VAE-based Generative Modeling for Music Audio

Wantong Zhang

University of California, Berkeley, the United States
wtzhang@berkeley.edu

Abstract. This project studies unsupervised timbre representation learning using a lightweight convolutional Variational Autoencoder (VAE) trained on log-mel spectrograms of monophonic instrument notes. Timbre is a multidimensional perceptual attribute shaped by harmonic structure, spectral envelope, and transient behavior, and compact latent embeddings of timbre are useful for controllable synthesis and analysis. Building on the VAE framework and the β-VAE objective, a baseline CNN β-VAE is re-implemented and evaluated on the nsynth-mini dataset. Qualitative inspection of reconstructions shows that the baseline captures coarse energy patterns but suffers from harmonic oversmoothing and transient smearing, particularly in high-frequency partials. To address these limitations under constrained computation, two minimal modifications are introduced: U-Net-style skip connections to improve time–frequency detail propagation and a mel-axis gradient difference loss to penalize frequency-domain blurring. Comparative results indicate visibly sharper harmonic stacks and better localized onsets without adversarial training. The report concludes with practical considerations for stable training and directions for extending the approach to richer perceptual objectives and more complex musical textures.

Keywords: Variational autoencoder, β-VAE, Timbre, Log-mel spectrogram, Audio synthesis.

1 Introduction

This research investigates unsupervised timbre representation learning with a convolutional Variational Autoencoder (VAE) trained on log-mel spectrograms of monophonic instrument notes. Timbre encodes the perceptual texture that differentiates instruments at fixed pitch and loudness; unlike pitch and loudness, timbre is shaped by a combination of harmonic structure, spectral envelope, and transient behavior, which makes it difficult to model with simple parametric descriptions. Learning compact and manipulable timbre embeddings is therefore foundational for controllable audio synthesis, feature analysis, and cross-modal generation [1].

Early musical sound generation relied on deterministic signal-processing pipelines and physical models that explicitly simulate acoustic mechanisms. While such approaches can be highly accurate for specific instruments, they require substantial hand engineering and do not scale well to large and heterogeneous datasets. Deep

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_101

generative models provide an alternative by learning representations directly from data. The original VAE formulation offers a principled latent-variable framework with a continuous latent space and a tractable training objective [2].

In the audio domain, NSynth demonstrates that learned latent spaces can capture perceptually meaningful dimensions of musical timbre using a WaveNet-based autoencoder, enabling smooth interpolation between instruments in latent space [3]. RAVE and related work further show that spectral objectives such as multi-resolution STFT losses can improve reconstruction fidelity while supporting real-time synthesis [4,5]. Perceptually driven approaches, such as Back-to-Ear, argue for optimizing reconstruction in learned auditory feature spaces to better align with human listening judgments [6,7].

Motivated by these developments, this work re-implements a lightweight CNN β-VAE baseline for mel-spectrogram reconstruction and introduces minimal-compute improvements that sharpen harmonics and transient details while keeping model complexity essentially unchanged. The remainder of this report reviews related work, describes the dataset and preprocessing pipeline, presents baseline behavior and limitations, and evaluates perceptually motivated enhancements.

2 Background Study

2.1 NSynth: WaveNet Autoencoder for Musical Timbre

NSynth introduces an encoder that compresses short waveform segments into a latent vector z and an autoregressive WaveNet decoder that reconstructs audio sample by sample. The learned latent space enables smooth interpolation between instruments, providing a continuous and interpretable representation of timbre. However, autoregressive decoding is computationally expensive, motivating lighter convolutional alternatives for faster training and inference [3].

2.2 RAVE: Real-Time High-Fidelity VAE

RAVE addresses the latency–quality trade-off via a two-stage design: an RVQ-VAE trained with a multi-resolution STFT loss to capture structure across time–frequency scales, followed by adversarial fine-tuning to enhance perceptual quality while preserving stability and interpretability [4].

2.3 Back-to-Ear: Perceptually Driven Reconstruction

Back-to-Ear proposes optimizing reconstruction in a learned auditory feature space $\Phi(\cdot)$ so that perceptual similarity is directly encouraged [7].

$$L\_perc = ||\Phi(\hat{x}) - \Phi(x)||_1 \quad (1)$$

This reframes the target from numerically accurate spectra to perceptually faithful audio, improving listening-based evaluation metrics such as mean opinion scores [7].

3 Methodology and Baseline Construction

3.1 Dataset and Preprocessing

Experiments use the nsynth-mini dataset, a compact public subset of NSynth containing monophonic notes across diverse instruments at 16 kHz [8]. Approximately 3,000 clips are sampled to ensure timbral diversity while keeping training feasible. Each waveform is transformed into a log-mel spectrogram via STFT and mel filterbank projection, converted to decibel scale, and normalized to [0,1]. The overall pipeline from waveform input to reconstructed mel spectrogram is shown in Fig. 1.

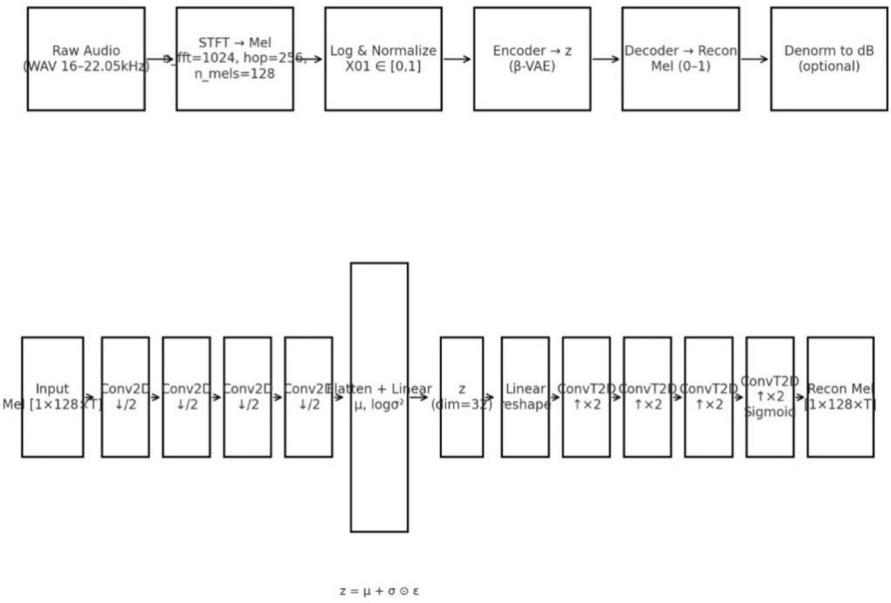

Fig. 1. Pipeline and architecture (Picture credit: Original)

3.2 Baseline CNN-βVAE

The baseline model uses a convolutional encoder with strided Conv2D layers and a symmetric transposed-convolution decoder. The encoder outputs the mean and log-variance of a diagonal Gaussian posterior, and the latent variable is sampled using the reparameterization trick. The model is trained using an $\ell 1$ reconstruction term with a β-weighted KL divergence to a standard normal prior [1,2].

4 Experiments and Model Improvements

4.1 Baseline Results and Analysis

The baseline CNN-βVAE is trained for 45 epochs with fixed β = 0.5 and without KL annealing. Checkpoints at 15, 30, and 45 epochs are compared. As shown in Fig. 2, the model learns coarse energy envelopes early in training but blurs harmonic structures and poorly localizes transient onsets. Later checkpoints exhibit slightly improved low-frequency stability, yet high-frequency partials remain attenuated and fine timbral detail fails to emerge. These patterns indicate harmonic oversmoothing and transient smearing under a purely pixelwise reconstruction objective [9, 10].

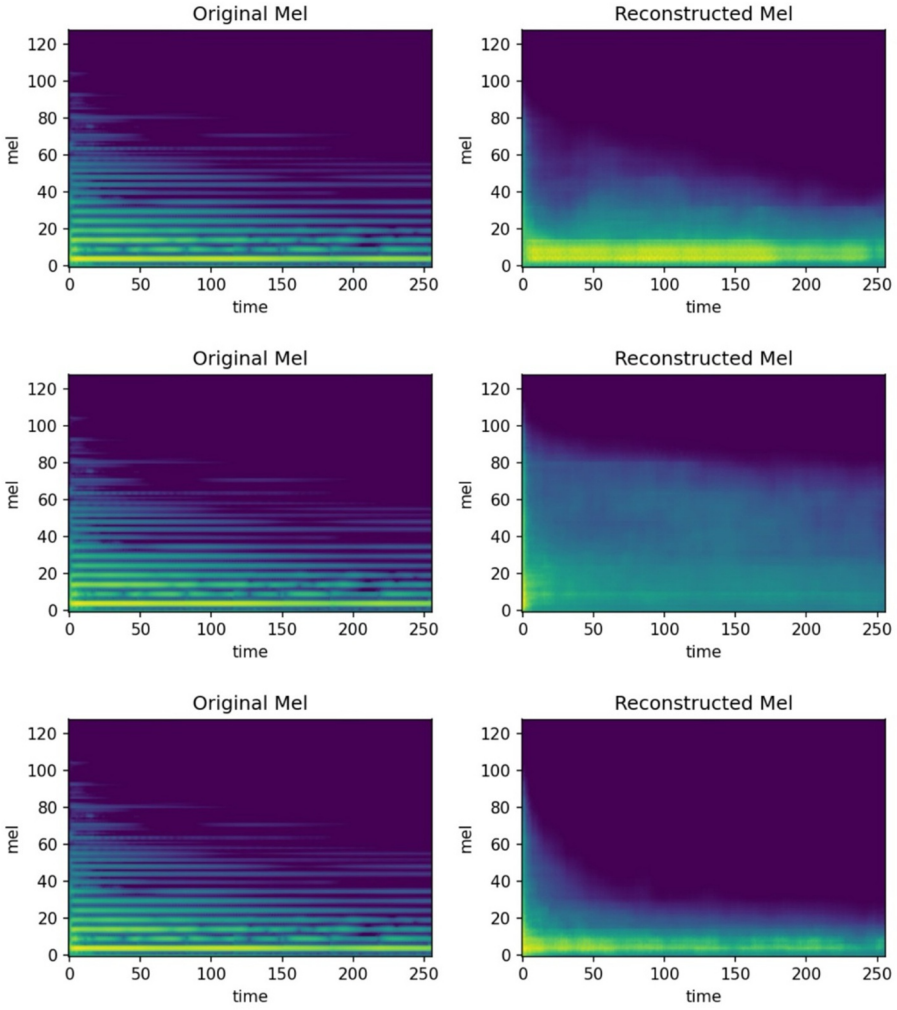

Fig. 2. Baseline reconstructions after 15, 30, and 45 epochs. (Picture credit: Original)

4.2 Model Enhancements and Training Stabilization

To address oversmoothed harmonics and blurred transients with minimal additional computation, two modifications are introduced. First, U-Net-style skip connections are added between matched encoder and decoder resolutions to allow local time–frequency detail to bypass the bottleneck. Second, a mel-axis gradient difference loss is added to penalize frequency-domain blurring. Training stability is improved using KL warm-up, log-variance clamping, and gradient clipping.

4.3 Comparative Visual Analysis

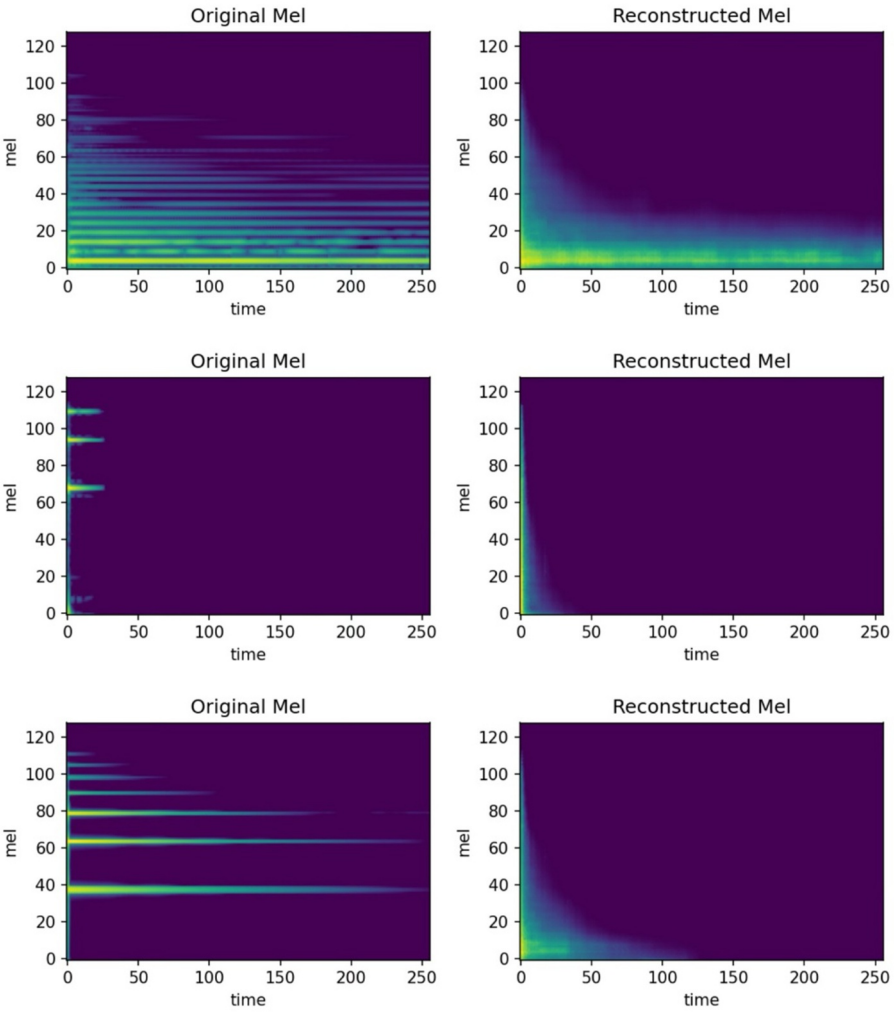

Fig. 3. Baseline reconstructions after 45 epochs (Picture credit: Original)

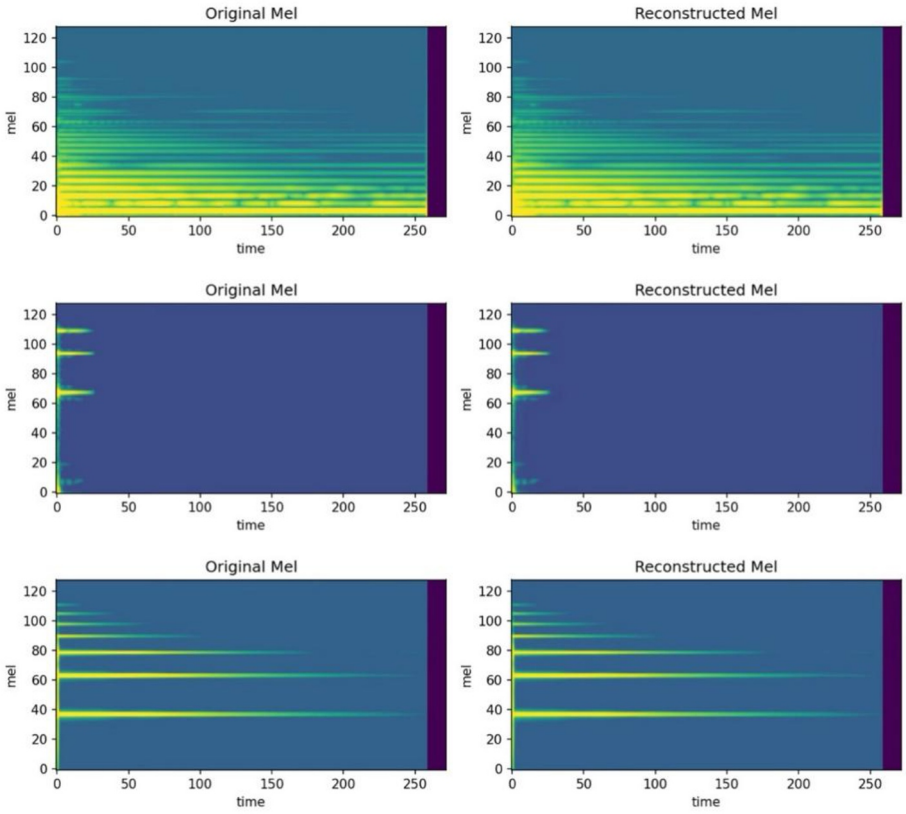

Fig. 4. Improved Model A reconstructions after 30 epochs (Picture credit: Original)

Representative reconstructions from the baseline after 45 epochs are shown in Fig. 3. While coarse energy patterns are captured, narrow harmonic peaks are averaged into broader frequency bands and transient structures are attenuated, demonstrating limited preservation of fine spectral–temporal detail. In contrast, Fig. 4 shows reconstructions from the improved Model A after 30 epochs. Harmonic stacks appear sharper and more vertically aligned, onsets are better localized in time, and high-frequency contrast is preserved without introducing noticeable noise artifacts.

5 Conclusions

This report re-implements a CNN-based β-VAE for musical timbre modeling and analyzes its limitations under an $\ell 1$ reconstruction objective. The baseline captures global energy structure but exhibits harmonic oversmoothing and transient blur. Lightweight skip connections and a mel-gradient difference loss mitigate these issues and yield visibly sharper reconstructions with stable, non-adversarial training. Future

work may explore systematic sweeps over β schedules, richer perceptual embedding spaces, and extensions to more complex musical textures.

References

1. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv:1312.6114. (2014).
2. Higgins, I., et al.: Ebeta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. ICLR (2017).
3. Engel, J., et al.: Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. arXiv:1704.01279 (2017).
4. Caillon, A., Esling, P.: RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv:2111.05011. (2021).
5. Yamamoto, R., et al.: Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. arXiv:1910.11480. (2020).
6. Défossez, A., et al.: High Fidelity Neural Audio Compression. arXiv:2210.13438. (2022).
7. Wang, K. D., et al.: Back to Ear: Perceptually Driven High Fidelity Music Reconstruction. arXiv (2025).
8. MTEB: NSynth-mini dataset. Hugging Face Datasets (2023).
9. Mathieu, M., et al.: Deep Multi-Scale Video Prediction Beyond Mean Square Error. arXiv:1511.05440 (2015).
10. Engel, J., et al.: DDSP: Differentiable Digital Signal Processing. arXiv:2001.04643. (2019).