWE，这表明我们的架构能够有效捕捉高频谐波相关性，并以比现有生成式或预测式基线更高的精度重建谱包络。

感知质量：从 ViSQOL 分数（表 2）来看，所提模型展现出稳健的感知表现。在 $8 \rightarrow 48$ 千赫兹下，我们取得了 **3.51** 的分数，与高性能 AP-BWE 基线相当。尽管 AP-BWE 在 12 千赫兹和 16 千赫兹基准测试中略占优势（< 0.01），我们的模型仍具有很强的竞争力，跻身顶尖表现之列，并在 16 千赫兹下保持 3.69 的分数。这表明，虽然我们的模型针对谱准确度进行了优化，但其感知质量与计算成本高得多的模型几乎无法区分。

讨论：结果表明，所提模型在所有评估的采样场景中均实现了极具竞争力的谱保真度，其对数谱距离始终低于所有基线方法。尽管 AP-BWE 在某些输入采样率下取得了略高的 ViSQOL分数，但感知差异仍然很小。重要的是，我们的方法以显著更低的计算成本取得了这些结果，并在统一框架内支持任意输入采样率。与现有的扩散模型和基于 GAN 的方法相比，这种兼具卓越谱准确度、出色感知质量和极高效率的特性，实现了良好的质量—效率权衡。

4.3 域外采样率下的鲁棒性

虽然模型主要是在常见的采样锚点（8、12、16 千赫兹等）上训练的，但我们架构的一项关键优势是能够处理任意输入带宽。为评估这种“零样本”泛化能力，我们在各种域外（OOD）采样率下测试了模型。

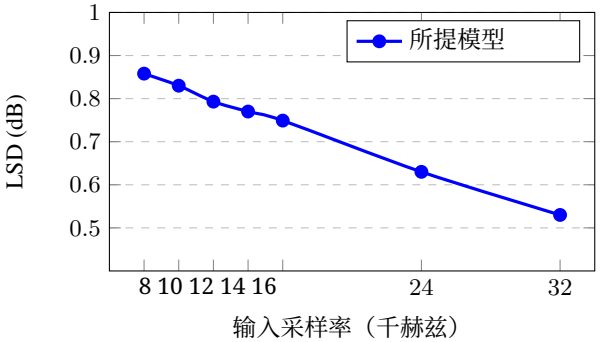

图2： 模型在域内和域外（*OOD*）采样率下的性能（对数谱距离）。曲线表明，随着输入带宽增加，保真度呈线性提升，无论该采样率是否在训练期间出现过。

域外泛化分析。如图2所示，随着输入采样率的提高，对数谱距离（LSD）单调下降。值得注意的是，模型在域外采样率（例如 10、14、24 和 32 千赫兹）下的表现遵循与域内采样率相同的线性趋势。我们并未针对这些采样率进行专门训练。例如，在 10 千赫兹（域外）下，模型的对数谱距离为 0.830，介于 8 千赫兹和 12 千赫兹的基准值之间。

这种出色的域外表现主要由两个因素驱动。首先，在输入生成器之前将所有输入重采样至 48 kHz 后，ConvNeXt 主干网络会将带宽扩展视为固定网格上的频谱补全任务。其次，也是更重要的一点，**受林克维茨启发的频率细化器** 支持动态交越。与可能针对固定的 8 kHz 或 16 kHz 边界进行优化的传统滤波器不同，我们的细化器采用平滑过渡，同时保留重采样输入中已有的能量。通过在频域中合并原始低频带与生成的高频带，细化器确保“幻觉”内容始终以真实信号为锚点，从而避免在非标准采样率下使用固定采样率带宽扩展系统时常见的不连续性或“金属感”伪影。

4.4. 消融研究

我们开展了一项消融实验，以评估不同的频谱合并策略如何影响最终输出质量。我们将受林克维茨-莱利启发的方法与三种基线配置进行比较：

• **无细化器：** 直接使用模型的原始输出。由于未将低频锚定到真实值，模型通常会在低频区域产生轻微伪影。

• **朴素高通/低通：** 在频域中实现的简单砖墙式截止。这会产生陡峭的过渡，但通常会在截止频率附近造成振铃伪影。

• **标准巴特沃斯：** 对四阶巴特沃斯响应的频域近似。虽然比朴素截止更平滑，但将两个频带相加时，通常会在交越点产生 +3 dB 的幅度尖峰。

•**所提方法（受 LR 启发）：** 我们的方法采用平方幅度响应，确保交越处的叠加结果平坦，从而将生成的高频与低频锚点平滑衔接，避免在交越点产生幅度突增。

细化器变体	对数谱距离 ↓ (8→48)
无细化器（直接输出）	0.897
朴素高通/低通分频器	0.865
标准巴特沃斯细化器	0.861
受林克维茨-莱利启发（所提方法）	**0.850**

表3： *VCTK* 验证集上的消融实验结果。所提细化器通过确保分频处的幅度响应平坦，取得了最低的频谱距离。

表3中的结果表明，与“无细化器”基线相比，即使是基础频谱锚定也能显著改善对数谱距离（0.897 对 0.865）。然而，巴特沃斯滤波器固有的幅度峰值限制了其性能。受 LR 启发的细化器将对数谱距离降至 0.850，证实在合并过程中保持平坦的幅度响应对于实现高保真度带宽扩展至关重要。

4.5. 效率

我们在中央处理器和 GPU 环境下评估了所提出模型的计算复杂度和推理速度，以考察其在现实场景中的部署可行性。我们将该方法与 AudioSR、NVSR 和 AP-BWE 进行比较，所有基准测试均采用固定的 4 秒音频时长。为确保公平，我们排除预处理步骤（如重采样），并为所有模型设置预热阶段。我们采用扩散模型推荐的采样步数。

表4：在*8*核中央处理器上的推理效率对比。实时因子的测量方式为$(latency/duration)$。速度 $= 1/$实时因子。

方法	参数	实时因子 ↓	速度 ↑
AudioSR	258M	50.0000	0.02×
NVSR	∼99M	0.2218	4.5×
AP-BWE	3*30M	0.0495	20.2×
所提方法	**15M**	**0.0053**	**190.5×**

如表4所示，我们的模型在中央处理器上的速度几乎提升了 10×，优于此前最先进的高效模型（AP-BWE）。在 NVIDIA A100 GPU 上（表5），效率提升更为显著。我们的模型在批次大小为 1 时，仅需 2.5 毫秒即可处理 4 秒的音频（1600×实时）。

表5：*NVIDIA A100* 上的 *GPU* 效率与吞吐量对比。延迟按整个批次报告。实时因子四舍五入，并按 $latency/(duration \times BS)$ 计算。

方法	BS	延迟（秒）	实时因子 ↓	速度 ↑
AudioSR	1	8.4700	2.1175	0.47×
NVSR	1	0.0410	0.0103	97.6×
AP-BWE	1	0.0135	0.0034	296.3×
AP-BWE	32	0.2879	0.0023	444.6×
所提方法	1	**0.0025**	**0.0006**	**1600.0×**
所提方法	32	**0.0102**	**0.0001**	**12549.0×**

至关重要的是，我们的架构能够从批处理中特别受益。在批次大小为 32 时，我们的模型总吞吐量超过 12，500× 实时，处理 128 秒音频仅需 10.2 毫秒。相比之下，AP-BWE 在相同批次大小下需要 287.9 毫秒（0.0023 实时因子）。因此，所提模型更适用于高吞吐量的云端处理和实时边缘应用。这种高效率源于我们精简的 Vocos 架构以及林克维茨-莱利细化器；后者避免了生成对抗网络和扩散模型基线方法所需的繁重多阶段上采样和迭代采样。

5. 结论

我们提出了一种基于 Vocos 的带宽扩展模型，支持 8–48 kHz 范围内的任意输入采样率，同时实现了极高的吞吐量。通过将输入重采样至 48 kHz、采用基于 ConvNeXt 的 Vocos 生成器，并应用轻量级的受林克维茨启发的频域细化器，所提模型在频谱和感知质量方面达到了与强基线方法相当的水平。未来的工作包括在音乐和噪声条件下进行评估，并探索适用于不同任务的自适应细化器。

6. 致谢

我们衷心感谢 Yongyi Zang 提供的宝贵反馈，并为这篇 arXiv 投稿提供必要的背书。此外，我们感谢开源社区在项目代码库首次发布后给予的支持和建设性讨论。早期采用者的反馈和参与对于完善最终稿件、提升这项工作的实践层面清晰度起到了重要作用。

7. 参考文献

[1] H. Liu、K. Chen、Q. Tian、W. Wang 和 M. D. Plumbley，“Au- dioSR：大规模多功能音频超分辨率”，载于*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2024年，第1076–1080页。[2] Y.-X. Lu、Y. Ai、H.-P. Du 和 Z.-H. Ling，“通过并行幅度和相位预测实现高质量、高效率的语音带宽扩展”，*IEEE/ACM*音频、语音与语言处理汇刊，第33卷，第236–250页，2024年。[3] H. Siuzdak，“Vocos：弥合时域与基于傅里叶变换的神经声码器之间的差距，实现高质量音频合成”，*arXiv*预印本 *arXiv:2306.00814*，2023年。[4] S. H. Linkwitz，“用于非共轴驱动单元的有源分频网络”，《音频工程学会期刊》，第24卷，第1期，第2–8页，1976年。[5] C. Veaux、J. Yamagishi 和 K. MacDonald，“CSTR VCTK语料库：供CSTR语音克隆工具包使用的英语多说话人语料库（版本0.92）”，爱丁堡大学。信息学院。语音技术研究中心（CSTR），技术报告，2017年。[6] J. Kong、J. Kim 和 J. Bae，“HiFi-GAN：用于高效、高保真语音合成的生成对抗网络”，载于神经信息处理系统进展（*NeurIPS*），第33卷，2020年，第17 022–17 033页。[7] H. Liu、W. Choi、X. Liu、Q. Kong、Q. Tian 和 D. Wang，“神经声码器是语音超分辨率所需的一切”，载于*Interspeech 2022*论文集，2022年，第4227–4231页。[8] M. Chinen、F. S. Lim、J. Skoglund、N. Gusev 和 A. Hines，“ViSQOL v3：一种用于语音和音频的客观音频质量指标”，载于*2020*年*IEEE*第*22*届多媒体信号处理国际研讨会 *(MMSP)*。IEEE，2020年，第1–6页。