HR-Mel: High-Resolution Mel Spectrogram with Frequency-Dependent Compression for Music Representation – A Preliminary Study

Kio
kiolaaoz223@gmail.com

December 7, 2025

Abstract

Standard log-mel spectrograms commonly used in music processing apply uniform compression across all frequencies, which may not optimally balance reconstruction fidelity across the spectrum. We propose HR-Mel (High-Resolution Mel), a 3-band mel-based representation with frequency-dependent compression designed to improve reconstruction fidelity while maintaining compact dimensionality. HR-Mel divides the spectrum into psychoacoustically-motivated bands (0–1.5 kHz, 1.5–6 kHz, 6–20 kHz) with 40, 32, and 24 bins respectively, applying log1p compression for low and mid bands, and sqrt(log1p) for the high band. In a preliminary evaluation on 12 music tracks (44.1 kHz) at fixed STFT parameters (n_fft=2048, hop=441), HR-Mel achieves 25.3% lower overall STFT reconstruction error compared to standard 96-bin log-mel (0.2845 vs 0.3810), driven primarily by improvements in the low-frequency band (<1.5 kHz: 27% improvement). This exploratory study provides initial evidence that frequency-dependent compression merits further investigation for neural audio codec front-ends, with comprehensive evaluation needed to assess practical utility.

1 Introduction

Music representation is fundamental to neural audio processing, serving as the input for tasks ranging from music generation to audio compression. The mel-scale spectrogram has become the de facto standard, leveraging psychoacoustic principles to approximate human auditory perception [1]. However, standard mel representations face a critical trade-off: they either sacrifice frequency resolution for computational efficiency or require prohibitively large dimensionality to capture fine spectral details.

This limitation is particularly problematic for music signals, where frequency content across the spectrum carries perceptually significant information. While low-frequency fundamentals require precise amplitude representation, high-frequency content including attack transients and acoustic "air" (brilliance >10 kHz) define timbre and spatial qualities essential to music perception. However, standard mel representations apply uniform log-compression across all frequencies, which fails to optimally balance these competing requirements across the spectrum.

We introduce HR-Mel (High-Resolution Mel), a rule-based mel representation that addresses this gap through two key innovations: (1) a 3-band frequency partitioning scheme with non-uniform bin allocation aligned with spectral energy distribution in music signals, and (2) frequency-dependent nonlinear compression that provides differentiated treatment across bands. Unlike recent learned multi-band ap-

proaches [2, 3], HR-Mel employs fixed, interpretable compression functions requiring no training, making it immediately deployable as a drop-in replacement for standard mel front-ends.

This paper presents a preliminary investigation into frequency-dependent mel compression. Our evaluation covers 12 music tracks with magnitude-only reconstruction analysis. While limited in scope, the results reveal interesting patterns: substantial improvements in low-frequency reconstruction (27%), with modest trade-offs in mid and high bands. Our contributions are:

- A 96-bin mel representation achieving 25.3% lower overall reconstruction error than uniform 96-bin log-mel
- Frequency-dependent compression strategy: log1p for 0–6 kHz, sqrt(log1p) for 6–20 kHz
- Band-wise analysis revealing frequency-specific reconstruction characteristics
- Ablation study of alternative bin allocations (32/32/32, 24/32/40)
- Open-source implementation enabling further research

2 Related Work

2.1 Mel-Scale Representations

The mel-scale, introduced by Stevens et al. [1], approximates the nonlinear frequency response of human hearing. Mel-frequency cepstral coefficients (MFCCs) [4] became ubiquitous in speech recognition, typically using logarithmic compression to model loudness perception.

2.2 Multi-Band Audio Processing

Multi-band approaches have proven effective across audio tasks. Band-Split RNN [5] explicitly partitions spectrograms into subbands for music source separation, with bandwidth choices guided by target instrument characteristics. In the neural codec domain, BSCodec [2] processes frequency bands through parallel encoder-quantizer-decoder modules, while MUFFIN [3] introduces multi-band spectral RVQ (MBS-RVQ) with psychoacoustic quantization allocation. However, these methods employ learned, end-to-end architectures requiring extensive training.

2.3 Neural Audio Codecs

Neural audio codecs [6–8] learn implicit perceptual models but typically apply uniform compression in latent space.

HR-Mel differs fundamentally from these approaches: it is a fixed, non-learned representation employing explicit frequency-dependent compression designed specifically for 44.1 kHz music front-ends.

3 Method

3.1 Design Rationale

Standard mel spectrograms apply uniform logarithmic compression across all frequencies:

$$M_{\text{standard}} = \log(1 + M_{\text{power}}) \tag{1}$$

where M_{power} is the mel-filtered power spectrogram. This uniform treatment fails to account for two key properties of music signals:

Energy Distribution: Music signals exhibit concentrated energy in low-mid frequencies (fundamentals and lower harmonics) with sparser but perceptually critical high-frequency content (upper harmonics, transients).

Representational Needs: While low frequencies benefit from increased bin count for precise amplitude resolution, high frequencies require careful compression to preserve sparse transient details.

3.2 HR-Mel Specification

HR-Mel employs a 3-band structure with frequency-dependent compression at fixed 44.1 kHz parameters (n_fft=2048, hop_length=441, win_length=2048):

Band 1 (0–1.5 kHz, 40 bins): Captures fundamentals and lower harmonics. Uses log1p compression:

$$M_1 = \log(1 + M_{\text{power},1}) \tag{2}$$

Band 2 (1.5–6 kHz, 32 bins): Covers mid-range harmonics and formants. Uses log1p compression:

$$M_2 = \log(1 + M_{\text{power},2}) \tag{3}$$

Band 3 (6–20 kHz, 24 bins): Targets high-frequency detail including attack, brilliance, and air. Uses sqrt(log1p) compression:

$$M_3 = \sqrt{\log(1 + M_{\text{power},3})} \tag{4}$$

The final representation concatenates all bands: $M_{\text{HR}} = [M_1; M_2; M_3]$ with total dimensionality 96.

Decoding: Each band is decoded with the inverse operation:

$$\hat{M}_{\text{power},1} = \exp(M_1) - 1 \tag{5}$$

$$\hat{M}_{\text{power},2} = \exp(M_2) - 1 \tag{6}$$

$$\hat{M}_{\text{power},3} = \exp(M_3^2) - 1 \tag{7}$$

3.3 Reconstruction Method

For evaluation, we reconstruct STFT power spectra from encoded representations using the Moore-Penrose pseudo-inverse. Given mel basis $B \in \mathbb{R}^{N_{\text{mel}} \times N_{\text{fft}}}$ and decoded mel power $\hat{M}_{\text{power}}$, reconstruction proceeds as:

$$\hat{S}_{\text{STFT}} = \max(B^{\dagger} \hat{M}_{\text{power}}, 0) \tag{8}$$

where $B^{\dagger}$ is the pseudo-inverse computed via SVD, and the max operation enforces non-negativity (physical constraint for power spectra). This reconstruction is magnitude-only; phase information is not considered in this preliminary study.

The pseudo-inverse provides a least-squares optimal reconstruction in the STFT domain, allowing fair comparison across different mel configurations with identical reconstruction methodology.

3.4 Compression Function Analysis

The sqrt(log1p) compression in Band 3 modifies the effective dynamic range relative to log1p. For small magnitudes (weak high-frequency energy), sqrt(log1p) yields larger encoded values than log1p, providing higher sensitivity. For large magnitudes (strong transients), the square root operation compresses more aggressively, limiting growth. However, our empirical results (Section 5) reveal that this design primarily benefits low-frequency reconstruction, with high frequencies showing modest degradation. This suggests that factors beyond compression function—such as bin allocation and basis construction—dominate high-frequency performance.

4 Experimental Setup

4.1 Dataset

This preliminary study evaluates on 12 music tracks from a personal collection, converted to mono at 44.1 kHz. Tracks average 2–3 minutes in duration and span diverse styles including vocal and instrumental content. While limited in size and diversity compared to standard benchmarks (e.g., MUSDB18), this dataset serves as an initial proof-of-concept for the frequency-dependent compression approach.

4.2 Baseline Representations

We compare HR-Mel against:

- **STFT Power:** 1025-bin power spectrogram (reference)
- **Mel-80:** Standard 80-bin power mel
- **Log-Mel-80:** 80-bin with log1p compression
- **Mel-96:** 96-bin power mel (matched dimensionality)
- **Log-Mel-96:** 96-bin with uniform log1p compression

All representations use identical STFT parameters (n_fft=2048, hop=441, win=2048) and mel basis construction (Slaney normalization, fmax=20,000 Hz).

4.3 Evaluation Metrics

Overall Reconstruction Error: We measure relative Frobenius norm error after pseudo-inverse reconstruction:

$$\epsilon_{\text{rel}} = \frac{\|S_{\text{STFT}} - \hat{S}_{\text{STFT}}\|_F}{\|S_{\text{STFT}}\|_F} \tag{9}$$

where $\hat{S}_{\text{STFT}}$ is reconstructed via pseudo-inverse as described in Section 3.3.

Band-Wise Analysis: To understand frequency-specific performance, we compute reconstruction error separately for three frequency ranges: low (<1.5 kHz), mid (1.5–6 kHz), and high (>6 kHz), by masking STFT bins outside each range.

Note on Phase: This evaluation considers magnitude reconstruction only. Phase information, critical for perceptual audio quality, is not addressed in this preliminary investigation.

5 Results

5.1 Overall Performance

Table 1 presents overall reconstruction error across all representations. HR-Mel (40/32/24 configuration) achieves the lowest error (0.2845 ± 0.0410), representing 25.3% improvement over same-dimensionality Log-Mel-96 (0.3810 ± 0.0395) and 35.9% improvement over 80-bin baselines (0.4438 ± 0.0404).

Notably, Mel-80 and Log-Mel-80 show identical errors, as do Mel-96 and Log-Mel-96. This indicates that log1p compression is perfectly reversible and introduces no information loss in the reconstruction process—the gains from HR-Mel stem from the band-specific design, not merely from compression choice.

5.2 Band-Wise Analysis

Table 2 reveals striking frequency-dependent patterns. HR-Mel achieves substantial gains in low frequencies (<1.5 kHz: 27.0% improvement), while showing modest degradation in mid (1.2%) and high (2.6%) bands compared to Mel-96.

This pattern indicates that HR-Mel's 40-bin allocation to the low band (vs uniform mel's effective 30 bins <1.5 kHz) provides substantial reconstruction benefits where music signals have highest energy density. The high band's 24 bins prove insufficient to match uniform 96-bin mel's implicit high-frequency resolution.

Table 1: Overall reconstruction performance (12 tracks, mean ± std)

Representation	Bins	Rel. Error ↓
STFT	1025	0.000
Mel-80	80	0.4438 ± 0.0404
Log-Mel-80	80	0.4438 ± 0.0404
Mel-96	96	0.3810 ± 0.0395
Log-Mel-96	96	0.3810 ± 0.0395
HR-Mel	**96**	**0.2845 ± 0.0410**

Table 2: Band-wise reconstruction errors (12 tracks, 40/32/24 config). Δ is improvement relative to Mel-96 for each band.

Band	Mel-96	HR-Mel	Δ
Low (<1.5 kHz)	0.3737 ± 0.0391	**0.2729 ± 0.0345**	+27.0%
Mid (1.5–6 kHz)	0.6701 ± 0.0372	0.6780 ± 0.0419	-1.2%
High (>6 kHz)	0.7186 ± 0.0548	0.7372 ± 0.0557	-2.6%

5.3 Ablation Study

To understand the effect of bin allocation, we tested two alternative configurations. Table 3 summarizes these results, with Δ computed relative to Mel-96 for each band.

32/32/32 (Balanced): Equal bins per band. Results show degraded low-band performance (-7.4%) with negligible high-band improvement (+0.3%), confirming that uniform allocation is suboptimal.

24/32/40 (High-Focused): Increased high-band bins to 40. While high-band error improved modestly (+3.1%), low-band performance suffered se-

Table 3: Ablation study: alternative bin allocations. Δ shows improvement over Mel-96 per band.

Config	Low	Mid	High
40/32/24 (Original)			
Mel-96	0.3737	0.6701	0.7186
HR-Mel	**0.2729**	0.6780	0.7372
Δ	**+27.0%**	-1.2%	-2.6%
32/32/32 (Balanced)			
HR-Mel	0.4007	0.6780	**0.7159**
Δ	-7.4%	-1.2%	**+0.3%**
24/32/40 (High-Focused)			
HR-Mel	0.4840	0.6780	0.6958
Δ	-30.5%	-1.2%	+3.1%

vere degradation (-30.5%), demonstrating the critical importance of adequate low-frequency representation.

6 Discussion

6.1 Why HR-Mel Works (and Where It Doesn't)

The 40/32/24 configuration's strong overall performance (25.3% improvement) is driven almost entirely by low-frequency gains. This aligns with music signal statistics: fundamentals and lower harmonics <1.5 kHz carry the bulk of signal energy, and allocating 40 bins (vs 30 for uniform mel) provides better resolution where it matters most for reconstruction error.

However, the high-band degradation (-2.6%) reveals a limitation: the 24-bin allocation combined with sqrt(log1p) compression proves insufficient to match uniform mel's implicit high-frequency resolution. Several factors may contribute:

Bin Scarcity: 24 bins spanning 6–20 kHz (14 kHz bandwidth) provide coarser resolution than uniform mel's effective allocation.

Compression Characteristics: While sqrt(log1p) theoretically provides different dynamic range characteristics than log1p, it may not optimally preserve the sparse high-frequency transients characteristic of music signals.

Basis Interaction: The multi-band mel basis construction may introduce artifacts at band boundaries that uniform mel avoids.

The ablation study confirms that bin allocation dominates compression function choice: increasing high-band bins to 40 improves high-frequency reconstruction (+3.1%) but at severe cost to low frequencies (-30.5%).

6.2 Comparison to Recent Work

HR-Mel's novelty lies in its simplicity and specificity:

Unlike BSCodec and MUFFIN [2, 3], which employ learned multi-band quantization within end-to-end neural codecs, HR-Mel is a fixed transformation requiring no training. This makes it immediately deployable and interpretable.

Unlike Band-Split RNN [5] for source separation, HR-Mel targets compact representation for generation tasks, optimizing reconstruction fidelity rather than source isolation.

6.3 Implications for Neural Codec Design

The results suggest several insights for mel-based codec front-ends:

Asymmetric Allocation: Devoting more capacity to low frequencies where music signals have highest energy yields better overall reconstruction than uniform allocation.

Frequency-Specific Optimization: Different frequency ranges may benefit from different design choices (bin count, compression, basis construction). One-size-fits-all approaches likely suboptimal.

Task-Dependent Trade-offs: The 27% low-frequency improvement vs 2.6% high-frequency degradation represents a specific trade-off. Applications prioritizing transient preservation (e.g., percussion-focused generation) may prefer different configurations.

6.4 Limitations and Future Work

This preliminary study has several significant limitations:

Limited Dataset: Results are based on 12 music tracks from a personal collection. Generalization to diverse genres, recording qualities, and production styles remains unvalidated. Comprehensive evaluation requires large-scale datasets (e.g., MUSDB18, FMA, MusicCaps).

Magnitude-Only Reconstruction: Our evaluation uses pseudo-inverse reconstruction of power spectra without phase information. Perceptual quality assessment requires phase reconstruction (Griffin-Lim or neural vocoder) and listening tests with human subjects.

No Perceptual Metrics: We report only Frobenius norm error. Standard perceptual metrics (PESQ, SI-SDR, ViSQOL) and subjective evaluation (MUSHRA) are essential for validating practical utility, as reconstruction error may not correlate with perceptual quality.

Fixed Parameters: HR-Mel is designed for 44.1 kHz with specific STFT parameters. Adapting to 48 kHz stereo or other configurations requires empirical validation. The band boundaries (1.5 kHz, 6 kHz) are heuristic rather than optimized.

No Downstream Evaluation: Integration with actual neural codecs (e.g., EnCodec, DAC), music generation models (MusicGen), or other downstream tasks is untested. The representational benefits may or may not translate to end-task performance.

Incomplete Ablation: We tested only three bin configurations. Systematic grid search over band boundaries, bin allocations, and compression functions remains unexplored. Alternative compressions (e.g., power-law families) warrant investigation.

Despite these limitations, the preliminary results—particularly the 27% low-frequency improvement—suggest frequency-dependent compression warrants further investigation. Future work should address:

- Large-scale evaluation across diverse music datasets with genre stratification
- Perceptual metrics and listening tests with phase reconstruction
- Integration into neural audio codecs as mel front-end replacement
- Development and evaluation of a dedicated neural audio codec (DevCodec) using HR-Mel as its front-end
- Extension to 48 kHz stereo for high-fidelity music applications
- Systematic optimization of band boundaries, bin allocation, and compression functions
- Evaluation on downstream tasks: music generation, audio coding, source separation
- Analysis of learned vs fixed multi-band approaches

7 Conclusion

We introduced HR-Mel, a 3-band mel representation with frequency-dependent compression that achieves 25.3% lower overall reconstruction error than uniform 96-bin log-mel in a preliminary evaluation on 12 music tracks. The design employs 40/32/24 bin allocation across low/mid/high frequency bands with log1p and sqrt(log1p) compression, requiring no training.

Band-wise analysis reveals that reconstruction error improvements concentrate in low frequencies (<1.5 kHz: 27% improvement), with modest high-frequency degradation (2.6%). Ablation studies confirm that bin allocation is the dominant factor, with balanced or high-focused configurations producing higher overall error.

This exploratory study provides initial evidence that asymmetric bin allocation can improve reconstruction fidelity for music signals, particularly in low-frequency regions. However, the limited evaluation scope (12 tracks, magnitude-only reconstruction, no perceptual metrics) means these results represent preliminary findings rather than conclusive validation.

The primary contribution is introducing frequency-dependent compression with asymmetric bin allocation as a design direction for mel representations,

and providing open-source implementation for community exploration. While the low-frequency reconstruction improvements are consistent within our test set, comprehensive validation with large-scale datasets, perceptual evaluation, and downstream task integration is required to determine practical utility.

As neural audio models continue to develop, front-end representation design remains an important consideration for balancing quality and efficiency. HR-Mel explores one approach—prioritizing low-frequency reconstruction fidelity through asymmetric capacity allocation—with measurable trade-offs that practitioners can assess for specific applications.

The code and analysis scripts are available at the project repository.

Acknowledgments

The author thanks the open-source audio processing community for foundational tools including librosa, NumPy, and SciPy.

References

[1] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude of pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.

[2] H. Wang, J. Shi, J. Tian, B. Li, K. Yu, and S. Watanabe, "BSCodec: A band-split neural codec for high-quality universal audio reconstruction," *arXiv preprint arXiv:2511.06150*, 2025.

[3] D. Ng, K. Zhou, Y.-W. Chao, Z. Xiong, B. Ma, and E. S. Chng, "Multi-band frequency reconstruction for neural psychoacoustic coding," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2025.

[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[5] Y. Luo and J. Yu, "Music source separation with band-split RNN," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1893–1901, 2023.

[6] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[7] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.

[8] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 495–507, 2022.